De la teoría a la práctica:

Implementación del **Enfoque Basado en Argumentos** en el Examen de Ingreso a la Educación Superior (ExIES)

Karla Karina Ruiz Mendoza

astra editorial

De la teoría a la práctica:

implementación del Enfoque Basado en Argumentos del Examen de Ingreso a la Educación Superior (ExIES)

astra
editorial

De la teoría a la práctica:

implementación del Enfoque Basado en Argumentos del Examen de Ingreso a la Educación Superior (ExIES)

Karla Karina Ruiz Mendoza

astra
editorial

De la teoría a la práctica: implementación del Enfoque Basado en Argumentos del Examen de Ingreso a la Educación Superior (ExIES).
Autora: Karla Karina Ruiz Mendoza —*Baja California, México. 2026.*

194 p. 23 cm.

Primera edición

ISBN: **979-13-88142-59-8**

DOI: **https://doi.org/10.61728/AE26000701**

La presente obra fue dictaminada bajo el sistema de doble ciego y cuenta con el aval de los dictámenes de pares académicos en el campo de las ciencias sociales en México.

Edición y corrección: **Astra ediciones**

HECHO EN MÉXICO | MADE IN MEXICO

Listado de acrónimos

Acrónimo	Significado
ACT	American College Testing
AERA	American Educational Research Association
AIU	Argumento de Interpretación y Uso
ANUIES	Asociación Nacional de Universidades e Instituciones de Educación Superior
APA	American Psychological Association
CBC	Ciclo Básico Común
CENEVAL	Centro Nacional de Evaluación para la Educación Superior
CERQual	Confidence in Evidence from Reviews of Qualitative research
CGSEGE	Coordinación General de Servicios Estudiantiles y Gestión Escolar (UABC)
DIF	Differential Item Functioning (Funcionamiento Diferencial del Ítem)
DEMRE	Departamento de Evaluación, Medición y Registro Educacional (Chile)
EAI	Examen de Alto Impacto
EBA	Enfoque Basado en Argumentos
ECD	Evidence-Centered Design (Diseño Centrado en la Evidencia)
EMS	Educación Media Superior
ENEM	Exame Nacional do Ensino Médio (Brasil)
EXANI / EXANI-II	Examen Nacional de Ingreso (México)
EXHCOBA / EXCOBA	Examen de Habilidades y Conocimientos Básicos
ExIES	Examen de Ingreso a la Educación Superior
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IB	International Baccalaureate (Bachillerato Internacional)

IELTS	International English Language Testing System
ICFES	Instituto Colombiano para la Evaluación de la Educación
INEE	Instituto Nacional para la Evaluación de la Educación
INEGI	Instituto Nacional de Estadística y Geografía
IIDE	Instituto de Investigación y Desarrollo Educativo
MCCEMS	Marco Curricular Común de la Educación Media Superior
MCER	Marco Común Europeo de Referencia
NCME	National Council on Measurement in Education
PAES	Prueba de Acceso a la Educación Superior (Chile)
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSU	Prueba de Selección Universitaria (Chile, antecedente de la PAES)
RSL	Revisión Sistemática de la Literatura
SAT	Scholastic Assessment Test
SEP	Secretaría de Educación Pública (México)
TOEFL	Test of English as a Foreign Language
TCT	Teoría Clásica de los Tests
TRI	Teoría de Respuesta al Ítem
UABC	Universidad Autónoma de Baja California
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura
VIF	Variance Inflation Factor

Contenido

Introducción

El acceso a la educación superior suele definirse mediante procesos de selección basados en exámenes estandarizados, especialmente en contextos con alta demanda y en escenarios de transformación educativa (Mattos et al., 2024). Estos instrumentos, denominados Exámenes de Alto Impacto (EAI), están diseñados para generar resultados que inciden de manera decisiva en las trayectorias académicas y profesionales de los estudiantes; su relevancia radica en que permiten evaluar objetivamente competencias específicas, certificar logros educativos y regular de manera transparente la admisión a la educación superior, facilitando así la toma de decisiones educativas de gran relevancia social (French et al., 2024; Gutiérrez Domínguez, 2024; Jones y Ennes, 2018; UNESCO, 2021).

Desde una perspectiva internacional, los EAI se han consolidado como instrumentos clave en los procesos educativos de evaluación del aprendizaje y selección académica, manteniéndose vigentes por su capacidad para aportar datos comparables y confiables (Mattos et al., 2024). Estos instrumentos posibilitan la asignación eficiente de recursos y contribuyen a la garantía de calidad y rendición de cuentas en los sistemas educativos (Asociación Americana de Investigación Educativa [AERA], Asociación Americana de Psicología [APA] y Consejo Nacional para la Medición en la Educación [NCME], 2014; Gutiérrez Domínguez, 2024; UNESCO, 2021).

Existen múltiples ejemplos reconocidos, en Estados Unidos destacan el Scholastic Assessment Test (SAT) y el American College Testing (ACT) —aplicados desde 1926 y 1959 respectivamente— que se utilizan para la admisión universitaria (Marini et al., 2023; Sackett et al., 2008); en Colombia, el examen Saber 11 —aplicado desde 1998 y reformado en 2009— es de carácter obligatorio para el ingreso a la educación superior (ICFES, 2024a, 2024b); en Chile, la Prueba de Selección Universitaria (PSU entre 2003 a 2020), y Prueba de Acceso a la Educación Superior

(PAES desde 2022; DEMRE–U. de Chile, 2021); y en el ámbito internacional de dominio del inglés, el Test of English as a Foreign Language (TOEFL), desde 1964; y el English Language Testing System (IELTS), desde 1980; solicitados a quienes aspiran a cursar estudios en universidades de habla inglesa (Dang y Dang, 2021; Ihlenfeldt y Rios, 2023). Por su parte, en México, desde 2008, el Examen Nacional de Ingreso II (EXANI-II) se aplica ampliamente como método de admisión universitaria en diversas instituciones del país (Ceneval, 2022).

En este panorama, la Universidad Autónoma de Baja California (UABC) se ha sumado a la incorporación de exámenes de este tipo, donde cada año enfrenta una alta demanda regional para sus 146 licenciaturas, posicionándose como una de las instituciones de educación superior más relevantes del país, al ocupar el sexto lugar nacional en matrícula universitaria pública, con 66,885 estudiantes, cifra que representa el 5.30 % del total nacional (ANUIES, 2024). En 2023, por ejemplo, se obtuvo un total de 30 640 solicitudes para ingresar al nivel licenciatura (Pedroza et al., 2024a, 2024b), de las cuales se admitió al 64.53 % de los sustentantes (UABC-CGSEGE, 2024); distribuidos en los campus de Mexicali, Tijuana y Ensenada.

A fin de responder a esta elevada demanda, y considerando la diversidad de programas educativos, la UABC diseñó en el año 2017, por medio del Instituto de Investigaciones de Desarrollo Educativo (IIDE), el Examen de Ingreso a la Educación Superior (ExIES). El objetivo principal del ExIES consistió en evaluar competencias fundamentales en Lectura, Lengua Escrita y Matemáticas, alineándose con el Marco Curricular Común de la Educación Media Superior (MCCEMS) (Caso-Niebla et al., 2017). Esta primera versión del ExIES se dejó de aplicar en el año 2020 y, gracias al mantenimiento constante y la aplicación de pilotajes, particularmente en el ciclo 2022-2 (Pedroza et al., 2022), el ExIES se incluyó formalmente en el proceso de admisión en el ciclo 2023-1 (Pedroza et al., 2024a), siendo una prueba muy joven en comparación con pruebas internacionales, e incluso nacionales, como el SAT o el ACT.

El problema se formuló desde una concepción contemporánea de validez: la validez no es una propiedad fija de un instrumento, sino el grado en que la evidencia y la teoría respaldan las interpretaciones de

puntajes para usos propuestos (AERA et al., 2014; Messick, 1989). En EAI, esta precisión conceptual es decisiva porque una documentación incompleta puede invisibilizar riesgos de inequidad, usos no previstos o consecuencias indeseadas.

Si bien los Estándares (AERA et al., 2014) presentan una organización sistemática de los distintos tipos de evidencia de validez —proporcionando un marco conceptual estructurado—, también reconocen la necesidad de adaptabilidad contextual en la aplicación de tales lineamientos. Asimismo, los Estándares indican que la validación se concibe como un proceso sistemático mediante el cual se formulan y analizan argumentos que respaldan o cuestionan tanto la interpretación prevista de los resultados de una prueba como su pertinencia para los fines de uso establecidos (AERA et al., 2014, p. 12), enunciado que fundamenta la adopción del Enfoque Basado en Argumentos (EBA). Como señala Kane (2015), con frecuencia las especificaciones sobre los usos e interpretaciones previstos de los puntajes se presentan de forma incompleta o ambigua, lo cual genera vacíos conceptuales difíciles de identificar. Frente a ello, el EBA ofrece una estructura argumentativa más coherente y sistemática para sustentar la validez.

En este contexto, y dada la ausencia de un procedimiento único establecido por los Estándares para la construcción y evaluación de la interpretación de los puntajes o del argumento, el EBA se configuró como una propuesta metodológica sustentada en la lógica informal, estructurándose alrededor de un Argumento de Interpretación y Uso (AIU). Este AIU encadena distintas inferencias que vinculan los puntajes obtenidos con interpretaciones específicas y decisiones derivadas de ellos. Para evaluar la solidez del AIU, Kane (2006, 2013, 2015) propone tres criterios fundamentales: claridad (precisión y explicitud en las interpretaciones y usos previstos), coherencia (consistencia de la evidencia que respalda cada inferencia) y plausibilidad (credibilidad global del argumento basado en teoría y evidencia). Finalmente, el resultado de evaluar el AIU es la formulación del Argumento de Validez, cuyo propósito es establecer el grado en que las interpretaciones y decisiones basadas en los puntajes resulten adecuadas, justificadas y equitativas (Kane, 2006, 2013, 2015).

Por lo anterior, es relevante contar con una guía desde una visión sumativa, pero que también puede aplicarse en evaluaciones formativas.

En este caso, al ser aplicado en el ExIES, es una forma práctica e institucional: permite (a) sistematizar fuentes existentes (manuales, reportes técnicos, normativa y bases de datos), (b) identificar vacíos de evidencia que importan para la defendibilidad del uso de puntajes, y (c) traducir ese diagnóstico en prioridades de mejora bajo criterios de confiabilidad/precisión e imparcialidad, tal como lo exigen los Estándares (AERA et al., 2014). En términos metodológicos, este tipo de sistematización responde a una limitación reportada en la literatura: la variabilidad en cómo se redacta el AIU, se seleccionan inferencias y se organiza la evidencia, lo que dificulta comparar estudios y replicar procedimientos (Dursun y Li, 2021; Lavery et al., 2020).

Asimismo, en una decisión de admisión, hablar del concepto de validez permitió acercarse inevitablemente a la pregunta por la verdad: no la verdad absoluta de una medición perfecta, sino la verdad práctica de una interpretación que debía ser pública, argumentada y revisable (AERA et al., 2014; Rorty y Habermas, 2012). En este sentido, la validación rara vez produjo "verdades finales"; más bien, sostuvo juicios orientados por la utilidad y por la evidencia disponible (Cronbach, 1971).

En este marco, la validez puede leerse como un intento por acercar el puntaje a la realidad del constructo sin confundirlos. El examen no "contiene" la habilidad: la representa mediante tareas, condiciones y modelos; por ello, la objetividad no se redujo a neutralidad técnica, sino que se construyó cuando el razonamiento se volvió público y fue capaz de resistir la crítica. Hacer explícitas las razones —datos, garantías, respaldos y reservas— permitió que otras personas revisaran, refutaran o fortalecieran la interpretación propuesta (AERA et al., 2014; Kane, 2013; Messick, 1989; Toulmin, 2003). Desde una perspectiva pragmatista y falibilista, una afirmación se volvió más defendible en la medida en que se sostuvo ante evidencia nueva y en que hizo visibles sus límites, sin prometer certezas finales (Peirce, 1878; Rorty y Habermas, 2012).

De ahí que esta guía también pueda leerse como un ejercicio de historia: reconstruye cómo, en un momento específico (2023-2), se encadenaron decisiones técnicas, normas institucionales y datos para sostener un uso de puntajes. En términos de Koselleck (2000), el ExIES acumuló estratos de tiempo en forma de versiones, reportes y ajustes; cada estrato reconfiguró lo que pudo afirmarse con mayor o menor seguridad.

Cuando esos estratos no se documentan, la institución se apoya en una memoria frágil y selectiva: se recuerda el resultado (el puntaje), pero se olvida el razonamiento que lo hizo interpretable. Ricoeur (2004) ayuda a nombrar esta tensión entre memoria, historia y olvido: registrar evidencia y supuestos no es un trámite administrativo, sino una forma de cuidado epistemológico y ético. El EBA, al obligar a explicitar inferencias y reservas, convierte la experiencia del examen en memoria argumentada y facilita que los equipos aprendan, corrijan y rindan cuentas ante la comunidad.

Cabe destacar que el presente libro es el resultado de una tesis doctoral. Pensando en que el fin es la pedagogía del enfoque, la organización va en tres bloques organizativos. Primero, el Capítulo 1 establece el marco teórico reconstruyendo la evolución del concepto de validez (de forma breve pero necesaria), el proceso de validación y la lógica del EBA. Segundo, el Capítulo 2 describe la ruta metodológica y los instrumentos de trabajo (tablas, criterios y escalas) para construir y evaluar el AIU. Tercero, los Capítulos 3 al 8 desarrollan el caso ExIES por inferencia y el Capítulo 9 cierra con una síntesis global y orientaciones para replicar el procedimiento en otros ciclos o pruebas. Aunque el ejemplo central corresponde a un examen de admisión (sumativo y de alto impacto), la lógica del EBA es transferible: puede emplearse para justificar interpretaciones y usos en otros contextos educativos cuando se requiere transparencia, rendición de cuentas y mejora continua (AERA et al., 2014; Kane, 2013). Por ende, este texto constituye un esfuerzo para funcionar como guía para el evaluador de pruebas de manera representativa y visual, donde la teoría es parte fundamental del propio procedimiento.

Capítulo 1

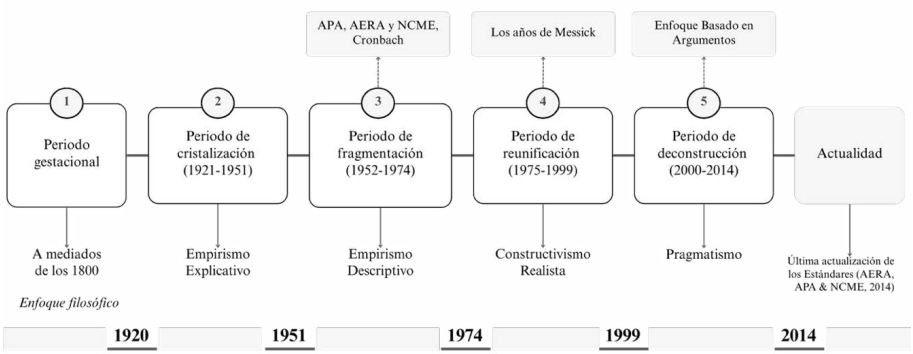
Validez, proceso de validación y EBA

A lo largo del tiempo, la noción de validez ha experimentado cambios teóricos y metodológicos. Newton y Shaw (2014) y García et al. (2017) señalan que esta evolución no ha sido lineal, sino que ha respondido a contextos históricos y exigencias cambiantes de la comunidad académica. Markus y Borsboom (2013) subrayan que, desde una perspectiva filosófica, la conceptualización de la validez varía según la corriente epistemológica dominante en cada periodo.

A partir de estas perspectivas complementarias, la Figura 1 presenta y sintetiza visualmente los periodos históricos clave y sus principales enfoques filosóficos, sirviendo como referencia para describir la evolución conceptual del término validez, desde mediados del siglo XIX hasta la actualidad. Esta revisión histórica es indispensable para comprender plenamente el EBA de Kane (2006, 2013, 2015, 2020), ya que evidencia el tránsito desde enfoques predominantemente estadísticos y específicos hacia modelos más integrales que enfatizan interpretación, uso y consecuencias de las evaluaciones educativas.

Figura 1

Línea del tiempo del concepto validez (1800-2014)



Nota. Elaboración propia que combina los enfoques filosóficos descritos en *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning* (K. A. Markus y D. Borsboom, 2013) con la periodización propuesta en *Validity in Educational y Psychological Assessment* (P. Newton y S. Shaw, 2014), como la reafirmación de *The evolution of concept of validity* (M. Kane y B. Bridgeman, 2021).

Periodo gestacional y de cristalización

En sus inicios, la validez se entendía como la correspondencia lógica entre lo que una prueba medía y el constructo o atributo que se pretendía representar (García et al., 2017). Aunque en este periodo inicial (mediados del siglo XIX hasta 1920) aún no existía una definición formal del término, se sentaron las bases fundamentales para el desarrollo de la psicometría y la investigación cuantitativa. Figuras clave como Francis Galton y James McKeen Cattell destacaron al explorar empíricamente las relaciones entre capacidades humanas y características físicas (Markus y Borsboom, 2013; Newton y Shaw, 2014). Karl Pearson (1896), por ejemplo, contribuyó de manera decisiva al desarrollar el coeficiente de correlación, lo que permitió el análisis sistemático y empírico de la relación entre las puntuaciones obtenidas en pruebas y distintos criterios externos (Newton y Shaw, 2014). Estas contribuciones fueron cruciales para el desarrollo de instrumentos como la escala de inteligencia Binet-Simon, consolidando la noción implícita de que los instrumentos debían medir adecuadamente el atributo pretendido (Newton y Shaw, 2014; Watson, 2002).

El término cristalización (1921-1951) hace referencia a cómo durante este periodo se consolidaron formalmente conceptos clave que sentaron las bases del enfoque moderno de validez. El movimiento estadounidense en medición buscó clarificar qué significaba concretamente que una prueba fuera considerada válida, mediante definiciones operativas claras. Lissitz (2009) y Newton y Shaw (2014) explican que dos perspectivas dominaron este periodo: una perspectiva lógica que analizaba si los ítems representaban adecuadamente el dominio o constructo evaluado, dando origen a la validez de contenido; y una perspectiva empírica basada en correlaciones con criterios externos específicos, conocida posteriormente como validez de criterio.

Durante este periodo, diversos investigadores aportaron definiciones relevantes, las mismas que se presentan en la Tabla 1. Garrett (1937), citado por Lissitz (2009), enfatizó claramente que la validez de una prueba radicaba en la fidelidad con que medía lo que pretendía medir, aunque sin considerar explícitamente sus implicaciones sociales. Bingham (1946), también referido en Lissitz (2009), priorizó la evidencia empírica en

forma de correlaciones con criterios externos. Por otro lado, Guilford (1946), citado en Messick (1989), propuso una postura más radical al considerar que cualquier resultado correlacionado con una variable externa podía interpretarse como evidencia suficiente de validez.

Tabla 1.
Definiciones del concepto de validez en el periodo de cristalización

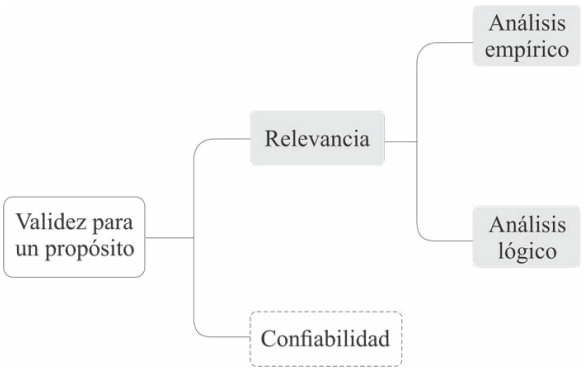
Autor	Año	Definición
Garrett	1937	"(...) la validez de un test es la fidelidad con la que mide lo que pretende medir (...)" (Lissitz, 2009, p. 23).
Bingham	1946	"(...) la correlación de las puntuaciones de un test con alguna otra medida objetiva de lo que el test quiere medir (...)" (Lissitz, 2009, p. 23).
Guilford	1946	"(...) una prueba es válida para cualquier cosa con la que se correlaciona (...)" (citado en Messick, 1989, p. 18).

Nota. Elaboración y traducción propia basado en "Validity" (S. Messick, 1989, en R. L. Linn [Ed.], *Educational Measurement*, 3.ª ed., pp. 13-103) y en *The Concept of Validity: Revisions, New Directions, and Applications* (R. Lissitz, 2009).

Hacia el final de este periodo, Cureton (1951) realizó un aporte importante al integrar estas diversas perspectivas en un marco unificado. Propuso considerar dos componentes fundamentales: la relevancia lógica, referida a la representatividad del dominio evaluado (validez de contenido), y la relevancia empírica, centrada en la correlación de la prueba con un criterio externo específico (validez de criterio) (Chapelle, 2021). Cureton destacó también la importancia de la confiabilidad, subrayando que mediciones consistentes eran fundamentales para tomar decisiones válidas (Chapelle, 2021).

Este marco integrador sentó bases sólidas para futuras concepciones unificadoras, como la validez basada en argumentos propuesta posteriormente por Kane, que establece un enfoque lógico que articula diversos tipos de evidencias para sustentar interpretaciones y usos específicos de los resultados de pruebas (Chapelle, 2021). La Figura 2 ilustra gráficamente el modelo propuesto por Cureton (1951), destacando los componentes clave que configuraron significativamente la evolución conceptual de la validez.

Figura 2.
Diagrama sobre los componentes de la validez según Cureton (1951)



Nota. Adaptado de *Argument-based validation in testing and assessment* [traducción propia] (p. 5), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Periodo de fragmentación y consolidación

Tras la publicación inicial de las recomendaciones técnicas por parte de la APA (1954) y posteriormente con la primera edición de los *Standards for Educational and Psychological Tests and Manuals* (AERA et al., 1966), surgió una etapa denominada periodo de fragmentación (1952-1974), caracterizada por la identificación explícita y formalización de diversos tipos específicos de validez. Durante esta fase, predominó un enfoque empírico descriptivo, donde la validez se clasificó claramente en tres categorías principales: validez de contenido, validez de criterio (predictiva y concurrente) y validez de constructo, cada una con objetivos y métodos de validación diferenciados (Newton y Shaw, 2014). Esta fragmentación reflejó un esfuerzo sistemático por clarificar metodológicamente cómo las pruebas debían justificar su adecuación para diferentes usos y contextos evaluativos. La Tabla 2 muestra la caracterización de estos enfoques según surgían en este periodo y cómo se modificó su pregunta central.

Tabla 2.
Enfoques de validez y sus características según el periodo de fragmentación.

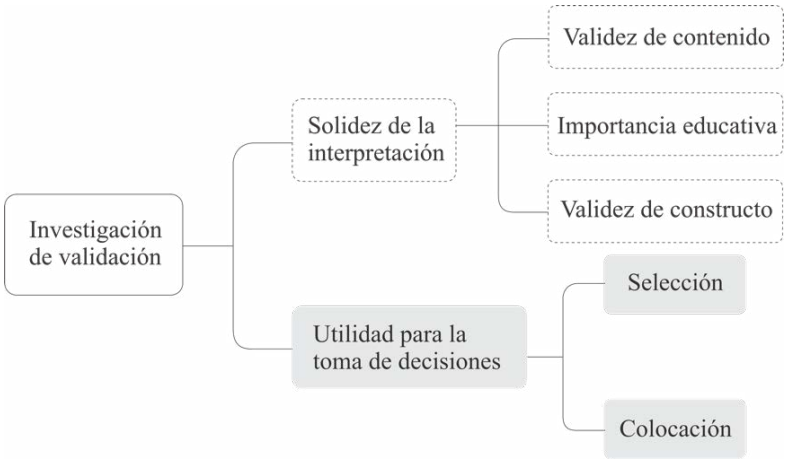
Enfoque de validez	Pregunta que responde	Evidencia característica	Ejemplo
Validez de contenido	¿Los ítems cubren de forma representativa el dominio o temario que se desea medir?	Juicio de expertos, mapeo de contenidos, índices de relevancia	Matriz de especificaciones (blueprint, en inglés) de un examen curricular
Validez predictiva (subtipo de validez de criterio)	¿Las puntuaciones anticipan con precisión un desempeño futuro relevante?	Correlaciones entre la prueba y un criterio externo medido en el futuro	Puntuación de ingreso vs. promedio del primer año universitario
Validez concurrente (subtipo de validez de criterio)	¿Las puntuaciones se relacionan con un criterio externo medido al mismo tiempo?	Correlaciones con otra prueba validada o con calificativos actuales	Examen diagnóstico correlacionado con calificación semestral
Validez de constructo	¿La prueba realmente mide el constructo teórico que afirma medir?	Análisis factorial, relaciones convergentes/discriminantes, método de grupos conocidos	Estructura interna de una prueba de ansiedad

Nota. Elaboración propia basada en *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (AERA et al., 1966/1974), "Construct Validity in Psychological Tests" (L. J. Cronbach & P. E. Meehl, 1955) y *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning* (K. A. Markus & D. Borsboom, 2013).

Cronbach (1971), en la segunda edición de *Educational Measurement*, recalcó que la validez no era una propiedad inherente a la prueba, sino específicamente de las interpretaciones derivadas de sus resultados. Este enfoque destacó la relevancia de la validación para tomar decisiones informadas en contextos prácticos, particularmente en procesos de selección y colocación educativa o laboral. La Figura 3 presenta el esquema propuesto por Chapelle (2021), para definir los tipos específicos de investigación necesarios para sustentar distintas inferencias interpretativas y decisiones derivadas de las puntuaciones, según Cronbach (1971).

Este llamado periodo de fragmentación (Newton y Shaw, 2014), que concluye en 1974 con una nueva edición de los Estándares, afirmó la relevancia de la validez de contenido, de criterio y de constructo sin considerarlas excluyentes. No obstante, su presentación en categorías separadas fomentó la idea de que se trataba de formas de validez distintas en lugar de evidencias complementarias.

Figura 3.
Tipos de investigación sobre validación definidos por Cronbach en 1971



Nota. Adaptado de *Argument-based validation in testing and assessment* [traducción propia] (p. 7), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Hacia una visión unificada: Messick y las fuentes de evidencia

Partiendo de lo anterior, la fase de re-unificación (1975-1999) se conoce como la etapa de Messick (Messick, 1989; Newton y Shaw, 2014). Durante este tiempo se destacó especialmente la validez de constructo, entendiéndola no como una simple categoría aislada, sino como el eje integrador de distintas facetas y evidencias relacionadas con la interpretación de los puntajes en pruebas educativas y psicológicas. Desde la perspectiva del constructivismo realista, Messick (1989) sostuvo que la validez es un concepto unitario respaldado por múltiples fuentes de

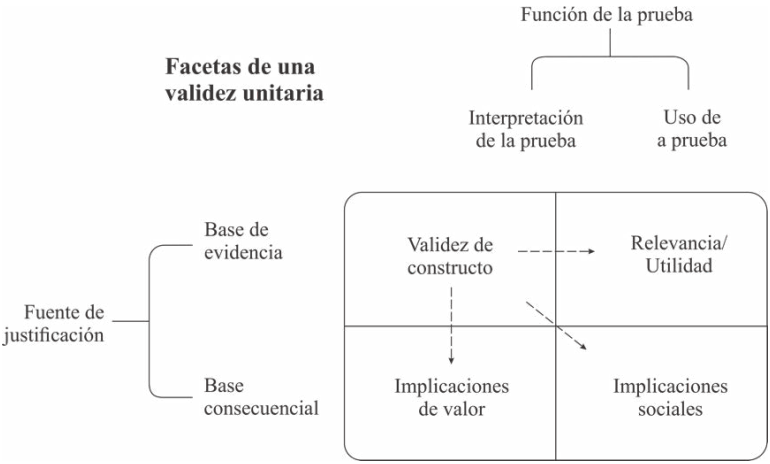
evidencia, incluyendo de forma explícita las consecuencias sociales y éticas derivadas del uso de las pruebas.

Previamente, otras contribuciones fueron claves para que Messick concretara esta visión. House (1980), por ejemplo, destacó el papel fundamental que desempeña la argumentación persuasiva en la evaluación, concibiéndola no simplemente como una demostración científica orientada hacia una comunidad racional universal, sino como una narrativa dirigida a audiencias específicas. En este sentido, la validez dependía de la coherencia y credibilidad del argumento presentado, así como de la capacidad estética (uso de metáforas, imágenes y narrativa) para generar sentido e interpretar hallazgos de manera significativa. Asimismo, House (1980) introdujo la noción de justicia, que en el ámbito de la evaluación en español suele traducirse como imparcialidad debido a las dificultades conceptuales relacionadas con la percepción pública sobre cómo se interpretan y utilizan los puntajes obtenidos en las pruebas. Para House (1980), diferentes teorías filosóficas sobre justicia (como el utilitarismo, el pluralismo intuicionista y la equidad o imparcialidad) ofrecen marcos desde los cuales evaluar la legitimidad y pertinencia de las decisiones tomadas a partir de una prueba. Por ejemplo, una evaluación puede considerarse válida en la medida en que represente adecuadamente los intereses de todos los grupos involucrados, especialmente los más vulnerables o menos favorecidos. Así, la validez no solo implica criterios técnicos y empíricos, sino también consideraciones éticas relacionadas con la equidad y el impacto social.

En este contexto, Messick amplió el concepto de validez, integrando aspectos técnicos, empíricos y éticos, argumentando que los puntajes de una prueba adquieren significado a partir de la teoría subyacente del constructo evaluado, la evidencia empírica disponible y las consecuencias potenciales del uso de estos puntajes (Messick, 1989). Su contribución implicó reconocer que las consecuencias sociales y éticas no son aspectos periféricos, sino centrales para evaluar la pertinencia y legitimidad de cualquier instrumento evaluativo. Al mismo tiempo, implicó la reflexión hacia el proceso de validación, es decir, la puesta operativa y práctica, ya que cada vez este concepto se complejizaba aún más.

La Figura 4 ilustra gráficamente el modelo propuesto por Messick, mostrando cómo distintas fuentes de evidencia y consideraciones éticas confluían en un concepto unificado de validez, alejándose de la fragmentación tradicional por tipos específicos; haciendo con la concepción fuera fragmentada y vista como distintos tipos de validez.

Figura 4.
Diagrama de las facetas de la validez definido por Messick en 1989



Nota. Adaptado de *Argument-based validation in testing and assessment* [traducción propia] (p. 10), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Este modelo influyó significativamente en las ediciones posteriores de los Estándares (AERA et al., 1999), consolidando una visión integrada y comprehensiva del proceso de validación.

Periodo de deconstrucción, hacia un enfoque basado en argumentos

El periodo de deconstrucción (2000-2014), como se iba mencionando, se caracterizó por un intenso cuestionamiento hacia el enfoque unitario y multifacético de validez propuesto por Messick (1989). Aunque el modelo integrador de Messick había consolidado un marco teórico amplio al incluir múltiples fuentes de evidencia y las consecuencias sociales y

éticas, surgieron importantes dificultades operativas y prácticas debido a su complejidad conceptual y metodológica (Newton y Shaw, 2014). En este contexto, según Newton y Shaw (2014), diversos investigadores comenzaron a cuestionar la viabilidad de aplicar una concepción tan amplia y abstracta de validez en contextos reales y específicos. Así, se mantuvo una simbiosis compleja entre el concepto propio de validez y el proceso de validación (teórico-práctico).

Entre las críticas más influyentes estuvieron las planteadas por Borsboom y colaboradores (2004, 2009) —desde una perspectiva ontológica—, quienes enfatizaron la necesidad de retornar a una perspectiva más concreta y operativa. Estos autores argumentaron que la validez debería centrarse en la capacidad intrínseca del instrumento para medir efectivamente el constructo que pretende evaluar, introduciendo así un enfoque ontológico centrado en la relación causal entre el constructo y las respuestas obtenidas.

Simultáneamente, Embretson (2007) propuso un enfoque cognitivo para el proceso de validación, subrayando la importancia de analizar con precisión los procesos internos que subyacen a las respuestas de los individuos en las pruebas. Este enfoque cognitivo ayudó a clarificar cómo se representaban los constructos en las tareas específicas de evaluación, proporcionando así una herramienta analítica adicional para evaluar la calidad interpretativa y la pertinencia de los resultados.

En paralelo, Mislevy, Steinberg y Almond (2003) introdujeron el Diseño Centrado en la Evidencia (ECD, por sus siglas en inglés), un enfoque metodológico sistemático orientado hacia la especificación precisa de los atributos que deben medirse, los tipos de evidencia necesarios y cómo interpretar adecuadamente dicha evidencia en contextos particulares. El ECD enfatizó la importancia de explicitar claramente cómo y por qué las tareas específicas proporcionan evidencia relevante sobre los constructos evaluados.

A partir de estas críticas y desarrollos alternativos, surgió el EBA, propuesto por Kane (1992, 2006, 2009, 2013). El EBA proporcionó un marco explícito y sistemático que permitía articular claramente las inferencias, los supuestos y las evidencias requeridas para respaldar interpretaciones específicas de los puntajes obtenidos en pruebas y evaluaciones. Este

enfoque retomó la propuesta argumentativa inicial de Cronbach (1988), llevándola a un nivel de precisión y claridad metodológica que permitía abordar las limitaciones identificadas en el modelo de Messick.

La entrada del EBA y de enfoques similares como el Argumento para el Uso de la Evaluación (AUA) de Bachman y Palmer (2010) permitió establecer estructuras claras y explícitas para validar interpretaciones particulares en diversos contextos, superando parcialmente las limitaciones del enfoque unitario. La flexibilidad y la especificidad del EBA permitieron adaptar el proceso de validación a contextos concretos, facilitando una mejor articulación entre teoría, evidencia empírica y aplicaciones prácticas.

No obstante, la postura realista-causal, defendida Markus y Borsboom (2013), ha seguido estando presente de forma paralela a estas otras propuestas como el EBA o el AUA, ya que concibe la validez como una propiedad objetiva, estable e independiente de los usos sociales de la prueba; es decir, una prueba es válida si y solo si el constructo que se afirma medir realmente existe y causa las respuestas observadas. Esta visión, que se alinea con un paradigma positivista clásico, busca resultados concluyentes, abarcadores y generalizables, similares a los objetivos de la historia total en la historiografía, que aspiraba a ofrecer explicaciones amplias, sistemáticas y unificadoras de la realidad. En la práctica evaluativa, esta perspectiva favoreció el diseño de instrumentos como los exámenes estandarizados de ingreso a la universidad que intentan resumir el mérito académico en un único puntaje, prescindiendo del contexto educativo o social del aspirante. Sin embargo, con el paso del tiempo generó críticas por ignorar la complejidad de los procesos de aprendizaje, así como las condiciones socioculturales que afectan el desempeño, produciendo decisiones de alto impacto basadas en supuestos de neutralidad y universalidad difíciles de sostener.

Estos desarrollos influyeron directamente en la formulación del concepto actual de validez en los estándares publicados en 2014 (AERA et al., 2014), en los que la validez se define explícitamente como el grado en que la evidencia empírica y la teoría respaldan las interpretaciones propuestas para los resultados obtenidos en función de su uso previsto.

Así, el periodo de deconstrucción y la entrada del EBA fueron parte de la reconfiguración hacia un concepto contemporáneo más pragmático, operativo y sensible al contexto, centrado en la interpretación y uso específico de los resultados de las evaluaciones.

Proceso de validación

Como se revisó en la sección anterior, la validez ha enfrentado el problema de operacionalización, es decir, el de convertir sus definiciones teóricas en indicadores y procedimientos empíricos claros. El concepto de proceso de validación cubre esta situación y, de la misma forma, también ha experimentado cambios metodológicos y conceptuales reflejados en distintas ediciones de los estándares (véase Tabla 3). En términos estrictos, en la actualidad, el proceso de validación es la indagación sistemática mediante la cual se reúne y evalúa esa evidencia para justificar dichas interpretaciones y usos (AERA et al., 2014; Kane, 2013).

Inicialmente, la edición de 1955 proporcionó recomendaciones técnicas básicas enfocadas en la documentación técnica de pruebas (AERA y NCME, 1955). Posteriormente, en las ediciones de 1966 y 1974, se adoptó una perspectiva jerarquizada que destacó el desarrollo, uso y reporte organizado de resultados, generando niveles explícitos de importancia dentro del proceso de validación (AERA et al., 1966, 1974).

A partir de 1985, se produjo un cambio hacia un enfoque más integral y acumulativo, abandonando las jerarquías explícitas y enfatizando la necesidad de considerar múltiples fuentes de evidencia en el proceso de validación, aunque aún sin una estructura sistemáticamente articulada (Eignor, 2013). Aunque, el gran cambio surgió en su versión del año 1999, al eliminar completamente las categorizaciones jerárquicas, enfocándose en la contextualización detallada y amplia de diversas fuentes de evidencia tales como contenido, procesos de respuesta, estructura interna, relaciones externas y consecuencias del uso de las pruebas (AERA et al., 1999; Zhu, 2001). No obstante, esta edición carecía aún de un marco argumentativo explícito y sistemático (Kane, 2020).

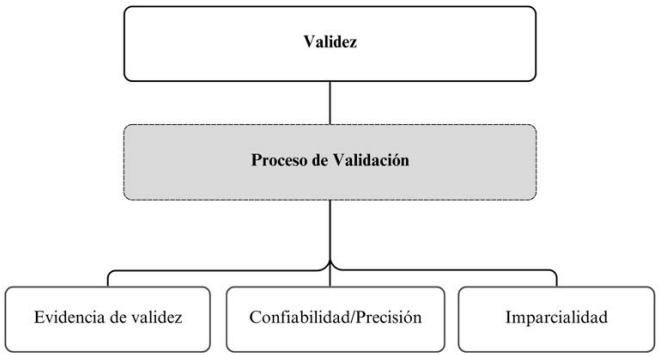
Tabla 3.
Evolución histórica de la concepción del proceso de validación en los estándares

Año	Enfoque dominante	Síntesis del proceso de validación
1954/1955	Recomendaciones técnicas iniciales	Documentación técnica básica sobre las pruebas (AERA y NCME, 1955).
1966/1974	Jerarquía de estándares	Énfasis en desarrollo, uso y reporte jerarquizado de resultados (AERA et al., 1966, 1974).
1985	Linealidad de estándares	Enfoque integral y acumulativo sin jerarquías explícitas (Eignor, 2013).
1999	Múltiples fuentes sin estructura argumentativa	Contextualización detallada de múltiples fuentes de evidencia sin estructura argumentativa explícita (AERA et al., 1999; Zhu, 2001).
2014	Enfoque argumentativo holístico	Validación explícitamente estructurada mediante argumentos que integran evidencia empírica y teórica, enfatizando aspectos éticos y sociales (AERA et al., 2014; Kane, 2020).

Nota. Elaboración propia basada en *Technical Recommendations for Achievement Tests* (AERA & National Council on Measurements Used in Education [NCMUE], 1955), *Standards for Educational and Psychological Tests and Manuals* (AERA et al., 1966), *Standards for Educational and Psychological Tests* (2.ª ed.; AERA et al., 1974, 1999, 2014), "The *Standards for Educational and Psychological Testing*" (D. R. Eignor, 2013, en *APA Handbook of Testing and Assessment in Psychology*, vol. 1) y "Validity Studies Commentary" (M. Kane, 2020).

A partir de la edición más reciente de los Estándares (AERA et al., 2014), representada en la Figura 5, el proceso de validación actual se articula como un modelo integrado y sistemático. Es decir, el proceso de validación ocupa un lugar central dentro del concepto amplio de validez y se fundamenta en la obtención de evidencias robustas y articuladas alrededor de tres dimensiones principales: evidencia de validez, confiabilidad/precisión e imparcialidad.

Figura 5.
Integración del Proceso de Validación dentro del concepto de Validez



Nota. Elaboración propia basada en *Standards for Educational and Psychological Testing* (AERA et al., 2014). Copyright 2014 de AERA.

Si bien las evidencias de validez son centrales, estas se acompañan transversalmente con la imparcialidad y la confiabilidad/precisión, dado que los Estándares (AERA et al., 2014) establecen que la imparcialidad es un requisito esencial de la validez que debe gestionarse a lo largo de todo el ciclo de diseño, validación, aplicación y uso de la prueba, y exigen asimismo que cada interpretación prevista de los puntajes esté respaldada por evidencia adecuada de su consistencia y exactitud.

Evidencias de validez. Según la AERA et al. (2014), se evita definir una tipología rígida de validez enfatizando, en cambio, cinco fuentes principales de evidencia (AERA et al., 2014):

A. Evidencia basada en el contenido de la prueba: Este tipo de evidencia se centra en analizar los temas abordados, la forma en que están redactados los ítems y el formato que presentan, ya que estos aspectos constituyen el contenido del instrumento.
B. Evidencia basada en los procesos de respuesta: Este enfoque examina cómo responden los participantes a la prueba, considerando aspectos como las estrategias utilizadas para resolver los problemas, el tiempo que emplean en cada ítem o incluso el seguimiento ocular. Estas observaciones permiten identificar en qué medida habilidades ajenas al constructo evaluado pueden afectar el desempeño de los examinandos.

C. Evidencia basada en la estructura interna: Indican el grado de relación entre los ítems y los componentes de la prueba para saber si se alinean al constructo y, por lo tanto, a las interpretaciones de los puntos dados en la prueba.
D. Evidencia basada en relaciones con otras variables: Este tipo de evidencia sugiere la relación entre variables con otras variables externas para lograr un análisis de los puntajes; es decir, es una evidencia que busca coherencia. Sobre este tipo de evidencias podemos subtitularlas en tres: evidencia convergente (evidencia de un mismo constructo o similar para determinar puntajes) y discriminante (se da en la relación de los puntos de la prueba y medidas de constructos diferentes); relaciones prueba-criterio (dependerá de la confiabilidad, relevancia y validez de la interpretación); y generalización de validez (comúnmente refiere a los metaanálisis o estudios estadísticos que ayuden a generalizar un criterio).
E. Evidencia de validación y consecuencias de las pruebas: Se analizan las consecuencias probables de las pruebas, por lo que se debe realizar un análisis de las consecuencias de las consecuencias; si bien se pueden adquirir beneficios y una brújula para la toma de decisiones en instituciones o escuelas, hay que ser cautelosos.

Confiabilidad. Por su parte, la confiabilidad es un concepto que se ha abordado de manera sistemática desde los primeros Estándares —1955, 1966, 1974, 1999 y 2014—, siendo un principio central debido a, como ya se ha mencionado, la necesidad de contar con instrumentos precisos y estables; asimismo, ha sido un concepto bastante claro desde su concepción.

Los Estándares (AERA et al., 2014) establecen que la confiabilidad refiere a la consistencia, estabilidad y precisión de las puntuaciones obtenidas mediante un instrumento en distintas circunstancias. Según AERA et al. (2014), la confiabilidad es crucial porque garantiza que las interpretaciones basadas en los resultados sean robustas y no producto del azar o factores contextuales específicos. La confiabilidad puede verse afectada por diversas fuentes de error, incluyendo variaciones en las condiciones de aplicación, inconsistencias entre calificadores y fluctuaciones temporales en las respuestas de los evaluados.

La idea de confiabilidad se remonta a Spearman (1904), quien introdujo la teoría clásica de los tests (TCT), señalando que las puntuaciones observadas tienen un componente verdadero y uno de error. Posteriormente, Lord y Novick (1968) formalizaron la TCT, y a partir de allí se desarrollaron enfoques avanzados como la teoría de respuesta al ítem (TRI) (Lord, 1980) y la teoría de la generalizabilidad (Cronbach et al., 2004), que profundizan en la medición mediante métodos estadísticos avanzados (Raykov y Marcoulides, 2011).

Diversos métodos son utilizados para evaluar la confiabilidad, como se presenta en la Tabla 4. Entre ellos destacan el alfa de Cronbach, que evalúa la consistencia interna entre ítems; el método de prueba-reprueba, que mide la estabilidad temporal; formas equivalentes, que buscan asegurar equivalencia entre versiones paralelas de un instrumento; la teoría de la generalizabilidad, que analiza fuentes específicas de error; y la TRI, que proporciona análisis detallados y específicos por ítem y habilidad evaluada.

Tabla 4.
Métodos principales para la evaluación de la confiabilidad

Método	Descripción	Autores	Ventajas	Limitaciones
Alfa de Cronbach	Mide la consistencia interna de un instrumento al estimar cuán correlacionados están los ítems entre sí (coeficiente α\alpha).	Cronbach (1951)	Sencillo de calcular. Muy utilizado en investigaciones.	Asume unidimensionalidad. Puede subestimar o sobreestimar la fiabilidad real.
Prueba-Reprueba	Administra el mismo instrumento al mismo grupo en dos ocasiones distintas correlacionando los puntajes.	Nunnally (1978)	Útil para evaluar la estabilidad en el tiempo. Interpretación sencilla.	Requiere aplicar la misma prueba dos veces. Puede verse afectado por factores de memoria o maduración.
Formas Equivalentes	Utiliza dos versiones paralelas de un mismo instrumento; se aplican ambas versiones y luego se correlacionan las puntuaciones.	Thorndike (1916)	Disminuye el efecto de la memoria. Asegura equivalencia si las formas están bien diseñadas.	Difícil de desarrollar formas realmente equivalentes. Costo mayor en tiempo y recursos.
Teoría de la Generalizabilidad	Extiende la teoría clásica, analizando diversas fuentes de error (ítems, ocasiones, calificadores, etc.) a través de diseños factoriales.	Cronbach et al. (1972); Cronbach et al. (2004)	Permite un análisis más completo y detallado de la varianza de medición. Identifica fuentes específicas de error.	Puede requerir diseños y análisis estadísticos más complejos. Exige recolección extensa de datos.
Teoría de Respuesta al Ítem (TRI)	Modelo que estima la probabilidad de que un individuo responda correctamente (o en forma positiva) a un ítem, considerando propiedades del ítem y del evaluado.	Rasch (1960); Lord y Novick (1968); Embretson y Reise (2000)	Ofrece información detallada e invariancia de parámetros. Permite estimar la confiabilidad específica por ítem y por niveles del rasgo evaluado.	Supone el ajuste del conjunto de ítems a un modelo matemático complejo. Requiere muestras grandes y software especializado.

Nota. Elaboración propia basada en *Introduction to Classical and Modern Test Theory* (L. Crocker & J. Algina, 2008) y en *Psychometric Theory* (3.ª ed.; J. C. Nunnally & I. H. Bernstein, 1994). Copyright 2008 de Routledge y 1994 de McGraw-Hill.

Es importante mencionar que, aunque un instrumento puede ser confiable sin ser necesariamente válido, la validez requiere necesariamente de confiabilidad para asegurar que las mediciones reflejen adecuadamente las habilidades y conocimientos evaluados (Miller et al., 2009).

Imparcialidad. En cuanto al principio de imparcialidad, se introdujo de manera más explícita y detallada a partir de la edición de los Estándares del año 1999 (AERA et al., 1999), y en línea con la propuesta de Messick (1989), estableciéndose como un principio fundamental junto con validez y confiabilidad. Esta conceptualización fue reforzada y ampliada en la edición de 2014, donde la imparcialidad se establece como una condición esencial para la interpretación válida de los puntajes, particularmente en contextos donde se busca garantizar equidad para todos los examinados, independientemente de sus características personales o contextuales.

De acuerdo con los Estándares (AERA et al., 2014), la imparcialidad exige que ningún individuo o grupo sea sistemáticamente beneficiado o perjudicado por factores irrelevantes al constructo que se pretende medir. En este sentido, la ausencia de imparcialidad constituye una reserva directa a la validez, ya sea por subrepresentación del constructo o por la inclusión de varianza irrelevante, lo cual puede distorsionar las inferencias derivadas de los puntajes.

Para preservar la imparcialidad a lo largo del proceso evaluativo, es indispensable identificar y monitorear de manera sistemática estas reservas. La vigilancia sobre posibles fuentes de sesgo debe formar parte integral del diseño, la aplicación, la puntuación y la interpretación de los instrumentos, asegurando así que las decisiones basadas en los puntajes no estén influenciadas por factores ajenos al constructo evaluado (AERA et al., 2014, pp. 54–57).

Este principio abarca todas las fases del proceso evaluativo —desde el diseño del instrumento hasta la interpretación de los resultados— y se orienta a eliminar o mitigar sesgos asociados con variables como género, origen étnico, idioma, nivel socioeconómico u otras condiciones personales no pertinentes. Para su implementación, se recomiendan prácticas como: (a) revisión experta del contenido para identificar sesgos culturales o lingüísticos; (b) análisis estadísticos de Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés); (c) ajustes en la administración que favorezcan la equidad en las condiciones de aplicación; y (d) desarrollo de criterios interpretativos sensibles a la diversidad del contexto evaluativo.

Limitaciones del proceso de validación actual

A pesar de los avances significativos introducidos en la edición 2014 de los Estándares (AERA et al., 2014), el proceso de validación aún enfrenta limitaciones importantes. Uno de los principales desafíos radica en la falta de lineamientos detallados para operacionalizar el marco argumentativo en contextos prácticos específicos. Si bien el documento enfatiza que la validación implica construir y evaluar argumentos sobre la interpretación de los puntajes de prueba (AERA et al., 2014), no se explicita de manera suficiente cómo estos argumentos deben estructurarse o evidenciarse en escenarios reales de evaluación (Lavery et al., 2020; Durson y Li, 2021). Esta ambigüedad metodológica representa —y ha representado— un obstáculo para la implementación rigurosa del proceso, especialmente para usuarios de pruebas en contextos educativos o clínicos con recursos limitados o experiencia técnica limitada.

Además, el enfoque actual, aunque holístico y articulado, depende en gran medida del juicio profesional para determinar qué tipos de evidencia son pertinentes para una interpretación dada. Esta dependencia puede derivar en prácticas desiguales de validación, especialmente si los usuarios no poseen formación robusta en evaluación psicométrica. Aun cuando los Estándares establecen que toda interpretación propuesta debe estar sustentada por evidencia suficiente y adecuada, no se definen umbrales claros respecto al volumen, tipo o calidad mínima de dicha evidencia, lo que puede generar dudas sobre cuándo una interpretación puede considerarse válidamente respaldada (AERA et al., 2014, pp. 21–22). En consecuencia, y a pesar del paso del tiempo, se sigue requiriendo una mayor sistematización de guías prácticas que permitan traducir el marco teórico-argumentativo en procedimientos de validación replicables, coherentes y transparentes, por lo que el EBA ha representado una forma aceptable de abordarlo (Lavery et al., 2020; Durson y Li, 2021).

Enfoque basado en argumentos

Los principios definidos en los Estándares (AERA et al., 2014)—validez, confiabilidad e imparcialidad—constituyen el fundamento conceptual del EBA, ya que el EBA responde a la parte operativa de la validez, es

decir, es un enfoque que permite abordar el proceso de validación. Estos principios, dentro del proceso de validación, permiten construir interpretaciones justificadas mediante evidencia sólida (validez), asegurar resultados consistentes y replicables (confiabilidad) y garantizar equidad en las decisiones evaluativas (imparcialidad), integrando así dimensiones técnicas, éticas y sociales (Chapelle, 2021).

Michael Kane desarrolló desde los años noventa una propuesta sistemática que establece el EBA como un referente central. Los trabajos iniciales de Kane, como *An Argument-Based Approach to Validation* (1990) y *An Argument-Based Approach to Validity* (1992), así como sus obras posteriores, *The Argument-Based Approach to Validation* (2006) y *Validating the Interpretations and Uses of Test Scores* (2013), sentaron las bases teóricas fundamentales para conceptualizar la validez como un proceso estructurado de construcción y evaluación de argumentos.

Esta perspectiva ha contribuido decisivamente a consolidar la idea, retomada en los Estándares (AERA et al., 2014), de que la validez se refiere no al instrumento en sí, sino a la solidez de las inferencias que se derivan de los puntajes para usos específicos. Esta propuesta metodológica ganó aceptación progresiva en la comunidad académica al abordar explícitamente cómo validar los supuestos e inferencias derivados de los puntajes (Kane y Bridgeman, 2021).

El EBA surge, por tanto, como respuesta directa a la creciente relevancia de considerar no solo la utilidad técnica, sino también las consecuencias sociales y éticas derivadas del uso de las evaluaciones (Cronbach, 1971; Messick, 1989). Mientras Messick enfatizó el constructo evaluado y sus implicaciones éticas y sociales, Kane concentró la discusión en cómo operacionalizar la validación —esto es, traducir los principios de la validez en una investigación articulada por un AIU (inferencias, garantías, supuestos, respaldos y reservas)—, proporcionando un modelo explícito y flexible. Para estructurar este proceso argumentativo, Kane (2006, 2013) recurrió a la lógica argumentativa propuesta por Toulmin (1958/2003), empleada inicialmente en el ámbito jurídico, planteando dos niveles claramente diferenciados:

1. El AIU, que formula las hipótesis sobre la lectura de los puntajes y las decisiones que derivan de ellos.

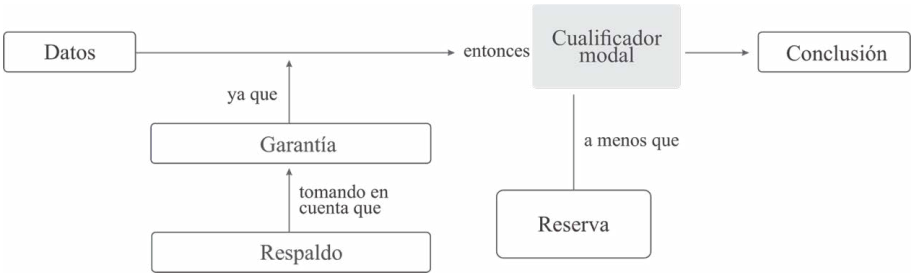
2. El argumento de validez, que analiza la coherencia de cada inferencia y las justificaciones que la respaldan.

Los eslabones lógicos o inferencias deben ser consistentes y plausibles para sostener que una prueba mide lo que se propone (Kane, 2011, 2013), que a la vez considera las evidencias como refutaciones que pudieran surgir.

Modelo de Toulmin

El modelo de Toulmin (1958/2003) (véase la Figura 9) —tradicionalmente compuesto por afirmación, garantía y respaldo empírico— demuestra cómo los datos observados justifican la conclusión, considerando posibles refutaciones. En esta línea, si se afirma "Ana es somalí; por ende, Ana no es católica romana", la garantía se basa en la presunción de que la mayoría de la población somalí practica el islam, y el respaldo estadístico confirma la escasa presencia de somalíes católicos. Con estos elementos, el razonamiento se defiende de manera coherente.

Figura 6.
Modelo de Toulmin



Nota. Adaptado de *The Uses of Argument* [traducción propia] (p. 92), por S. E. Toulmin, 2003, Cambridge University Press (obra original publicada en 1958). Copyright 2003 de Cambridge University Press.

Criterios del EBA según Kane

A partir de esta lógica, Kane (2011) propone tres criterios esenciales para la interpretación argumentativa de las pruebas: (1) la claridad del argumento, que consiste en explicitar garantías y respaldos; (2) la coherencia, encaminada a asegurar la solidez lógica de las inferencias; y (3) la plausibilidad o verosimilitud de cada inferencia, la cual se fundamenta en hipótesis aceptadas o en evidencia empírica. Al aplicar estos criterios al ámbito de la evaluación, se organiza de forma sistemática la recolección de evidencias de validez, la confiabilidad, la pertinencia de los ítems y el impacto de los puntajes en las decisiones, que tiene relación con la imparcialidad (Kane, 2006, 2013). El resultado es la elaboración de cadenas argumentales —también llamadas redes de inferencias— que abarcan desde la descripción técnica de la prueba hasta el análisis de sus consecuencias de uso (Kane, 2015, 2016).

Para articular este proceso, Kane (2006, 2016) plantea cuatro inferencias principales que, al ser verificadas, fortalecen o refutan el argumento de validez: (1) la inferencia de puntuación, (2) la inferencia de generalización, (3) la inferencia de extrapolación y (4) la inferencia de implicaciones. Diferentes autores, entre ellos Cook et al. (2015) y Chapelle (2021), han descrito los métodos habituales para recolectar evidencia en cada una de esas inferencias, coincidiendo, en cierta medida, con Kane.

Estas cuatro inferencias establecieron un tránsito lógico que inicia con la medición puntual de un individuo —mediante reglas específicas de puntuación— hasta las consecuencias derivadas de utilizar ese puntaje en una decisión práctica (Kane, 2013). Cook et al. (2015) ilustran este proceso de forma progresiva: primero se recaban observaciones singulares (por ejemplo, ítems de opción múltiple o tareas de escritura), después se obtiene un puntaje global (generalización), se extrapolan los resultados a escenarios o desempeños futuros (extrapolación) y, finalmente, se valoran las implicaciones de usar ese puntaje para la admisión, la certificación u otras determinaciones.

Las siete inferencias de Chapelle

Chapelle (2021), por su parte, ha ampliado el EBA para aplicarlo, sobre todo, en pruebas de competencia lingüística como el TOEFL. Su esquema incluye siete inferencias, añadiendo pasos para la definición de dominio, la explicación y la utilización de resultados (Figura 7 y Tabla 5). Aunque su propuesta difiere en el número de etapas, mantiene la lógica central: cada inferencia plantea supuestos, precisa evidencias (cuantitativas o cualitativas) y deriva en una conclusión. Así, el concepto de validez implica la construcción y evaluación de argumentos que justifiquen por qué un examen es adecuado para una interpretación específica en un contexto determinado (AERA et al., 2014; Chapelle, 2021). Es importante mencionar que las evidencias de validez, la confiabilidad/precisión e imparcialidad, en el EBA, se encuentran implícitas, pues se pueden vincular las evidencias con cada inferencia, pero sin forzar una correspondencia, ya que los tipos de evidencia.

Tabla 5.
Términos para expresar argumentos de validez de lo general a lo particular

	Estándares	Argumento de validez	Definiciones
General	Interpretación y uso	Interpretación y Uso	Declaración general del propósito de la prueba.
		Significados de Puntuación	Expresiones generales que denotan aspectos del significado.
	Proposiciones (afirmaciones)	Afirmaciones	Enunciados generales sobre interpretación y uso.
		Inferencias	Términos técnicos generales que denotan los pasos en el razonamiento.
		Garantíasa	Enunciados que indican que una inferencia puede autorizarse en un contexto determinado.
		Supuestos	Enunciados que aclaran qué evidencia es necesaria.
Particular	Evidencia	Respaldo	Fragmentos de texto, tablas o figuras en descripciones extendidas de hallazgos.

Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 36), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications. aLas refutaciones son las declaraciones correspondientes a las garantías que indican las condiciones bajo las cuales una inferencia no puede ser autorizada en un contexto particular.

En la Figura 7, Chapelle (2021) esquematiza el Argumento de Validez, integrando propuestas de Messick (1989) y Kane (2006) dentro del modelo lógico de Toulmin (1958/2003). Este esquema muestra siete inferencias encadenadas, en cada una de las cuales se presentan simultáneamente evidencias de validez, confiabilidad e imparcialidad. Estas inferencias se articulan progresivamente, ya que la conclusión de una sirve como sustento para la siguiente. Y son una aproximación a la aplicación en pruebas de inglés; por ende, como afirman los Estándares (AERA et al., 2014), el proceso de validación nunca termina; por lo tanto, podrían existir más inferencias, siempre y cuando se encontraran datos que permitan realizar estas nuevas inferencias.

Figura 7.
Esquema del argumento de validez con las siete inferencias de Chapelle

Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 104), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Por ejemplo, a partir de puntuaciones apropiadas (producto de procedimientos imparciales y mediciones confiables) se sostiene la inferencia de generalización, que establece que los puntajes reflejan consistentemente el desempeño auténtico en el dominio evaluado. A su vez, esta inferencia sustenta las siguientes etapas, hasta llegar al uso práctico de los puntajes en contextos específicos. Sin embargo, el argumento puede enfrentar refutaciones cuando surgen evidencias empíricas sobre sesgos, falta de precisión o inadecuación en las inferencias previas (Chapelle, 2021). Si dichas refutaciones se confirman, el argumento se limita a las poblaciones o contextos para los que se sostiene la validez. De este modo, el Argumento de Validez se entiende como dinámico, abierto al fortalecimiento mediante nuevos datos y revisión constante de sus supuestos.

Así, Chapelle (2021) enfatiza el carácter progresivo de este encadenamiento: la conclusión de una inferencia documentada explícitamente sirve como base lógica para la siguiente. Messick (1989), Kane (2006, 2013) y la propia Chapelle (2021) insisten en que los puntajes tienen sentido solo si están claramente asociados a un uso práctico. Por ello, el argumento final no solo evidencia cómo los puntajes representan el constructo, sino cómo dicha representación se traduce en aplicaciones concretas. En este sentido, el EBA opera como una metodología documental, pues no prescribe métodos específicos, sino que articula la integración lógica de diversas técnicas empíricas, tanto cuantitativas como cualitativas, en un documento que sustenta cada inferencia del Argumento de Validez. En la Tabla 6 se ofrecen ejemplos concretos de investigaciones que ilustran cómo diferentes técnicas pueden aportar evidencia sólida para respaldar cada paso del argumento.

Tabla 6.
Ejemplos cualitativos y cuantitativos según la inferencia

Inferencia en el argumento de validez	Ejemplo de investigación cuantitativa que apoya una suposición	Ejemplo de investigación cualitativa que apoya una suposición
Implicación de consecuencias	Encuesta sobre las opiniones de los profesores acerca del valor de un examen de rendimiento obligatorio para mejorar el aprendizaje de los estudiantes.	Entrevistas con estudiantes sobre sus prácticas de preparación para el examen después de la implementación de un nuevo examen de alto impacto.
Utilización	Estadísticas descriptivas de las puntuaciones del examen mostrando una discriminación aceptable en las puntuaciones de corte propuestas.	Observaciones en el aula de estudiantes ubicados en ciertas clases basadas en las puntuaciones del examen.
Extrapolación	Correlación de las puntuaciones del examen con puntuaciones destinadas a reflejar el rendimiento objetivo.	Análisis del discurso comparando las características lingüísticas de las respuestas construidas por los examinados con su rendimiento en tareas similares en el dominio objetivo.

Explicación	Modelado de ecuaciones estructurales que prueba el papel de los componentes teorizados del constructo.	Relatos retrospectivos de procesos durante la realización del examen recogidos mediante la técnica de "pensar en voz alta".
Generalización	Un estudio G investigando la confiabilidad obtenida con diferentes números de tareas y evaluadores.	Estudio de "pensar en voz alta" sobre los procesos de decisión de los evaluadores mientras califican respuestas construidas en diferentes formas de examen.
Evaluación	Análisis de ítems para calcular dificultad, discriminación y ajuste al modelo.	Estudio observacional de protocolos de seguridad mientras se llevan a cabo en centros de exámenes.
Definición del dominio	Encuesta a expertos en contenido sobre la importancia del contenido prospectivo del examen.	Grupo focal realizado con expertos en contenido para explorar el rango de cobertura de contenido deseado para un examen.

Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 115), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Concebir al EBA como una metodología documental conlleva importantes consecuencias prácticas. En primer lugar, proporciona flexibilidad, dado que permite utilizar una diversidad amplia de métodos empíricos adaptados a distintos contextos de evaluación, siempre y cuando se integren coherentemente en la estructura inferencial del argumento. Además, fomenta la transparencia y la rendición de cuentas, pues la lógica argumental queda documentada explícitamente, facilitando así su comprensión y evaluación crítica tanto por especialistas como por audiencias generales. Asimismo, el carácter acumulativo del documento permite incorporar de manera progresiva nuevos hallazgos investigativos, convirtiéndolo en un registro histórico actualizado —un cuaderno de bitácora o portafolio de evidencias— que respalda la validez del instrumento a lo largo del tiempo.

Para validar una prueba siguiendo este esquema, se inicia definiendo de forma clara el dominio a evaluar —por ejemplo, las habilidades matemáticas que se busca medir—. Luego se verifica la correspondencia entre la prueba y dichas habilidades (evaluación) y se comprueba que las puntuaciones se mantienen estables en distintas circunstancias (generalización). Con la inferencia de Explicación, se dilucida si los puntajes

reflejan realmente el constructo teórico propuesto y, mediante la Extrapolación, se investiga si esos resultados pueden predecir o relacionarse con el desempeño real en contextos externos (por ejemplo, el rendimiento futuro de un estudiante en una materia avanzada). Finalmente, se analiza la forma de utilizar los puntajes (Utilización) y se ponderan las consecuencias derivadas de ello, garantizando la imparcialidad y beneficios para quienes participan en la evaluación (Implicación de Consecuencias).

Recomendaciones para planificar el AIU en pruebas existentes y nuevas

Las guías que se presentan a continuación funcionan como plantillas para planificar un AIU según el momento del examen: si se revisa una prueba ya operativa (como el ExIES) o si se diseña una prueba nueva. Su valor no es "llenar formatos", sino anticipar decisiones, identificar inferencias críticas y alinear —desde el inicio— la evidencia que se requerirá para sostener el Argumento de Validez (Chapelle, 2021; Cook et al., 2015; Kane, 2013).

En evaluación formativa, estas guías pueden adaptarse a escala de aula: el AIU suele centrarse en retroalimentar aprendizajes y ajustar la enseñanza, por lo que las inferencias se formulan con ciclos de evidencia más breves y con criterios que deben revisarse según el contexto, la equidad y la transparencia del uso de resultados (AERA et al., 2014; Brookhart, 2013; Shepard, 2006). En la Tabla 9 se presenta la Guía para planificar un argumento de interpretación/uso sobre una prueba existente. Y, en la Tabla 10, se presenta una guía para planificar un argumento de interpretación/uso sobre una prueba nueva.

Tabla 9.
Guía para planificar un argumento de interpretación/uso sobre una prueba existente

Preguntas de análisis (A)	Acción (B)	Respaldo (C)	Investigación objetivo (E)
1. ¿Se analizó un dominio para crear tareas relevantes para la interpretación y uso de la prueba?	Si no, no hay reclamación ni inferencia de definición de dominio.	¿Qué evidencia tienes de que el dominio fue analizado apropiadamente?	Evaluar la evidencia existente y planear reunir más.
2. ¿La administración y calificación de la prueba afectan las puntuaciones de la prueba?	Si no, no hay reclamación ni inferencia de evaluación.	¿Qué evidencia tienes de que estos factores no han influido inapropiadamente las puntuaciones?	Evaluar la evidencia existente y planear reunir más.
3. ¿Se pretende que las puntuaciones de la prueba reflejen consistencia en el rendimiento?	Si no, no hay reclamación ni inferencia de generalización.	¿Qué estimaciones tienes sobre la magnitud de la inconsistencia para cada fuente?	Evaluar la evidencia existente y planear reunir más.
4. ¿Se ha definido un constructo para servir como base para la interpretación de la puntuación?	Si no, no hay reclamación ni inferencia de explicación.	¿Qué evidencia tienes sobre el constructo que la prueba mide?	Evaluar la evidencia existente y planear reunir más.
5. ¿Has definido el dominio para el cual tus puntuaciones son relevantes?	Si no, no hay reclamación ni inferencia de extrapolación.	¿Qué evidencia tienes sobre cómo las puntuaciones reflejan el rendimiento en el dominio objetivo?	Evaluar la evidencia existente y planear reunir más.
6. ¿Tienes un uso para tus puntuaciones de la prueba?	Si no, no hay reclamación ni inferencia de utilización.	¿Qué evidencia tienes sobre la utilidad de las puntuaciones de la prueba para estos usos?	Evaluar la evidencia existente y planear reunir más.
7. ¿Has identificado los efectos o implicaciones intencionados de tus puntuaciones de la prueba?	Si no, no hay reclamación ni inferencia de consecuencia.	¿Qué evidencia tienes sobre las consecuencias de las puntuaciones de la prueba?	Evaluar la evidencia existente y planear reunir más.

Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 111), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Tabla 10.
Guía para planificar un argumento de interpretación/uso sobre una prueba nueva

Preguntas de análisis (A)	Acción (B)	Garantías, reclamos y supuestos (C)	Investigación objetivo (D)
1. ¿Existe un dominio que deba analizarse para proporcionar insumos para la creación de tareas de prueba relevantes?	Si no, no se hacen reclamaciones sobre la definición del dominio.	¿Qué reclamo harás sobre el dominio? ¿Cuáles son las garantías y suposiciones?	¿Cómo analizarás el dominio para proporcionar respaldo a las suposiciones?
2. ¿La administración y calificación de la prueba afectarán las puntuaciones de la prueba?	Si no, no hay reclamo de evaluación ni inferencia.	¿Qué reclamo harás sobre cómo la puntuación refleja el rendimiento previsto? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás respaldo para las suposiciones?
3. ¿Deberían las puntuaciones de la prueba reflejar la consistencia del rendimiento?	Si no, no hay reclamo de generalización ni inferencia.	¿Qué reclamo harás sobre la consistencia de las puntuaciones de la prueba? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás evidencia sobre la consistencia de la puntuación para respaldar las suposiciones?
4. ¿Un constructo servirá como base para la interpretación de la puntuación?	Si no, no hay reclamo de extrapolación ni inferencia.	¿Qué reclamo harás sobre el constructo que reflejan las puntuaciones? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás evidencia para respaldar las suposiciones sobre el constructo que la prueba pretende medir?
5. ¿Habrá un dominio objetivo que sirva como base para la interpretación de la puntuación?	Si no, no hay reclamo de extrapolación ni inferencia.	¿Qué reclamo harás sobre el rendimiento objetivo que reflejan las puntuaciones de la prueba? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás apoyo para las suposiciones sobre cómo las puntuaciones reflejan el rendimiento en el dominio objetivo?

6. ¿Para qué se utilizarán las puntuaciones de la prueba?	Si no, no hay reclamo de utilización ni inferencia.	¿Qué reclamo harás sobre la utilidad de las puntuaciones de la prueba? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás respaldo para las suposiciones sobre la utilidad de las puntuaciones de la prueba?
7. ¿Cuáles son los efectos previstos de la prueba?	Si no, no hay reclamo de consecuencia ni inferencia.	¿Qué reclamo harás sobre las consecuencias previstas de las puntuaciones de la prueba? ¿Cuáles son las garantías y suposiciones?	¿Cómo proporcionarás respaldo para las suposiciones sobre la utilidad de las puntuaciones de la prueba?

Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 113), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Capítulo 2

Diseño metodológico

Este capítulo describe la ruta metodológica seguida para aplicar el EBA en la revisión del ExIES 2023-2. La ruta se ancló en el objetivo de construir un argumento de validez, con reservas y recomendaciones, a partir de evidencia documental y resultados técnicos disponibles. Además, se basa en antecedentes, una Revisión Sistemática de la Literatura, ya publicados sobre el estado actual del EBA (véase Ruiz Mendoza et al., 2025).

El enfoque de este trabajo es evaluativo y documental: recupera fuentes existentes (técnicas, normativas e institucionales), las organiza en un AIU y, después, somete cada inferencia a escrutinio mediante supuestos y respaldos (Kane, 2006, 2013; Chapelle, 2021). Con fines de guía, el capítulo se organiza en cuatro movimientos: 1) procedimiento y criterios de evaluación; 2) identificación de fuentes de datos; 3) descripción del instrumento; y 4) formulación del AIU y su traducción a inferencias.

Procedimiento

El procedimiento se organizó conforme a la lógica del EBA. Para la elaboración de este proceso de validación se retomaron los tres pasos de Kane (2006, 2013, 2020): 1) formulación del AIU; 2) evaluación crítica del argumento; 3) conclusión global sobre la validez; y, además, se retomaron elementos propuestos por Chapelle (2021) como la etapa 7; las etapas 2, 3, 4 y 5 se plantean para facilitar el seguimiento al procedimiento realizado. Estas siete etapas articuladas fueron diseñadas para cumplir funciones específicas dentro del argumento interpretativo, integrando tanto datos cuantitativos como cualitativos.

En términos metodológicos, las evidencias no se interpretaron en el vacío: se leyeron a la luz de teorías de medición y evaluación. Por ejemplo, la definición de dominio se apoyó en criterios de validez de contenido y representatividad (Kane, 2006; Sireci, 1998; Lawshe, 1975); la generalización se respaldó en teoría clásica, teoría de la generalizabilidad e IRT/TRI (Lord y Novick, 1968; Cronbach et al., 1972; Brennan, 2006; Rasch, 1960; Bond & Fox, 2015); y la explicación y extrapolación se vincularon con la validez de constructo, la estructura interna y las relaciones con variables externas (Cronbach y Meehl, 1955; Messick, 1989; Hu y Bentler, 1999). Esta conexión teoría-evidencia permitió construir

un argumento de validez, en lugar de una lista de resultados (Kane, 2013; Chapelle, 2021). Así, y con el fin de poder replicarlo, a continuación, se presentan las etapas del procedimiento, como sus fases y productos.

Etapa 1: Definir el AIU

La intención de esta primera etapa fue precisar el uso e interpretación que se le daría al ExIES, entendiendo que ya existía un antecedente del objetivo del instrumento.

Fase 1. En esta fase se estableció la población objetivo, el contexto de aplicación del instrumento y las consecuencias previstas del uso del ExIES.

Fase 2. Se formuló la conclusión general sobre la interpretación y los usos de las puntuaciones, es decir, a partir del contexto y la necesidad, se estableció que el ExIES se utiliza para el ingreso a la universidad.

Fase 3. Una vez establecida la afirmación principal, se procedió a establecer las conclusiones específicas de cada inferencia: definición del dominio, evaluación, generalización, explicación, extrapolación, utilización e implicación de consecuencias. Estas elaboraciones partieron de las recomendaciones realizadas por Chapelle (2021), lo cual conllevó un proceso de reflexión continuo.

Producto. El producto central fue una adaptación del modelo argumentativo de Toulmin (1958/2003) al contexto del EBA según Chapelle (2021).

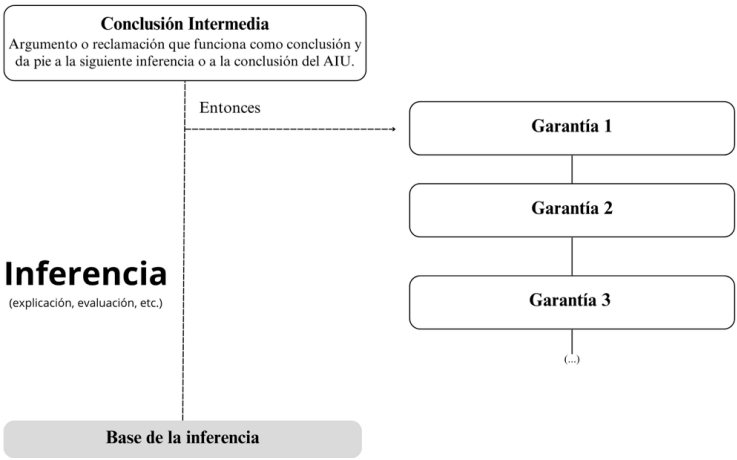
Etapa 2. Formular garantías y supuestos e identificar fuentes de datos

En la segunda etapa se especificaron las evidencias necesarias para respaldar cada inferencia planteada. Así, se formulan las garantías para después alinear los supuestos que, a su vez, relacionan lógicamente con las inferencias; es decir, se parte de seleccionar las fuentes posibles que fundamenten cada supuesto lo que permite la concatenación de argumentos en menor o mayor medida.

Fase 1. La Figura 8 muestra este primer paso de identificación de garantías; estas pueden ser diversas según la argumentación. Además, se

establecieron las conclusiones intermedias hasta llegar a la conclusión del argumento; en este caso se seleccionaron todas las inferencias; la última natural es la implicación de consecuencias.

Figura 8.
Esqueleto del argumento por inferencia como parte de la validez del argumento



Fase 2. Además, se identificaron diversas fuentes potenciales de información, incluyendo estudios cuantitativos, cualitativos y documentos normativos institucionales. Cada fuente fue relacionada por el supuesto correspondiente. En este sentido, el criterio fue revisar todo lo disponible y, en caso de ausencia, revisar la posibilidad de obtener las fuentes necesarias para el desarrollo de evidencias. De este acercamiento surgió la necesidad de obtener las fuentes de datos necesarias para la inferencia de extrapolación.

Para esta fase, donde el fin es plantearse preguntas clave para comprender qué es lo que se tiene y dónde colocarlo, es útil tomar como referencia la Tabla 6, donde se dan los ejemplos de investigaciones cualitativas o cuantitativas por inferencia como guía para identificar las evidencias.

Fase 3. En esta fase se identificaron las fuentes de datos a desarrollar; en el caso del ExIES fue la evidencia de validez predictiva, por lo que se procedió a preparar el estudio a partir de las bases de datos seleccionadas. En este caso, también aplican las tablas mencionadas en la Fase 2.

Producto. El producto resultante fue una tabla completa con inferencias, garantías, supuestos y fuentes de datos. Por otra parte, un ejemplo de la estructura se muestra en la Tabla 7.

Tabla 7.
Esqueleto de la estructura argumentativa con fuentes de datos del ExIES

Conclusión	Descripción de la conclusión de la inferencia.		
Inferencia	**Garantía**	**Suposiciones**	**Fuentes de datos**
Definición de dominio	G1.1. Descripción de la garantía	S.1.1.1 Descripción del supuesto	F1.1.1 Descripción del dato
Descripción del argumento			
(…)	(…)	(…)	(…)

Etapa 3. Desarrollo de fuentes de datos

Fase 1. La tercera etapa tuvo un fin muy claro, desarrollar la fuente faltante a partir de otras fuentes de datos, por lo que implicó una fase, la elaboración o desarrollo del estudio, guía, manual, reporte, etcétera.

Producto. Estudio de extrapolación.

Etapa 4. Desarrollo de respaldos

La cuarta etapa consistió en analizar cada fuente de información en relación con los supuestos e inferencias definidos anteriormente.

Fase 1. Se analizaron las fuentes de datos asociadas a cada supuesto y, por ende, por inferencia.

Fase 2. Después se procedió a describir cada uno de los respaldos (según el supuesto) que fundamentan empíricamente cada inferencia, alineándolos con sus garantías.

Producto. El producto consistió en la descripción estructurada de los respaldos que fortalecen la solidez argumentativa del instrumento, facilitando la comprensión y análisis de cada inferencia, es decir, la descripción de los resultados según cada supuesto de la evidencia pertinente mediante descripciones empíricas y teóricas, así como con el uso de figuras y tablas.

Etapa 5. Valoración por inferencia

Esta etapa tiene como fin valorar las evidencias que sostienen cada inferencia, según sus garantías y supuestos.

Fase 1. Por lo anterior, la primera fase correspondió a construir el criterio de esta evaluación. Así, y de acuerdo con los Estándares (AERA et al., 2014), la validez no constituye una propiedad absoluta de una prueba, sino que se sustenta en un conjunto articulado de evidencias destinadas a justificar las interpretaciones y usos previstos de los puntajes obtenidos. Según Chapelle (2021), estas evidencias corresponden a declaraciones fundamentadas en respaldos derivados de diversas fuentes de datos, complementadas con bases teóricas o normativas evaluadas por expertos.

Con el fin de construir un argumento de validez claro, coherente y plausible, se propone explicitar criterios de evaluación que mejoren la legitimidad y precisión del AIU y, sobre todo, orienten la reflexión sobre las conclusiones, en especial cuando faltan teorías o referentes empíricos específicos. Ahora bien, estos criterios propuestos no sustituyen el juicio cualitativo expresado por la AERA et al. (2014): lo estructuran y transparentan. Funcionan, en todo caso, como un proceso de autoevaluación: guías para organizar la evidencia, hacer visibles los supuestos y vacíos, y priorizar los análisis necesarios en el proceso de mejora continua; sobre todo para un instrumento tan joven. El puntaje resultante es un indicador sintético para comparar y dar seguimiento; no es una medida de validez. Más que alcanzar la puntuación máxima, el objetivo es documentar avances y focalizar mejoras donde el argumento es más vulnerable para proseguir con el proceso de mejora continua.

Definición de la escala de evaluación. En los últimos años, diferentes disciplinas han desarrollado sistemas de categorización para valorar la solidez de la evidencia y orientar la toma de decisiones. En el campo de la medicina, por ejemplo, emplean la Clasificación de la Calidad de la Evidencia y Graduación de la Fuerza de las Recomendaciones (GRADE, por sus siglas en inglés) como método para clasificar la calidad de la evidencia y la fortaleza de las recomendaciones (Guyatt et al., 2011). De manera análoga, en contextos de investigación cualitativa, se ha adoptado el sistema de Confianza en la Evidencia procedente de Revisiones de

Investigación Cualitativa (CERQual, por sus siglas en inglés), centrado en valorar el grado de confianza en la evidencia derivada de síntesis cualitativas (Lewin et al., 2018).

Tanto GRADE como CERQual coinciden en emplear cuatro niveles (Alta, Moderada, Baja, Muy Baja) para evaluar la calidad o credibilidad de la evidencia, al tiempo que resaltan la consistencia, la coherencia y la pertinencia de los datos. Estos principios resultan altamente valiosos al diseñar procedimientos de evaluación en el ámbito educativo, pues brindan un marco de referencia para clasificar los hallazgos o interpretaciones en función de su solidez. A partir del EBA (Kane, 2013; Chapelle, 2021), se contempla una serie de inferencias que vinculan la forma en que se interpretan y utilizan los resultados de la prueba con los supuestos teóricos que las sostienen. Para valorar estas inferencias —ante la ausencia de propuestas específicas— se plantea examinarlas de manera sistemática, considerando tres criterios fundamentales propuestos por Kane (2013) y precedidos por Chapelle (2021):

- Claridad en la formulación de las relaciones teóricas y empíricas que enuncian los supuestos.
- Coherencia interna de los argumentos (garantías, datos y conclusiones).
- Plausibilidad, entendida como la calidad y pertinencia de los datos (tanto teórica como metodológica).

Sin embargo, para estos criterios se propone evaluar cada evidencia según su supuesto (por ende, garantías e inferencias) a través de estos tres criterios. Esto significa que la valoración se aplica a las evidencias, lo que permite hacer una valoración global de forma detallada; a cada pieza de evidencia se le evalúan los mismos criterios: claridad (¿está la evidencia descrita y documentada con precisión?), coherencia (¿es internamente consistente y congruente con otras evidencias?) y plausibilidad (¿es metodológica y teóricamente creíble para sostener la inferencia?).

Por ende, para integrar los tres criterios (claridad, coherencia y plausibilidad), se propone sumar las puntuaciones parciales, generándose un índice global por inferencia que oscila entre 3 y 12 puntos. Tal como se presenta en las escalas de evaluación de GRADE y CERQual, se definen cuatro categorías cualitativas, como se observa en la Tabla 8.

Estos criterios se enfocan en la valoración de la evidencia, es decir, los hallazgos que apoyan al supuesto.

Tabla 8.
Criterios EBA para la evaluación de las evidencias de las inferencias

Criterio	Definición	Puntuación (1 a 4)
Claridad	Indica en qué medida la evidencia del supuesto se enuncia de forma precisa, comprensible y específica, evitando ambigüedades o vacíos conceptuales.	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta
Coherencia	Valora la consistencia interna de la evidencia, verificando que no existan contradicciones.	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta
Plausibilidad	Mide la credibilidad o fundamentación teórica de la evidencia, es decir, si las evidencias respaldan razonablemente el uso de los puntajes o conclusiones.	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta

Los criterios se basan en la comparación sintética de los niveles de calificación de la evidencia en GRADE, CERQual y en la escala del EBA propuesta. Aunque cada metodología surge en un contexto distinto (investigación clínica, revisiones de evidencia cualitativa y evaluación educativa, respectivamente), todas comparten la clasificación en cuatro niveles y el énfasis en la consistencia y la solidez de la evidencia. Esta es solo una sugerencia de uso para la autoevaluación a considerar.

Relación con los estándares (AERA, APA y NCME, 2014). Con el propósito de realizar una evaluación plausible de las inferencias planteadas en el modelo EBA, se realizó una vinculación explícita entre dichas inferencias (Chapelle, 2021), los estándares propuestos por AERA, APA y NCME (2014), y relacionándolos con los tipos de evidencia de validez. Así, esta integración proporciona una estructura para analizar y evaluar de manera fundamentada cada inferencia, asegurando que los criterios empleados sean conceptualmente coherentes, metodológicamente sólidos y éticamente apropiados. A continuación, la Tabla 9 detalla dicha vinculación, acompañada de una breve justificación fundamentada a través de Chapelle (2021) y los Estándares (AERA et al., 2014).

Tabla 9.
Selección de estándares por inferencia y tipos de evidencias de validez

Inferencia	Estándares principales	Tipo de evidencia de validez	Justificación
Definición de dominio	1.0, 1.2, 4.1, 11.13, 11.14	Basada en el contenido.	Asegura que la prueba refleja con fidelidad el dominio relevante al constructo. Fundamenta la interpretación válida de los puntajes.
Evaluación	6.1–6.5, 7.2, 8.1–8.2, 9.3	Basada en procedimientos de administración y puntuación.	Verifica la calidad técnica de la aplicación y corrección, garantizando procedimientos justos, consistentes y replicables.
Generalización	2.1–2.11, 4.10, 5.2	Basada en confiabilidad/precisión.	Confirma que los puntajes son estables y reproducibles en formas, ocasiones o contextos comparables.
Explicación	1.13–1.16, 1.0, 1.1, 3.2, 5.1	Basada en estructura interna, procesos de respuesta y relaciones con otras variables.	Fundamenta que los puntajes reflejan el constructo pretendido, conforme a teorías y modelos empíricamente respaldados.
Extrapolación	1.19–1.21, 5.1	Basada en relaciones con otras variables (criterios externos).	Evidencia que los puntajes predicen o reflejan el desempeño en contextos o tareas fuera del entorno de prueba.
Utilización	12.1, 12.2, 11.4	Evidencia sobre uso e interpretación para decisiones.	Verifica que los puntajes se usen adecuadamente para decisiones específicas, con implicaciones prácticas y éticas.
Implicación de consecuencias	6.10, 13.1, 13.6	Basada en consecuencias del uso de la prueba.	Evalúa los efectos sociales, educativos o psicológicos del uso de la prueba, tanto previstos como no previstos.

Nota. Elaboración propia basada en *Argument-Based Validation in Testing and Assessment* (C. A. Chapelle, 2021) y en *Standards for Educational and Psychological Testing* (AERA, et al., 2014).

Fase 2. Una vez determinados los criterios, se procedió a valorar cada uno de los supuestos, por ende, sus evidencias, lo que implicó un análisis teórico como empírico de las evidencias.

Fase 3. Por último, se procede a valorar la inferencia de forma global en cuanto a su claridad, coherencia y plausibilidad, describiendo el porqué de la obtención de dicho resultado.

Producto. Para ilustrar la aplicación concreta de los criterios, la Tabla 10 presenta una plantilla diseñada (producto) para expresar la evaluación de cada inferencia según los tres criterios especificados (claridad, coherencia y plausibilidad).

Tabla 10.
Estructura de las escalas para la evaluación de inferencias

Supuestos a evaluar	Estándares para evaluar	Claridad	Coherencia	Plausibilidad	Puntaje global
Listar aquí los supuestos relacionados con las garantías.	Listar aquí los estándares relevantes (p. ej., Estándar 1.1, 4.7, etc.)	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta	1 = Muy baja 2 = Baja 3 = Moderada 4 = Alta	Suma

En la última columna de la Tabla 10, se calculó un índice global de la inferencia cuyo rango (3–5, 6–8, 9–10 o 11–12) determina la categoría cualitativa final (Muy Baja, Baja, Moderada o Alta); además, se obtiene el porcentaje (%) sumando el resultado de los criterios, dividiéndolo entre doce y multiplicando por cien. Así, para calcular el Porcentaje de Validez por Supuesto o Inferencia (LV), se utilizan fórmulas que comparan la suma de los puntos obtenidos frente a la suma de los puntos máximos posibles:

En esta expresión:

- Representa la suma de puntos logrados en cada uno de los criterios.
- Es la suma de los puntos máximos posibles.
- LV indica el porcentaje total de validez alcanzado por ese supuesto o inferencia.

Etapa 6. Valorar el argumento global

La etapa seis implicó integrar las evaluaciones individuales de cada inferencia en un argumento global, la cual se refiere a la valoración del argumento como unidad según Kane (2006, 2013).

Fase 1. Se encadenaron las valoraciones de la escala EBA de las inferencias en un argumento global, es decir, se reunieron las valoraciones de todas las inferencias por criterio, realizando promedios; para ello, véase el ejemplo de la Tabla 11.

Tabla 11.
Esqueleto para los resultados del argumento global

Inferencia	Claridad	Coherencia	Plausibilidad	Global	Puntaje	Interpretación
Definición de dominio	LV (LV%)	n/ N (LV%)	n/ N (LV%)	n/ N (LV%)	P	Alta
Implicación de consecuencias	(LV%)	(LV%)	(%)	n/ N (%)	P	Baja
Global	(LVg%)	(LVg%)	(LVg%)	n/ N (LV g%)	Pg	Moderada

Nota. Donde n=suma total obtenida por criterio; N=suma de los puntos máximos posibles; LV representa la fórmula del porcentaje de validez; P=(n/N) x12, el cual representa el criterio de puntuación.

Fase 2. Se realizó un análisis e interpretación de los resultados de la evaluación del Argumento Global a partir de la Tabla 11. Por ende, con el fin de interpretar el LV global (LVg), en la Tabla 12 se presenta una escala para valorar los puntajes de la prueba.

La selección de los cohortes propuestos (25.0–37.5%, 37.5–62.5 %, 62.5–87.5 % y >87.5–100 %) se fundamenta en tres razones metodológicas complementarias: primero, preserva la coherencia con la rúbrica

ordinal subyacente (criterios valorados 1–4), situando los puntos de corte en los puntos medios entre las categorías adyacentes (1.5, 2.5, 3.5), lo que en términos porcentuales equivale a 37.5 %, 62.5 % y 87.5 %; esta decisión permite minimizar arbitrariedades y respeta la interpretación cualitativa original de la escala (Brookhart, 2013). Segundo, respeta las limitaciones imposibles de la métrica sumativa: dado que el rango posible por suposición va de 3 a 12 (mínimo 25 %), el umbral inferior inicia en 25 % preservando la correspondencia entre puntaje bruto y LV (%), y los cortes centrales reflejan promedios por criterio que tienen sentido operativo (p. ej., que Moderada implique, en promedio, ≥2.5 por criterio). Tercero, la elección mantiene la prudencia exigida por enfoques de validación: reservar la categoría superior para valores cercanos al máximo (≥87.5 %) sigue la recomendación del EBA de no conceder una alta confianza salvo cuando la evidencia sea consistentemente robusta (Chapelle, 2021; Kane, 2013).

Tabla 12.
Escala para la interpretación del porcentaje de validez sobre los puntajes

Regla en %	Nivel	Interpretación
25 % – 37.5 %	Muy baja	La evidencia es muy limitada o incluso contradictoria. Se recomienda recabar más datos antes de utilizar los puntajes en decisiones de alta relevancia.
37.5 % – 62.5 %	Baja	Aunque se identifican vacíos, la evidencia resulta moderada. Se sugiere utilizar los puntajes con cautela o en conjunto con otras fuentes de datos.
62.5 % – 87.5 %	Moderada	La evidencia es mayormente sólida; los puntajes se consideran adecuados para la mayoría de los usos previstos, aunque se aconseja un seguimiento continuo.
> 87.5 % – 100%	Alta	La evidencia es coherente y prácticamente libre de contradicciones. Existe un elevado grado de confianza en la validez de la inferencia.

Fase 3. Finalmente, se analizan las reservas, por supuesto, y se elaboran recomendaciones con bases teóricas y empíricas que ayuden a sostener el argumento, que posteriormente son revisadas y valoradas por el equipo del ExIES para establecer un seguimiento a las mismas.

Producto. Se obtuvieron dos productos: la tabla del resultado del Argumento Global y una tabla que integra el supuesto, la evidencia, su recomendación y reservas; la Tabla 13 es un esqueleto de recomendaciones por inferencia que funciona para la mejora continua; este se expresa en el apartado de *Discusión y Conclusiones* del presente trabajo.

Tabla 13.
Esqueleto para el resumen de las evidencias y recomendaciones por supuesto

Suposición	Evidencia	Recomendaciones	Reservas
S1.	Descripción de la evidencia con datos o afirmaciones: qué, cómo y con base en qué teoría, normativa o documento.	Descripción de la recomendación general, por supuesto.	Descripción de la amenaza posible: hipotético.
S2.	…	…	…

Etapa 7. Ciclo iterativo: comunicación y mejora

Esta última etapa expresa la iteración de retroalimentación en las etapas anteriores, a partir de la comunicación constante tanto con el responsable del proyecto como con el equipo técnico.

Fase 1. Se presentan los resultados finales para acordar los siguientes pasos en la mejora continua del instrumento como de las interpretaciones de los puntajes. En este caso se presentaron los resultados finales y se obtuvo retroalimentación oportuna para realizar ajustes necesarios, como la revisión de resultados inexactos. Al ser una etapa iterativa, se espera que no solo sea útil para el equipo del ExIES, sino para los tomadores de decisiones. Además, es importante aclarar que el producto final de esta etapa son las recomendaciones sintetizadas posteriores a la evaluación mediante la escala propuesta.

Fuentes de datos

Las fuentes de datos que aquí se presentan fueron todas las disponibles para el proceso de validación; es importante señalar que estas fuentes son indispensables para el desarrollo de la evidencia según cada supuesto, y así poder evaluarlas. La mayoría de las fuentes de datos fueron desarrolladas y proporcionadas principalmente por el equipo técnico del ExIES: los manuales técnicos, guías específicas para evaluadores y sustentantes, especificaciones de elaboración y jueceo de ítems, así como reportes técnicos, bases de datos derivadas del proceso psicométrico históricos. Por otro lado, también se obtuvieron las bases de datos de los promedios de EMS y primer año universitario de la UABC para el estudio de la inferencia de extrapolación, el cual fue el único de elaboración propia. Las fuentes anteriores fueron obtenidas entre los ciclos 2023-2, 2024-1 y 2025-1. En la Tabla 14 se presentan las fuentes específicas clasificadas por tipo para el proceso de validación del ExIES, donde se señala el año de publicación que corresponde al de obtención. Asimismo, los documentos institucionales y normativos, como la Ley Orgánica, Estatuto General y Escolar de la UABC, fueron obtenidos directamente de fuentes en línea oficiales.

Tabla 14.
Tipo de fuentes utilizadas para el proceso de validación del ExIES

Tipo de fuente	Fuentes específicas
Manuales técnicos	Manual Técnico del Nuevo Examen de Selección (Caso et al., 2017); Manual Técnico del ExIES (Pedroza Zúñiga et al., 2022).
Guías y manuales específicos	Guía para la Evaluación de ítems del Nuevo Examen de Selección de aspirantes a ingresar a la Universidad Autónoma de Baja California (Caso y Díaz, 2016); Guía de estudios para el sustentante (Pedroza Zúñiga et al., 2023l).
Especificaciones	Especificaciones para la elaboración de ítems de Lectura, Lengua Escrita y Matemáticas (Pedroza Zúñiga et al., 2023n, 2023o, 2023p).

Manuales de desarrollo y jueceo de reactivos	Manual para el desarrollo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023a, 2023b, 2023c); manual para el jueceo de reactivos de lectura, lengua escrita y matemáticas del ExIES (Pedroza Zúñiga et al., 2023d, 2023e, 2023f).
Bases de datos	Base de datos completa de resultados Rasch y estadísticas ítem–forma (2023s); base de datos de los jueceos (Pedroza Zúñiga et al., 2023q); base de datos de organización de ítems, histórico del ExIES: control de ítems NDC-especificación-contenido (Pedroza Zúñiga et al., 2024c); base de datos del promedio del primer y segundo semestre de universidad (UABC, 2024); base de datos de promedios por alumno de bachillerato (EMS BC, 2024).
Reportes	Reporte Técnico 2023-1, 2023-2 (Pedroza Zúñiga et al., 2024a, 2024b); Reporte de aplicación del ExIES 2023-1 (Pedroza Zúñiga et al., 2023r); reporte sobre el funcionamiento diferencial del ítem (DIF) por sexo del ExIES 2023-2 (Pedroza Zúñiga y Gómez Monárrez, 2025c).
Documento de resultados	Resultados de validez concurrente con los puntajes de EXANI II de los informes particulares y generales del ExIES (Pedroza Zúñiga y Gómez Monárrez, 2025a, 2025b).
Documentos institucionales y normativos	Ley Orgánica UABC (2010); estatuto general UABC (2019); estatuto escolar UABC, Artículos 16, 18, 24 (UABC, 2021).
Otros documentos y recursos	Presentaciones de capacitaciones del ExIES (Pedroza Zúñiga et al., 2023h, 2023k); protocolos para incidencias (Pedroza Zúñiga et al., 2023m); tabla comparativa de cambios por área (elaboración propia); estudio de validez predictiva (elaboración propia).

Instrumento del objeto de estudio: ExIES

El ExIES fue desarrollado considerando las competencias disciplinares básicas y extendidas según el Marco Curricular Común de la EMS, cuyo uso se delimitó para el ingreso a la educación superior (Pedroza Zúñiga et al., 2024a). Este examen "(...) mide la capacidad que tienen los aspirantes para aplicar los conocimientos y habilidades que poseen y serán requeridos para atender con éxito las demandas propias de su

formación universitaria." (Caso-Niebla et al., 2017, p. 15), que busca evaluar tres áreas. Asimismo, en la Tabla 15 del ExIES se pueden observar las especificaciones.

- Lectura. Evalúa la capacidad para leer y comprender un amplio rango de textos literarios e informativos. Los temas y preguntas sobre los textos se enfocan en la elaboración de conexiones y comprensión individual de los textos, así como la interpretación y síntesis de información e ideas en textos con gráficos.
- Lengua escrita. Evalúa la capacidad para revisar y editar una amplia variedad de textos con contenido de naturaleza académica, además de medir la capacidad para expresar ideas en apego a las convenciones del español escrito. Los textos y preguntas se enfocan en la toma de decisiones de edición y revisión de los textos, y reconocimiento e identificación de errores de gramática, uso y puntuación, relacionados con el contexto de los textos.
- Matemáticas. Mide la capacidad para la aplicación, manejo y comprensión de conceptos matemáticos, y la habilidad para la resolución de problemas e interpretación de datos, tablas, planos, figuras y gráficos. Las preguntas se enfocan en la demostración de habilidades para la aplicación de procedimientos, comprensión profunda de conceptos matemáticos y resolución de problemas en amplia variedad de contextos.

Tabla 15.
Especificación de contenidos en el ExIES

Componente	Contenido	Descripción	Proporción
Lectura	Información e ideas	Evaluación del contenido informativo del texto.	40 %
	Formas discursivas	Análisis estructural del discurso.	40 %
	Intertextualidad	Síntesis de múltiples fuentes de información.	20 %
Lengua escrita	Expresión e ideas	Revisión del desarrollo del tema, precisión, lógica, cohesión y uso efectivo del lenguaje en el texto.	50 %
	Cumplimiento de reglas del español escrito	Edición de un texto para asegurar su conformidad con las convenciones gramaticales del español, estructura de oraciones, uso y puntuación.	50 %
Matemáticas	Herramientas algebraicas	Resolución de problemas mediante el empleo de ecuaciones y sistemas lineales, ya sea a través de la representación de cantidades o de la representación gráfica.	20 %
	Problemas, probabilidad y análisis de datos	Creación y análisis de relaciones, representación y análisis de datos cuantitativos y aplicación de probabilidades.	30 %
	Matemáticas avanzadas	Creación de expresiones algebraicas y uso de gráficos que representan funciones exponenciales no lineales o cuadráticas.	30 %
	Temas adicionales en matemáticas	Solución de problemas asociados al área y volumen, aplicación de definiciones, teoremas sobre líneas, ángulos, triángulos y círculos.	20 %

Nota. Adaptado de *Reporte técnico del ExIES 2023-1* (p. 10), por Pedroza Zúñiga et al., 2024a.

Por otro lado, la Tabla 16 detalla los Niveles de Demanda Cognitiva (NDC) del ExIES que ayudan a comprender el peso porcentual por área.

Tabla 16.
Especificación de niveles de demanda cognitiva (NDC) ExIES

NDC	Descripción	Lectura	Lengua Escrita	Matemáticas	Totales
Comprensión	Capacidad de comprender el significado del material.	-	-	8 %	8 %
Aplicación	Capacidad de utilizar el material aprendido en situaciones nuevas y concretas.	9 %	15 %	22 %	46 %
Análisis	Capacidad de descomponer el material en sus partes o componentes de manera que la organización de la estructura pueda ser entendida.	-	-	5 %	5 %
Síntesis	Capacidad de acomodar las partes entre sí para formar un nuevo conjunto.	5 %	-	6 %	11 %
Evaluación	Capacidad de juzgar el valor del material (declaración, novela, poema, informe de investigación para un propósito determinado.	15 %	15 %	-	30 %
	Totales	29 %	30 %	41 %	100 %

Nota. Adaptado de *Reporte técnico del ExIES 2023-1* (p. 11), por Pedroza Zúñiga et al., 2024a.

Especificaciones del instrumento

El ExIES se compone de ítems de opción múltiple que incluyen un enunciado y cuatro opciones de respuesta (una correcta y tres distractores). En su versión 2023-1 (Pedroza Zúñiga et al., 2024a), véase Tabla 17, se

compuso de dos formas (Forma A y Forma B), integradas por un total de 122 ítems distribuidos en tres áreas: 36 de lectura, 36 de lengua escrita y 50 de matemáticas. Además, en la aplicación 2023-1 se incluyeron 38 ítems piloto adicionales por forma (14 en lectura, 14 en lengua escrita y 10 en matemáticas), generando así diez subversiones con un total de 160 ítems cada una, con el propósito de mantener actualizado y robusto el banco de ítems general.

Tabla 17.
Componentes del ExIES 2023-1

Componente del ExIES	Descripción
Tipo de ítems	Preguntas de opción múltiple (4 opciones, 1 correcta y 3 distractores)
Estructura	122 ítems: 36 lectura, 36 lengua escrita, 50 matemáticas
Ítems ancla (30 %)	Comunes a todas las formas para equiparación de resultados
Ítems diferenciadores (70 %)	Distintos por forma, pero equivalentes en dificultad y contenido
Ítems piloto	Nuevos ítems en prueba para evaluar su idoneidad técnica
Subversiones (2023-1)	5 por forma consolidada, cada una con 160 ítems totales (122 base + 38 piloto)
Adaptación especial	Versión adicional adaptada para estudiantes con discapacidad visual (instrucciones específicas, ítems modificados, fuente aumentada)

Nota. Elaboración propia basado en *Reporte técnico del ExIES 2023-1*, por Pedroza Zúñiga et al. (2024a).

Participantes en la elaboración del instrumento

El proceso de elaboración del ExIES reúne a cinco tipos de participantes con funciones diferenciadas (véase Tabla 18). Primero, se encuentra el responsable del proyecto, quien presenta los resultados a las autoridades correspondientes, y quien también asume la coordinación de las mejoras a desarrollar. El segundo tipo es el equipo técnico del ExIES, quienes

se encargan de elaborar los manuales, especificaciones y reportes históricos, además de sostener la logística operativa del ExIES. El tercer y cuarto tipo de participantes son los responsables de la elaboración y supervisión de los ítems; entre el ciclo 2022-1 y 2023-1, hubo un total de 26 diseñadores y 20 jueces. Por último, los participantes finales, es decir, los aspirantes; la aplicación del ExIES, en el ciclo 2023-1, tuvo un total de 28 205 aspirantes evaluados. A continuación, se detalla cada uno de los tipos de participantes.

Tabla 18.
Participantes, funciones y relación con las inferencias del EBA

Tipo de participantes	n	Función principal	Actividades clave
Responsable del proyecto	1	Usuario clave y enlace institucional.	Retroalimentar hallazgos; coordinar mejoras y difusión.
Equipo técnico del ExIES	4	Generar documentación y bases históricas; apoyo técnico y logístico.	Coordinación operativa, coordinación de áreas disciplinares; manuales, especificaciones, reportes, entre otros documentos.
Diseñadores/as de ítems (2022-1 a 2023-1)	26	Elaborar reactivos alineados al dominio.	Redacción de ítems, ajuste tras retroalimentación.
Jueces de contenido (2022-1 a 2023-1)	20	Revisar calidad y pertinencia de los ítems.	Revisión ciega, veredicto de aceptación/rechazo.
Aspirantes de la convocatoria 2023-1	28, 205	Suministrar datos de desempeño.	Resolución del ExIES.
Aspirantes de la convocatoria 2023-2	2,291	Suministrar datos de desempeño.	Resolución del ExIES.
Aspirantes de la convocatoria 2022-2	1,937	Suministrar datos de desempeño.	Resolución del ExIES.

Nota. Elaboración propia y retomando usuarios del *examen de ingreso a la educación superior (ExIES) 2023-1*: Reporte técnico (p. 10), por Pedroza Zúñiga et al., 2024a.

Equipo ExIES. El equipo, véase Figura 9, está conformado por cinco integrantes.

El equipo se conforma por un responsable del proyecto y cuatro asistentes de investigación distribuidos en análisis de datos, coordinación operativa, una coordinación de comprensión lectora y lengua escrita, una coordinación de matemáticas, así como una persona como asesoría externa.

Diseñadores y jueces. Sobre el diseño y jueceo de ítems, los participantes se convocan por conveniencia, según las características de los perfiles profesionales y académicos. Durante el periodo 2022-1 al 2023-1, participaron 46 docentes en total: 26 fungieron como diseñadores y 20 como jueces. Se procuró la distribución equilibrada de estos docentes en las tres áreas de conocimiento, tal como se ilustra en la Tabla 19.

Tabla 19.
Distribución de diseñadores y jueces según área y periodo del ExIES (2022-2023)

Área	Rol	2022–1	2022–2	2023–1	Subtotal
Lectura	Diseñadores	5	2	1	8
	Jueces	3	2	1	6
Lengua escrita	Diseñadores	5	2	1	8
	Jueces	3	2	1	6
Matemáticas	Diseñadores	6	2	2	10
	Jueces	4	2	2	8
Total, por periodo	–	26	12	8	46

Nota. Elaboración propia *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (p. 10), por Pedroza Zúñiga et al., 2024a.

Población de aspirantes. La población para la aplicación del ExIES, de la versión 2023-1, fue de 28,205 aspirantes universitarios (Pedroza Zúñiga et al., 2024a), correspondientes a la convocatoria 2023-1. La Tabla 20 detalla esta distribución.

Tabla 20.
Proporción de aspirantes por campo según la clasificación del INEGI (2011)

Campo	Proporción
Agronomía y veterinaria	4.0 %
Artes y humanidades	2.5 %
Ciencias naturales, exactas y de la computación	2.5 %
Ciencias sociales, administración y derecho	32.3 %
Educación	1.4 %
Ingeniería, manufactura y construcción	17.5 %
Salud	32.0 %
Servicios	7.7 %
Total	100 %

Nota. Adaptado de la clasificación propuesta por el INEGI (2011). Las áreas con mayor proporción de aspirantes son *ciencias sociales, administración y derecho* (32.3 %) y *salud* (32.0 %), seguidas por *ingeniería, manufactura y construcción* (17.5 %). Las proporciones corresponden al 100% del total de aspirantes.

Consideraciones éticas

Se resguardaron los documentos técnicos y bases de datos, de acuerdo con las disposiciones de la UABC y el IIDE sobre difusión de información interna. Y para evitar el conflicto de interés, la investigadora es externa al ExIES, asegurando independencia en la evaluación. Asimismo, se atendieron criterios de confiabilidad, equidad y uso adecuado de los puntajes, dejando constancia documental de procedimientos y resultados; y se siguieron las consideraciones y criterios de los Estándares de la AERA et al. (2014) para el proceso de validación de pruebas.

Argumento de interpretación y uso

Una vez que se identifica el estado actual del examen o forma evaluativa, se desarrolla el AIU. En términos operativos, el AIU expresa la conclusión principal que se busca sostener con evidencias a lo largo de las inferencias (Kane, 2006, 2013). Para el ExIES, la investigación formuló la siguiente conclusión general:

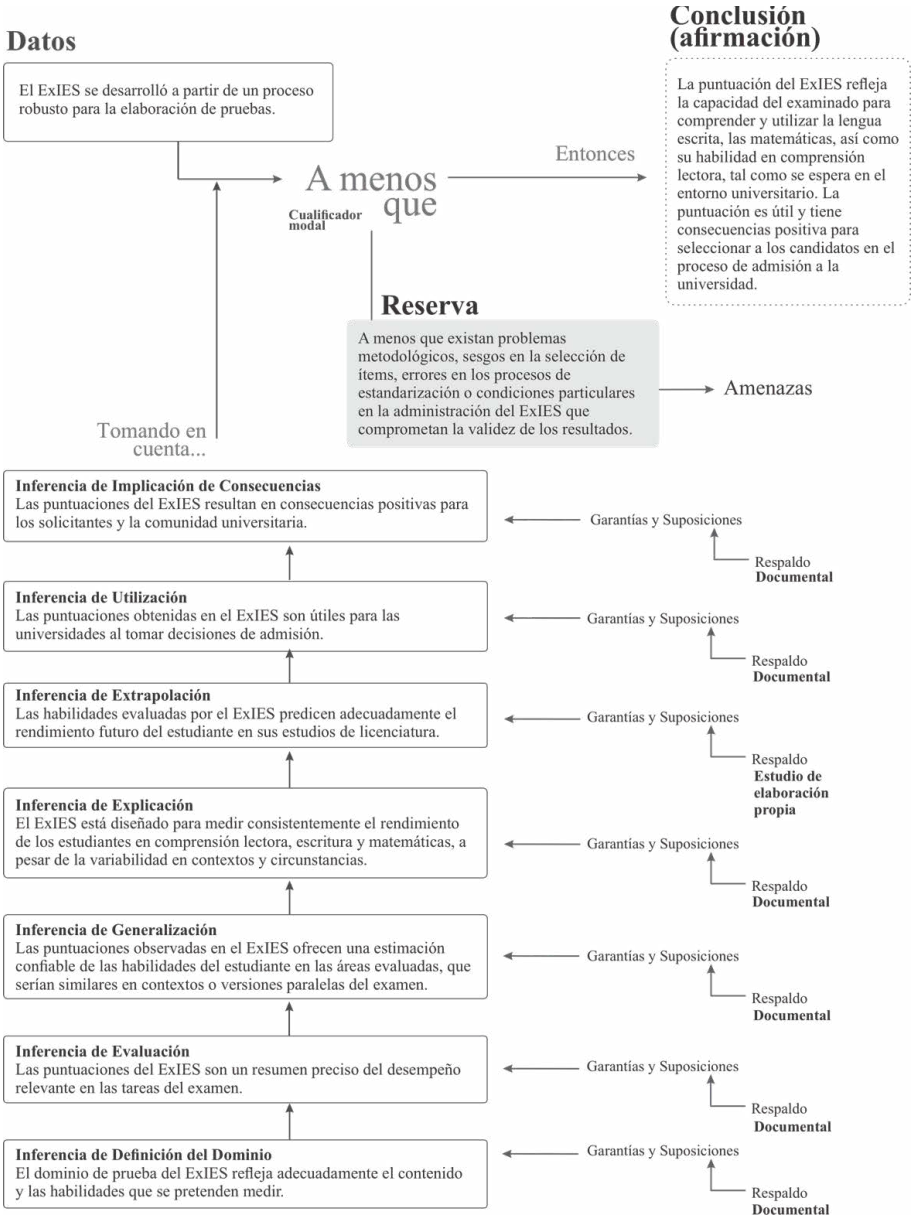
La puntuación del ExIES refleja la capacidad del examinado para comprender y utilizar la Lengua Escrita, las Matemáticas y la habilidad en Lectura, tal como se espera en el entorno universitario. La puntuación es útil y tiene consecuencias positivas en la selección de candidatos para la universidad.

En la Figura 3 se muestra un ejemplo de AIU expresado con la lógica de Toulmin; este formato facilita ubicar afirmaciones, datos y reservas antes de pasar al detalle por inferencias (Kane, 2013; Toulmin, 2003).

Para facilitar el uso de este ejemplo, del EBA al ExIES, cada capítulo posterior se organiza con una lógica recurrente: se explica el sentido de la inferencia, se enuncian los supuestos que la sostienen, se describen las fuentes de evidencia consultadas y se valora su fuerza, incluyendo reservas cuando la evidencia resulta insuficiente. Es decir, a partir de esta conclusión general, las inferencias siguientes especifican qué condiciones deben cumplirse para interpretar y usar los puntajes del ExIES de manera defendible, y qué evidencias permiten respaldar (o cuestionar) cada eslabón del argumento (AERA et al., 2014; Chapelle, 2021). Este patrón busca que el lector pueda seguir el EBA de manera secuencial, pero también consultar capítulos específicos cuando necesite justificar una decisión particular (Kane, 2013; Chapelle, 2021). Asimismo, si bien se expresan los resultados, hay algunas descripciones, figuras y tablas que se omiten por cuestión de espacio, pero que dan la idea suficiente de cómo organizar y describir las evidencias propias.

De forma complementaria, el libro puede leerse como una ruta de trabajo, como guía. Si el examen se encuentra en fase de diseño o rediseño, conviene iniciar por definición de dominio y evaluación para alinear constructo, tareas e ítems; si el examen está en operación, resulta útil revisar primero generalización y explicación para verificar consistencia y estructura interna; y, cuando el uso implica decisiones de admisión, extrapolación y consecuencias permiten discutir relaciones con desempeño posterior y efectos del uso de puntajes (AERA et al., 2014; Messick, 1989; Newton y Shaw, 2014; Shepard, 2016).

Figura 3.
Modelo de Toulmin: AIU del ExIES



Nota. Elaboración propia basada en *The Uses of Argument* (p. 92; Toulmin, 1958/2003) y en *Argument-Based Validation in Testing and Assessment* (p. 104; Chapelle, 2021).

Capítulo 3

Inferencia de definición de dominio

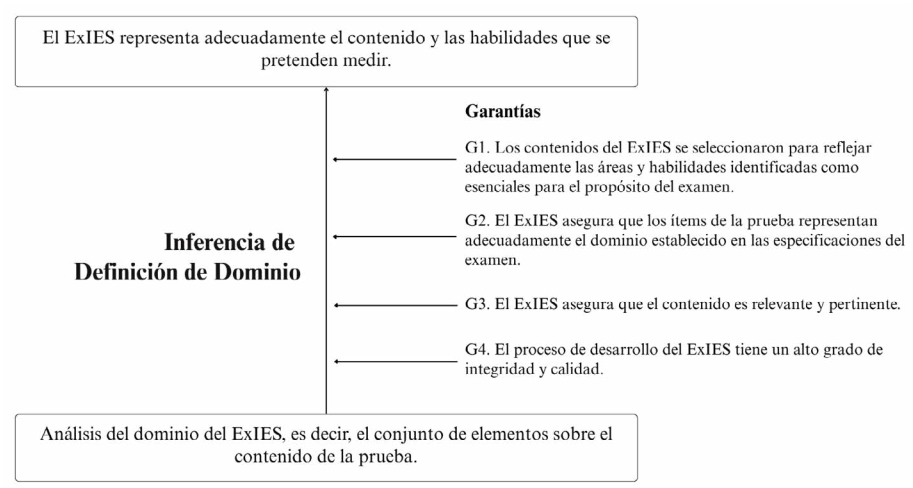
En el EBA, la inferencia de definición de dominio funcionó como el punto de partida del argumento de validez: delimitó el dominio objetivo y permitió justificar que el contenido del examen representó, de forma suficiente y pertinente, aquello que se pretendía evaluar (Chapelle, 2021; Kane, 2013; Sireci, 1998, 2008). En exámenes, esta inferencia es crucial porque definió qué saberes y habilidades se consideraron "relevantes" y, por tanto, qué se dejó dentro y fuera de la prueba.

Aplicarla supuso describir el dominio con base en documentos curriculares y normativos, traducirlo a especificaciones y mapas de contenido, y documentar controles de calidad (revisión experta, pilotaje, análisis estadísticos) que redujeron la subrepresentación del constructo y la varianza irrelevante (AERA et al., 2014; Messick, 1989). En la revisión del ExIES 2023-2, la pregunta guía fue: ¿en qué medida el contenido del examen representó el dominio que la UABC esperaba evaluar en aspirantes a educación superior? A continuación, se sigue la descripción de los elementos, pero también de los resultados propios del ExIES, esperando que funcionen como guía para otros procesos de validación.

Definición de las garantías, supuestos y fuentes de la inferencia de definición de dominio

La conclusión de la inferencia de definición de dominio afirma que el ExIES representa adecuadamente los contenidos y las habilidades que conforman el constructo teórico delineado. Para sustentar, se contemplaron cuatro garantías; véase la Figura 10.

Figura 10.
Argumento de la inferencia de definición de dominio como parte de la validez del argumento



A partir de las garantías, se integraron los supuestos y fuentes de datos, véase la Tabla 21, que respaldan la pertinencia, representatividad y calidad de los contenidos evaluados por el ExIES. La defensa de estos argumentos se presenta en el apartado siguiente, donde se desarrollan los respaldos a partir de las fuentes seleccionadas y así poder evaluarlos.

Se anticipa que en las investigaciones se puede hacer uso solo del formato de figura o tabla para expresar la inferencia, aunque se recomienda el formato de tabla, ya que es más completa; usar ambas también es conveniente y es una expresión visual que ayuda al lector a comprender el argumento.

Tabla 21

Estructura argumentativa para la inferencia de definición de dominio

Conclusión de definición de dominio	El ExIES representa adecuadamente el contenido y las habilidades que se pretenden medir.	
Garantía	**Suposiciones**	**Fuentes de datos**
G1.1. Los contenidos del ExIES se seleccionaron para reflejar adecuadamente las áreas y habilidades identificadas como esenciales para el propósito del examen.	S.1.1.1 El contenido de la prueba se define a partir de las competencias básicas de la educación media superior. S.1.1.2 Las especificaciones de la prueba se desarrollan a partir de un análisis exhaustivo de múltiples fuentes para garantizar que reflejen adecuadamente el contenido y la estructura del examen. S.1.1.3 Hay un proceso continuo de revisión y actualización del dominio de prueba y de las especificaciones del examen para asegurar que sigan siendo actuales y relevantes.	F1.1.1.1 Manual técnico del Nuevo Examen de Selección (Caso et al., 2017) F1.1.1.2 Guía para la evaluación de ítems del nuevo examen de selección de aspirantes a ingresar a la Universidad Autónoma de Baja California (Caso y Díaz, 2016) F1.1.2.1 Especificaciones de lectura, lengua escrita y matemática (Pedroza Zúñiga et al., 2023n, 2023o, 2023p) F1.1.2.2 Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) F1.1.3.1 Tabla comparativa de cambios por área y versiones en las especificaciones. (Elaboración propia) F1.1.3.2 Manual Técnico del ExIES (Pedroza Zúñiga et al., 2022).

G1.2. El ExIES asegura que los ítems de la prueba representan adecuadamente el dominio establecido en las especificaciones del examen.	S.1.2.1 Los ítems de la prueba se adhieren a las especificaciones que establecen las proporciones apropiadas de habilidades, conceptos y niveles de habilidad cognitiva requeridos. S.1.2.2 Se desarrollan estrategias de revisión detalladas para identificar y refinar los ítems de la prueba para que cumpla con las especificaciones. S.1.2.3 Hay un proceso de revisión constante para garantizar que los ítems de la prueba cumplan con las especificaciones y sean actualizados o revisados según sea necesario.	F1.2.1.1, 1.2.2.1 Manual para el desarrollo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023a, 2023b, 2023c); Base de datos completa de Resultados Rasch y estadísticas ítem–forma (2023s) F1.2.1.2 Especificaciones de lectura, lengua escrita y matemática (Pedroza Zúñiga et al., 2023n, 2023o, 2023p); F1.2.2-3 Manual para el jueceo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023d, 2023e, 2023f) F1.2.2.4 Base de datos completa de Resultados Rasch y estadísticas ítem–forma (2023s) F1.2.2-3 Base de datos de los jueceos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023q) F1.2.3.1 Base de datos de organización de ítems, histórico del ExIES: control de ítems NDC-especificación-contenido (Pedroza Zúñiga et al., 2024c)
G1.3. El ExIES asegura que el contenido es relevante y pertinente.	S.1.3.1 Los ítems de la prueba se revisan interna y externamente para identificar y eliminar cualquier contenido no relevante.	F1.3.1.1 Manual para el desarrollo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023a, 2023b, 2023c)
	S.1.3.2 Hay procesos para revisar y ajustar cualquier ítem potencialmente sesgado o inapropiado antes de su publicación.	F1.3.2.1 Manual para el jueceo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023d, 2023e, 2023f)

G1.4. El proceso de desarrollo del ExIES tiene un alto grado de integridad y calidad.	S.1.4.1 Los desarrolladores de ítems están calificados y entrenados en la construcción de estos. S.1.4.2 Hay un proceso riguroso para el desarrollo de ítems que involucra revisiones por múltiples expertos y especialistas. S.1.4.3 Se llevan a cabo análisis avanzados en formas operativas para monitorear la calidad del ítem.	E1.4.1.1 Manual para el desarrollo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023a, 2023b, 2023c) F1.4.1.2 Manual para el jueceo de reactivos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023d, 2023e, 2023f) F1.4.1.3 Presentación de las capacitaciones del ExIES en la elaboración de reactivos (Pedroza Zúñiga et al., 2023h) F1.4.2.1 Base de datos de los jueceos de Lectura, Lengua Escrita y Matemáticas del ExIES (Pedroza Zúñiga et al., 2023q) F1.4.2.1, 1.4.3.1 Análisis TRI, Reporte Técnico del ExIES 2023-1 y 2023-2 (Pedroza Zúñiga et al., 2024a, 2024b) F1.4.3.2 Base de datos histórica del ExIES (Pedroza Zúñiga et al., 2024c)

Desarrollo de respaldos de la inferencia de definición de dominio

Garantía 1.1: Los contenidos del ExIES se seleccionaron para reflejar adecuadamente las áreas y habilidades identificadas como esenciales para el propósito del examen.

S.1.1.1 El contenido de la prueba se define a partir de las competencias básicas de la educación media superior. Según el Reporte Técnico (Pedroza Zúñiga et al., 2024a), el ExIES 2023-1 parte de la propuesta del Manual Técnico 2017 (Caso et al., 2017). El primer acercamiento al contenido de la prueba fue un análisis por competencias que parte de los Acuerdos 442 y 444 de la SEP (2008a, 2008b); asimismo, se revisaron

competencias genéricas, disciplinares y profesionales, así como la convergencia con los lineamientos de la OCDE, DeSeCo y PISA (Caso et al., 2017). De este modo, el ExIES se estructuró en tres áreas: Lectura, Lengua Escrita y Matemáticas (Caso et al., 2017; Caso y Díaz, 2016). En la Tabla 22 se compara el enfoque disciplinar de la EMS con el Proyecto DeSeCo y la evaluación PISA.

Tabla 22
Tabla comparativa sobre los campos disciplinares

Campos disciplinares EMS	Proyecto DeSeCo, OCDE	Evaluación PISA
Matemáticas: álgebra, aritmética, cálculo, trigonometría, estadística	Competencia matemática y competencias en ciencias	Matemáticas: cantidad, espacio y forma, cambio y relaciones, probabilidad
Ciencias experimentales: física, química, biología, ecología	Pensamiento científico y habilidad para usar la tecnología de forma interactiva	Ciencias: física, química, ciencias biológicas, ciencias de la tierra y el espacio
Ciencias sociales: historia, derecho, sociología, política, antropología, economía, administración	Competencias sociales y cívicas	Contexto: Situación pública, vida y salud, tierra y medio ambiente
Humanidades: literatura, filosofía, ética, lógica, estética	Conciencia y expresiones culturales	Contexto: conciencia cultural
Comunicación: lectura y expresión oral y escrita, taller de lectura y redacción, lengua adicional al español, tecnologías de la información y la comunicación	Competencia en comunicación lingüística	Lectura: textos continuos, textos discontinuos

S.1.1.2 Las especificaciones de la prueba se desarrollan a partir de un análisis exhaustivo de múltiples fuentes para garantizar que reflejen adecuadamente el contenido y la estructura del examen. Para garantizar que las especificaciones de la prueba reflejen adecuadamente el contenido y la estructura del examen, se siguió un proceso que incluyó (Caso et al., 2017; Caso y Díaz, 2016; Pedroza Zúñiga et al., 2024a):

- Análisis curricular: Revisión detallada de los currículos de la EMS para alinear los contenidos del examen con lo que se enseña en las aulas.
- Revisión de estándares internacionales: Incorporación de competencias y marcos evaluativos de reconocidos programas internacionales como PISA para asegurar una comparación global y mantener un estándar alto.
- Consultas con expertos Colaboración con educadores y especialistas en contenido para asegurar la validez de contenido y la relevancia educativa.

Es importante resaltar que, a pesar de contar con el Manual Técnico (Caso et al., 2017), así como la Guía para la Evaluación de ítems (Caso y Díaz, 2016), no se cuenta con una tabla comparativa u otro tipo de evidencia donde se muestren los procedimientos seguidos y resultados obtenidos de las tres acciones anteriormente descritas; sin embargo, se describe que se realizó la consulta con expertos a partir del análisis curricular, así como la revisión de los estándares internacionales.

Según el reporte del ExIES 2023-1 (Pedroza Zúñiga et al., 2024a), la actualización y desarrollo de las especificaciones del instrumento parten de Lane et al. (2016), donde también se consideran los estándares 4.1, 4.2, 11.3 y 12.4 de la AERA, la APA y el NCME (2014). Además, entre 2022-2 y 2023-1 se elaboraron especificaciones para cada una de las áreas (Pedroza Zúñiga et al., 2023n, 2023o, 2023p). Cada especificación se estructura en secciones claramente definidas que incluyen la identificación y definición del contenido a evaluar, la delimitación precisa de los alcances, los conocimientos y habilidades previas requeridas, las actividades cognitivas implicadas, ejemplos de aplicación, la plantilla del ítem, peculiaridades en la redacción y elaboración de opciones, así como la bibliografía consultada (véase Tabla 23). Este proceso sistemático y documentado permite asegurar la validez de contenido de los reactivos y sustentar teóricamente cada decisión tomada en el desarrollo de la prueba.

Tabla 23.
Secciones de las especificaciones de ítems y su descripción

Sección de la especificación	Descripción
Identificación del contenido	Incluye el área, contenido, subcontenido y el nivel cognitivo que se evaluará, especificando a detalle qué se pretende medir.
Definición del contenido	Precisa de manera clara el constructo o habilidad que se busca evaluar, delimitando su significado dentro del contexto de la prueba.
Delimitación del contenido	Describe los alcances y límites del contenido a evaluar, estableciendo el enfoque y los aspectos que serán incluidos o excluidos del ítem.
Conocimientos y habilidades previas	Enumera los saberes, habilidades y competencias que el sustentante debe poseer para poder responder correctamente el reactivo.
Actividades cognoscitivas	Expone los procesos mentales o habilidades cognitivas necesarias para resolver el ítem, de acuerdo con su nivel de complejidad.
Ejemplos de aplicación	Proporciona ejemplos específicos o ilustrativos que muestran el tipo de contenido y las formas en que puede ser evaluado.
Plantilla del ítem	Presenta la estructura base y el formato del reactivo, indicando los elementos que debe contener, como instrucciones, texto base y opciones.
Peculiaridades de la plantilla	Detalla características particulares del diseño, redacción, extensión, vocabulario y elaboración de opciones o distractores.
Bibliografía consultada	Enumera las fuentes y referencias empleadas para fundamentar teóricamente la especificación y el diseño del reactivo.

Nota. Elaboración propia, basada en la revisión de las *Especificaciones de Lectura, Lengua Escrita y Matemáticas* (Pedroza Zúñiga et al.,2023n, 2023o, 2023p).

S.1.1.3 Hay un proceso continuo de revisión y actualización del dominio de prueba y de las especificaciones del examen para asegurar que sigan siendo actuales y relevantes. Según el *Reporte Técnico del ExIES 2023-1* (Pedroza Zúñiga et al., 2024a), a partir de las recomendaciones sobre la elaboración de pruebas de Lane et al. (2016) y los Estándares (AERA et al., 2014), de forma anual existe la revisión y evaluación a través del juicio de expertos en las áreas de: a) contenido, b) ponderaciones y c) niveles de demanda cognitiva, partiendo de la tabla de especificación de contenidos, realizando las adecuaciones pertinentes y, de forma posterior,

las equiparaciones sobre cada área evaluada. En este sentido, la versión piloto del ExIES, implementada en 2022-2 (Pedroza Zúñiga et al., 2022), experimentó cambios significativos en la versión oficial de 2023.

Tabla 24.
Cantidad de ítems con cambios por versión y componente

Componente	Versión 2022	Versión 2023	Variación neta	Observaciones principales
Lectura	15	11	–4	Se agrupan o eliminan varios subcontenidos en Información e ideas; se renombró Evaluación del tono textual como Evaluación del estilo, entre otros ajustes.
Lengua escrita	6	18	+12	Se detallan mucho más las reglas del español escrito (puntuación, ortografía, concordancia); el bloque Expresión de ideas escritas se subdivide en tópicos más específicos.
Matemáticas	40	39	–1	En Herramientas algebraicas y Problemas, probabilidad y análisis de datos se fusionan o eliminan ciertos temas; temas adicionales aumentan ligeramente (+1); "Matemáticas avanzadas" mantiene el mismo número total (14), con cambios de denominación.
Total	61	68	+7	Crecimiento global, principalmente por la gran desagregación en Lengua Escrita; más granularidad en la descripción de competencias y mayor reorganización global.

Nota. Elaboración propia basada en el apéndice D de las especificaciones, sobre los cambios comparativos entre versiones.

Garantía 1.2. El ExIES asegura que los ítems de la prueba representan adecuadamente el dominio establecido en las especificaciones del examen

S.1.2.1 Los ítems de la prueba se adhieren a las especificaciones que establecen las proporciones apropiadas de habilidades, conceptos y niveles de habilidad cognitiva requeridos. Como se señala en la Garantía 1, las especificaciones se establecen claramente, se someten a seguimiento, revisión continua y modificaciones periódicas. Además, las tablas de especificaciones son sometidas a un proceso de validación mediante jueceo de expertos. Para asegurar que los ítems se alineen a estas especificaciones, se cuenta con manuales específicos para el desarrollo de reactivos (Pedroza Zúñiga et al., 2023a, 2023b, 2023c). Por otro lado, también se cuentan con especificaciones por componente del ExIES. En la Tabla 25 se describe el contenido esencial de las especificaciones que sustentan la Suposición 1.2.1, señalando cómo cada una garantiza la adherencia de los ítems a las proporciones adecuadas de habilidades, conceptos y NDC.

Tabla 25.
Secciones de las especificaciones del ExIES

Sección	Contenido principal
1. Estructura y organización formal	Identificación de la especificación (coordinador, redactor, fecha, identificador único). Declaración del contenido macro a evaluar (p. ej., expresión de ideas escritas en lengua escrita). Definición o alcance del contenido, aclarando su relación con el dominio global del área.
2. Subcontenido y nivel cognitivo	Detalle del subcontenido específico (p. ej., uso de conectores, uso de oraciones subordinadas). Indicación del nivel de demanda cognitiva (comprensión, aplicación, evaluación, etc.), según la taxonomía de Bloom. Garantiza la alineación del ítem con el grado de complejidad esperado.
3. Descripción del contenido a evaluar	Especificación de los aspectos particulares que se miden (coherencia, concordancia, conectores, etcétera). Delineación de las habilidades previas que el sustentante debe poseer. Clarificación de la actividad cognoscitiva involucrada en la resolución correcta del ítem.
4. Plantilla del ítem (estructura base)	Definición de la forma general del ítem: instrucciones claras, extensión máxima de textos, longitud similar en distractores. Redacción esperada de opciones correctas e incorrectas (errores frecuentes como distractores). Especificación de la temática preferente (literatura, ciencias, etc.), vocabulario apto y restricciones (evitar textos con derechos restringidos).
5. Peculiaridades o lineamientos específicos	Condiciones adicionales (extensión de 50 palabras por párrafo, justificación en caso de superar ese límite, originalidad del texto). Orientaciones sobre referencias, ortografía, puntuación y coherencia interna de la pregunta.
6. Bibliografía y fuentes	Inclusión de las referencias consultadas: manuales, lineamientos de la RAE, taxonomía de Bloom (1956), diccionarios, etc. Presentación de la fundamentación teórica y metodológica que respalda el contenido evaluado.

Nota. Elaboración propia basada en la revisión de *especificaciones de lectura, lengua escrita y matemáticas* (Pedroza Zúñiga et al., 2023n, 2023o, 2023p).

S.1.2.2 Se desarrollan estrategias de revisión detalladas para identificar y refinar los ítems de la prueba para que cumpla con las especificaciones. Se cuenta con un Manual de Jueceo por componente (Pedroza Zúñiga et al., 2023d, 2023e, 2023f) para cada una de las áreas comprendidas, lectura, lengua escrita y matemáticas, donde se especifican seis secciones de acompañamiento:

1) Normas de seguridad.
2) Proceso de diseño, construcción y validación del ExIES.
3) Pasos para el desarrollo de los ítems.
4) Indicaciones generales para la construcción de ítems.
5) Tipos de ítems, según la competencia a evaluar.

A partir de ello, se especifican los criterios de evaluación individual, como la evaluación colegiada de los ítems, es decir, a través del jueceo; posteriormente, son recolectadas por las coordinaciones para dar seguimiento a los cambios. En este sentido, existen bases de datos sobre jueceos que cada coordinador alimenta de forma individual (Pedroza Zúñiga et al.,2023q); falta mayor sistematización para este proceso, por ejemplo, una tabla de seguimiento común, así como análisis globales de estos procesos. Los criterios que se siguen son los siguientes:

1) Basarse en la tabla de especificaciones del ítem.
2) Responder el ítem.
3) Verificar que los ítems cumplan los criterios para su evaluación.
4) Emitir el dictamen en su documento correspondiente, incluyendo comentarios y los indicadores señalados: Descartar, Modificar con Cambios Mayores, Modificar con Cambios Menores, Aceptar.
5) Justificar los errores o las problemáticas encontradas.
6) Proponer alguna mejora al ítem si lo requiriera.

S.1.2.3 Hay un proceso de revisión constante para garantizar que los ítems de la prueba cumplan con las especificaciones y sean actualizados o revisados según sea necesario. Según cada Manual para el jueceo de reactivos (Pedroza Zúñiga et al., 2023d, 2023e, 2023f), el proceso para el desarrollo del ExIES consta de cinco etapas generales (ver Tabla 26); a su vez, subdivididas en pasos específicos que delinean el trabajo que se

realiza. Algunos de estos pasos involucran la participación de expertos en el área de lectura, lengua escrita y matemáticas; la colaboración con estos actores garantiza que el examen evalúe información relevante y pertinente, pues su conocimiento y experiencia en el área les permite realizar valiosas aportaciones en este proceso de construcción. Con el fin de garantizar este proceso, se guardan reportes de los jueceos realizados en formato de tabla para dar seguimiento puntual a cada ítem.

Tabla 26.
Etapas y subetapas para la construcción del ExIES

Etapa	Pasos
1. Planeación general y diseño del instrumento	1. Plan inicial
	2. Diseño del instrumento
2. Construcción y validación de ítems	3. Elaboración de ítems
	4. Jueceo de ítems (independiente)
	5. Correcciones
	6. Jueceo de ítems (grupal)
	7. Pilotaje de los ítems
	8. Análisis de resultados de aplicación piloto
3. Aplicación Institucional	9. Capacitación de aplicadores
	10. Diseño
	11. Reproducción
	12. Administración de los instrumentos
4. Procesamiento y edición de los datos	13. Lectura de los instrumentos
	14. Validación y calificación de la información
5. Resultados y propiedades métricas	15. Plan de análisis
	16. Generación de resultados individuales y agregados
	17. Análisis psicométricos
	18. Elaboración del informe técnico

Nota. Reimpreso de *Manual para el jueceo de reactivos: Lectura* [manuscrito no publicado], por L. H. Pedroza Zúñiga, S. A. García Aldaco, C. Gómez Monárrez, M. A. Orozco Vergara, K. K. Ruiz Mendoza y A. P. Gutiérrez Zavala, 2023; de *Manual para el jueceo de reactivos: Lengua escrita* [manuscrito no publicado], por los mismos autores, 2023; y de *Manual para el jueceo de reactivos: Matemáticas* [manuscrito no publicado], por los mismos autores, 2023.

Garantía 1.3. El ExIES asegura que el contenido es relevante y pertinente

S.1.3.1 Los ítems de la prueba se revisan interna y externamente para identificar y eliminar cualquier contenido no relevante. En primera instancia, en cada uno de los manuales para la elaboración de los ítems (Pedroza Zúñiga et al., 2023d, 2023e, 2023f), se definen cuestiones de formato, sesgos, así como se provee de una rúbrica para la evaluación de ítems, la cual considera los aspectos de la Tabla 27 que ejemplifica lo realizado en el componente lectura.

Tabla 27.
Áreas del constructo de la rúbrica para la evaluación de ítems (2023) del ExIES

Área del constructo	Descripción	Ejemplo de elemento evaluado
Claridad y relevancia del texto	Evalúa si el texto proporciona la información necesaria para responder los ítems sin ambigüedades ni contenido superfluo.	"El texto no contiene información irrelevante para los ítems asociados".
Comprensibilidad	Asegura que el tema del texto sea comprensible para el público objetivo.	"El tema central del texto es comprensible para la población objetivo".
Ausencia de respuestas directas	Verifica que el texto no obvie las respuestas a los ítems.	"El texto no contiene las respuestas explícitas a los ítems directos".
Longitud y formato del texto	Determina si la longitud del texto es adecuada y justificada, y si sigue el formato establecido.	"El texto tiene 50 palabras o menos o se justifica una extensión mayor".
Ausencia de sesgos	Confirma que el texto está libre de sesgos culturales, de género, localismos, estereotipos o temas controversiales.	"El texto está libre de cualquier tipo de sesgo y estereotipo".
Adecuación al nivel cognitivo	Evalúa si el ítem refleja el nivel de demanda cognitiva establecido.	"El ítem refleja el nivel de demanda cognitiva acorde a la tabla de especificaciones".

Corrección disciplinar	Verifica que el ítem mantenga la brevedad y que plantee un problema central definido comprensible para la población objetivo.	"El ítem plantea un único problema central claro y comprensible".
Estructura y redacción	Evalúa la claridad en la redacción del ítem y la adecuación del vocabulario, la ortografía y la ausencia de sesgos.	"El ítem está redactado con claridad y usa un vocabulario adecuado sin errores ortográficos".
Diseño de respuestas	Revisa la coherencia y adecuación de las opciones de respuesta, incluyendo su longitud y plausibilidad.	"Las opciones de respuesta tienen longitudes similares y no dan pistas sobre la respuesta correcta".
Formato y presentación	Evalúa el uso correcto de la negrita y la indicación de los números de línea cuando se requiere.	"Las palabras o frases clave están en negrita y los números de línea están correctamente indicados".

Nota. Elaboración propia basada en *Manual para el desarrollo de reactivos: Lengua Escrita* (Pedroza Zúñiga et al., 2023b).

Aunado a lo anterior, según la Tabla 27, tomando en cuenta los manuales de jueceo (Pedroza Zúñiga et al., 2023d, 2023e, 2023f), existe un proceso para revisar de forma interna y externamente, a través de jueces, para identificar, cambiar o eliminar los contenidos no relevantes. Para la afirmación, la revisión interna y externa por jueces se realiza como parte de los pasos 4 y 6 descritos en el proceso de desarrollo del examen. En la fase de jueceo de ítems (independiente y grupal), los expertos evalúan cada ítem para asegurarse de que cumplan con los criterios técnicos y sean relevantes para las competencias que se buscan medir.

Los ítems irrelevantes o que no cumplen con los estándares establecidos son modificados o eliminados. Este enfoque es un método estandarizado en el campo de la psicometría para mejorar la validez de contenido del examen. En este sentido, se estableció en el Reporte técnico 2023-1 (Pedroza Zúñiga et al., 2024a) un seguimiento sobre la revisión de sesgos mediante análisis DIF, el cual se detalla en la inferencia de explicación.

S.1.3.2 Hay procesos para revisar y ajustar cualquier ítem potencialmente sesgado o inapropiado antes de su publicación. Este mismo

proceso de jueceo, que incluye revisiones independientes y colegiadas, funciona para identificar y corregir cualquier potencial sesgo en los ítems. Durante el jueceo, los expertos están atentos a cualquier elemento del ítem que pueda ser injusto o inapropiado para algún grupo de aspirantes. Estas descripciones se encuentran en las especificaciones generales y en la guía de evaluación de ítems individuales del manual de jueceo, según su área (Pedroza Zúñiga et al., 2023d, 2023e, 2023f); no obstante, carecen de lineamientos específicos como un manual o ejemplos de malos usos, ya que estos se subsanan a través de los jueceos. Asimismo, se señalan dos principales (Pedroza Zúñiga et al., 2023e, p.20): 1) Estar libre de expresiones idiomáticas locales que dificulten su comprensión, debido al contexto de frontera; y 2) evitar todo tipo de sesgo (cultural, social, de género, etcétera). Es decir, vocabulario o cualquier tipo de representación que ofenda a un grupo de sustentantes en particular y/o facilite la identificación de la respuesta correcta o dificulte contestar correctamente el reactivo.

Estas revisiones se complementan con un pilotaje —de los ítems nuevos— que proporciona una verificación empírica adicional: durante el pilotaje se realizan análisis de Funcionamiento Diferencial del Ítem (DIF), empleando técnicas como Mantel–Haenszel (Holland y Thayer, 1988), para identificar reactivos que muestran probabilidades distintas de respuesta entre subgrupos con igual nivel de habilidad; además, se monitorean los distractores y otras señales de comportamiento atípico en los ítems. Cuando el jueceo identifica problemas potenciales, se documentan las modificaciones (Pedroza Zúñiga et al., 2024a). Sin embargo, aún no se lleva a cabo el cambio u omisión de un ítem basado en los resultados del análisis DIF aún no se lleva a cabo, y por el momento solo se tiene como referencia.

Garantía 1.4. El proceso de desarrollo del ExIES tiene un alto grado de integridad y calidad

S.1.4.1 Los desarrolladores de ítems están calificados y entrenados en la construcción de estos. Según el Reporte técnico 2023-1 (Pedroza Zúñiga et al., 2024a), así como los manuales para el jueceo de los ítems (Pedroza Zúñiga et al., 2023d, 2023e, 2023f), los desarrolladores de ítems son

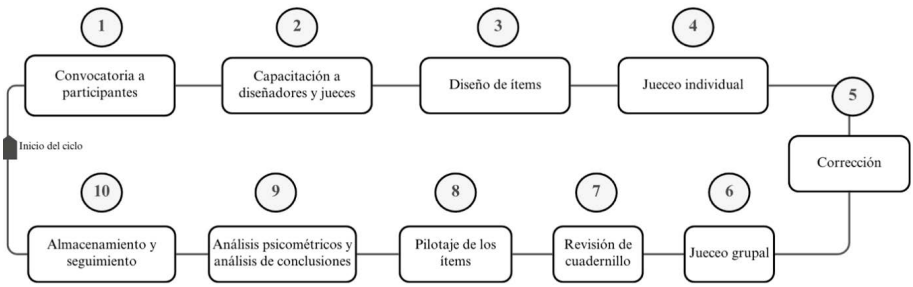
cuidadosamente seleccionados a partir de recomendaciones específicas para cada área de conocimiento. Se asegura que cada participante tenga experiencia relevante en proyectos evaluativos previos y las credenciales académicas y profesionales necesarias para la construcción de ítems de calidad. Los jueces, en cambio, son seleccionados una vez que han adquirido experiencia en la generación de ítems y según los resultados de las métricas de sus ítems (análisis de TRI). Una vez seleccionados, los desarrolladores de ítems pasan por un proceso de capacitación que incluye:

- Un curso en línea por área de conocimiento en una plataforma educativa.
- Sesiones de capacitación sincrónica en línea —vía Google Meet— dividida en una sesión general y sesiones específicas por área de conocimiento, dirigidas por coordinadores de área.
- Los siguientes materiales:
  1. Manual para la elaboración de ítems o para el jueceo de ítems.
  2. Presentación sintética de la elaboración de ítems, incluyendo ejemplos por contenido.
  3. Especificaciones de cada uno de los subcontenidos por área.

A partir de dichos resultados, se implementa un proceso de retroalimentación continua que identifica a los diseñadores que requieren apoyo adicional en su capacitación o, en casos puntuales, su reemplazo para la elaboración de nuevos ítems, asegurando así la consistencia y mejora en la construcción de ítems; por ende, se cumple con este supuesto adecuadamente.

S.1.4.2 Hay un proceso riguroso para el desarrollo de ítems que involucra revisiones por múltiples expertos y especialistas. El Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) describe un proceso completo para la elaboración de ítems, fundamentado en los aportes de Brijmohan et al. (2018) y Jornet et al. (2010), que busca salvaguardar la validez y confiabilidad de cada ítem. Tal como se muestra en la Figura 11, este ciclo inicia con la elaboración inicial de ítems por parte de desarrolladores capacitados, seguida de un jueceo independiente donde expertos revisan críticamente cada propuesta.

Figura 11
Ciclo del diseño de los ítems



Nota. Elaboración propia basada en los datos y descripciones del *Reporte técnico del ExIES 2023-1* (Pedroza Zúñiga et al., 2024a).

A partir de la retroalimentación obtenida, se realizan correcciones y se lleva a cabo un jueceo grupal para garantizar el consenso respecto a la calidad de los ítems. Después de estas etapas, se procede a la revisión de cuadernillos para confirmar que el contenido se alinee con los estándares del ExIES, y se efectúa un pilotaje con una muestra representativa de la población, lo que permite evaluar el desempeño de los ítems en condiciones reales. Finalmente, se realizan análisis psicométricos, particularmente mediante la TRI, para identificar las propiedades psicométricas de los ítems y seleccionar aquellos más adecuados. Este proceso metódico y secuencial garantiza la pertinencia y solidez del examen a lo largo de sus diferentes versiones.

S.1.4.3 Se llevan a cabo análisis avanzados en formas operativas para monitorear la calidad del ítem. Se aplican análisis avanzados y complementarios sobre las formas operativas con el propósito de monitorear y garantizar la calidad técnica de los ítems. El *Reporte Técnico del ExIES 2023-1* (Pedroza Zúñiga et al., 2024a) documenta que los parámetros se estiman a partir de modelos Rasch y que, a nivel ítem, se calculan estadísticos como la dificultad, la correlación punto-biserial y los índices de ajuste (infit y outfit, MNSQ y ZSTD), los cuales permiten detectar ítems mal ajustados o con funcionamiento atípico. Asimismo, estos estadísticos se realizan de forma semestral en la versión correspondiente de los Reportes Técnicos y se cuenta con una base de datos

donde se da seguimiento al histórico de los ítems (Pedroza Zúñiga et al., 2024c). Aunque no se menciona de forma explícita, los análisis realizados incluyen, entre otros, los siguientes procedimientos técnicos:

- Estadísticos ítem-a-ítem (TCT y Rasch): índice de dificultad (p), correlación punto-biserial y parámetros Rasch que ubican ítems en la escala de dificultad (0–1). Estos parámetros permiten seleccionar las combinaciones de ítems con mejores métricas para conformar subversiones (Pedroza Zúñiga et al., 2024a).
- Análisis de distractores: Evaluación de la frecuencia y patrón de selección de distractores para detectar distractores no funcionales o confusos y así mejorar la formulación de ítems. Además, se documenta un seguimiento histórico de ítems y un registro donde se monitorea esta información (Pedroza Zúñiga et al., 2024a).
- Confiabilidad y consistencia interna: estimación de coeficientes de confiabilidad (por ejemplo, alfa de Cronbach) para cada subversión y componente, como indicador de consistencia en las mediciones (Pedroza Zúñiga et al., 2024a).
- Equiparación de formas: procedimientos de calibración concurrente basados en TRI (implementados, por ejemplo, con Winsteps) para garantizar comparabilidad entre diferentes formas y subversiones (Pedroza Zúñiga et al., 2024a; Zieky, 1993).
- Análisis DIF (Pedroza Zúñiga y Gómez Monárrez, 2025c): detección de DIF mediante enfoques complementarios —por ejemplo, Mantel-Haenszel para comparaciones por sexo y análisis por terciles de habilidad— utilizando paquetes estadísticos en R (difR, mirt). En el reporte del análisis DIF del ExIES, se informa el uso de Odds Ratio/logOR para cuantificar la magnitud del DIF y se emplean umbrales empíricos (p. ej., Zieky, 1993) para clasificar DIF moderado o severo.

Como resultado operativo, estos análisis sirven para identificar ítems con oportunidad de mejora (los cuales son señalados en los anexos técnicos y en figura resumen), y sostener decisiones de retiro, reformulación o pilotaje adicional; por ejemplo, en la aplicación 2023-1 (Pedroza Zúñiga et al., 2024a) se reportaron oportunidades de mejora en 22 % de ítems de lectura, 19.5 % en lengua escrita y 10.6 % en matemáticas (datos derivados del procesamiento Rasch y análisis de distractores).

Evaluación de la inferencia de definición de dominio

De acuerdo con el *diseño metodológico*, la Tabla 28 presenta la valoración de los supuestos que integran la inferencia de Definición de Dominio. El resultado global fue de 84.8 %, con niveles moderados en claridad y plausibilidad (84.1 % cada una) y un nivel algo mayor en coherencia (86.4 %). La evidencia se caracteriza por ser consistente y sin contradicciones explícitas, aunque se identifican ambigüedades conceptuales y carencias de precisión, como la falta de homologación de categorías, referencias incompletas y aspectos de formato, lo que explica la clasificación global en un nivel moderado.

Tabla 28.
Evaluación de la inferencia de definición de dominio.

Suposición	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S1.1.1 Definición del contenido	El constructo está claramente definido y es comprensible (Estándar 1.0, 1.2).	3	4	3	10 (83.3 %, Moderada)
S1.1.2 Especificaciones del examen	Las especificaciones de la prueba se desarrollan a partir de un análisis exhaustivo de múltiples fuentes para garantizar que reflejen adecuadamente el contenido y la estructura del examen (Estándar 4.1).	4	4	4	12 (100 %, Alta)
S1.1.3 Revisión y actualización del contenido	Hay un proceso continuo de revisión y actualización del dominio de prueba y de las especificaciones del examen para asegurar que sigan siendo actuales y relevantes (Estándar 6.2).	3	3	3	9 (75 %, Moderada)
S1.2.1 Adherencia a las especificaciones	Los ítems de la prueba se adhieren a las especificaciones que establecen las proporciones apropiadas de habilidades, conceptos y niveles de habilidad cognitiva requeridos (Estándar 2.1).	3	3	3	9 (75 %, Moderada)
S1.2.2 Estrategias de revisión de ítems	Se desarrollan estrategias de revisión detalladas para identificar y refinar los ítems de la prueba para que cumpla con las especificaciones (Estándar 6.2).	3	3	3	9 (75 %, Moderada)
S1.2.3 Revisión continua de los ítems	Hay un proceso de revisión constante para garantizar que los ítems de la prueba cumplan con las especificaciones y sean actualizados o revisados según sea necesario (Estándar 6.2).	3	3	3	9 (75 %, Moderada)
S1.3.1 Revisión interna y externa de contenido	El contenido es evaluado por expertos y se eliminan ítems irrelevantes o desactualizados (Estándar 4.1).	3	4	4	1 1 (91.7 %, Alta)

S1.3.2 Procesos para ajustar ítems sesgados o inapropiados	Hay procesos para revisar y ajustar cualquier ítem potencialmente sesgado o inapropiado antes de su publicación. (Estándar 4.1).	3	4	4	1	1 (91.7 %, Alta)
S1.4.1 Capacitación de los desarrolladores de ítems	Los creadores de ítems poseen experiencia y reciben formación en la elaboración de ítems (Estándar 1.0, 4.1).	4	3	3	1	0 (83.3 %, Moderada)
S1.4.2 Proceso riguroso de desarrollo	La secuencia (elaboración, jueceo individual/grupal, pilotaje, análisis psicométrico) involucra expertos en cada etapa y se documenta en manuales y reportes, reforzando la integridad de la construcción (Estándar 4.1, 11.13).	4	4	3	1	1 (91.7 %, Alta)
S1.4.3 Análisis avanzados en formas operativas	Se aplican métodos psicométricos (TRI, índices de confiabilidad, equiparación) para monitorear la calidad de ítems y realizar ajustes (Estándar 11.14).	4	4	3	1	1 (91.7 %, Alta)
	Global	37 (84.1 %, Moderada)	37 (84.1 %, Moderada)	38 (86.4 %, Alta)	112/ 132 (84.8 %, Moderada)	

Como procedimiento replicable, la inferencia de Definición de Dominio se fortaleció cuando se documentó la cadena dominio-especificaciones-ítems y se conservó trazabilidad entre decisiones de diseño (qué se incluyó o excluyó) y el constructo. Para aplicar este paso en otros exámenes, se recomendó: definir el dominio con referentes curriculares; describirlo en especificaciones internas o públicas; realizar revisión de expertos (jueceo) con criterios claros; y registrar cambios por versión. Cuando alguno de estos elementos faltó, la inferencia quedó vulnerable y se sugirió generar actas y reportes técnicos breves por ciclo.

Capítulo 4

Inferencia de evaluación

La inferencia de evaluación se concentró en la calidad de las observaciones de rendimiento: si la aplicación, la seguridad, las condiciones de administración y la calificación permitieron registrar puntajes comparables y justos. Antes de discutir confiabilidad o estructura interna, este paso preguntó si el proceso de medición produjo datos legítimos para sostener las interpretaciones posteriores (AERA et al., 2014; Chapelle, 2021K; Kane, 2013).

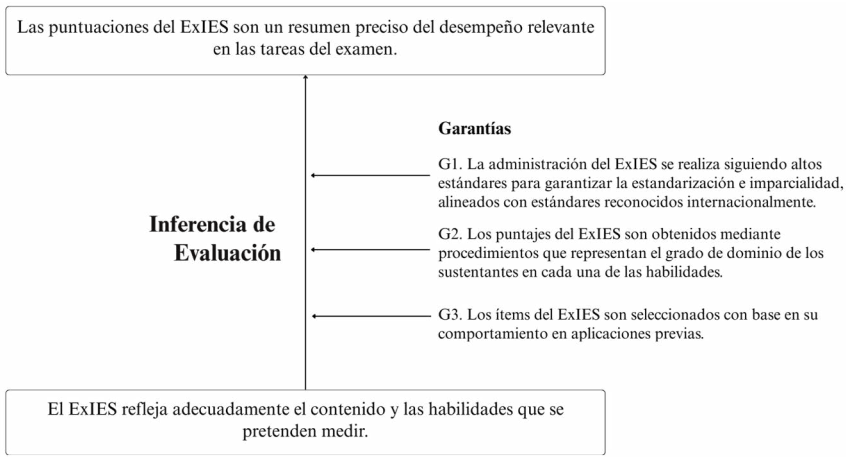
En exámenes de admisión, este análisis incluyó revisar manuales de aplicación, protocolos de estandarización, capacitación y supervisión, así como evidencias de control de calidad en la puntuación. En la revisión del ExIES 2023-2, la pregunta guía fue: ¿en qué medida las condiciones de administración y calificación garantizaron una evaluación justa y estandarizada? Las Tablas 27 y 28 organizan las garantías, supuestos y respaldos identificados.

En este capítulo se reportan los elementos operativos esenciales (manuales, protocolos y criterios) que permiten sostener la inferencia de evaluación en un examen a gran escala. Por el carácter de manual, no se reproducen bitácoras completas de aplicación ni todos los anexos de capacitación; se sintetizan los puntos críticos que suelen comprometer la calidad del puntaje. Como evidencia adicional, pueden desarrollarse auditorías de campo, estudios de consistencia entre calificadores (cuando aplica), evaluaciones de seguridad y revisiones de adecuaciones, accesibilidad o traducción/adaptación cuando el examen se usa en poblaciones diversas (AERA et al., 2014; Hambleton y Zenisky, 2011; Haladyna y Rodríguez, 2013).

Definición de las garantías, supuestos y fuentes de la inferencia de evaluación

En la Figura 12 se establece la conclusión de la inferencia de evaluación: las puntuaciones del ExIES resumen con precisión el desempeño relevante de los estudiantes en las tareas del examenl sustentado por la premisa base de que este refleja adecuadamente las habilidades y contenidos que pretende medir.

Figura 12.
Argumento de la inferencia de evaluación como parte de la validez del argumento.



Considerando la conclusión de evaluación, se sostienen tres garantías complementarias que sostienen esa conclusión: G2.1 (estandarización e imparcialidad) asegura que la administración se realice bajo condiciones uniformes para todos los sustentantes, previniendo sesgos o irregularidades; G2.2 (estimación de la habilidad) afirma que los puntajes reflejan el nivel real de los examinados mediante técnicas psicométricas sólidas —por ejemplo el modelo Rasch— de modo que exista coherencia entre desempeño observado y calificación en las áreas evaluadas (lectura, lengua escrita y matemáticas); y G2.3 (calidad y selección de ítems) enfatiza que los ítems incorporados provienen de procesos de pilotaje y evaluación psicométrica previos, garantizando que cada ítem cumpla criterios adecuados de dificultad y ajuste para que la prueba mida eficazmente las habilidades de los sustentantes.

Así, la Tabla 29 sintetiza de manera estructurada las garantías clave, los supuestos específicos y las fuentes de evidencia que respaldan esta inferencia. Las fuentes de datos reunidas comprenden documentos operativos, materiales de capacitación, protocolos incidentales y reportes técnico-psicométricos. A continuación, se describen los resultados por garantía, supuestos, según dichas fuentes.

Tabla 29.
Estructura argumentativa para la inferencia de evaluación del ExIES

Conclusión de evaluación	Las puntuaciones del ExIES son un resumen preciso del desempeño relevante en las tareas del examen	
Garantía	**Suposiciones**	**Fuentes de datos**
G2.1. La administración del ExIES se realiza siguiendo altos estándares para garantizar la estandarización e imparcialidad, alineados con estándares reconocidos internacionalmente.	S2.1.1 El personal del equipo de evaluación está formado para asegurar la administración del examen según las pautas establecidas. S2.1.2 Los candidatos cuentan con materiales de preparación y práctica para familiarizarse con las condiciones del examen. S.2.1.3 Se emplean procedimientos de seguridad para el manejo del examen durante la aplicación. S2.1.4 Se proporcionan instrucciones claras a los candidatos sobre posibles consecuencias de deshonestidad durante el examen. S2.1.5 Se lleva a cabo un proceso estandarizado que permite las mismas condiciones de aplicación.	F2.1.1.1 Manual del aplicador (Pedroza Zúñiga et al.,2023i) F2.1.1.2 Manual del supervisor (Pedroza Zúñiga et al.,2023j) F2.1.1.3 Presentación de capacitación del aplicador y supervisor (Pedroza Zúñiga et al.,2023k) F2.1.2-4 Guía del sustentante (Pedroza Zúñiga et al.,2023l) F2.1.3-5 Protocolos para incidencias en caso de siniestro o emergencia del ExIES (Pedroza Zúñiga et al.,2023m) F2.1.3.2 Reporte de aplicación del Examen de Ingreso a la Educación Superior (ExIES) 2023-1 (Pedroza et al., 2023r) F2.1.5.1 Documentación de estadísticas y dificultad de ítems y personas, Reporte técnico 2023-1, Reporte Técnico 2023-2 (Pedroza Zúñiga et al., 2024a, 2024b)

G2.2 Los puntajes del ExIES son obtenidos mediante procedimientos que representan el grado de dominio de los sustentantes en cada una de las habilidades.	S2.2.1 La técnica de Rasch permite estimar la habilidad de los sustentantes basada en sus respuestas.	F2.2.1.1 Técnica Rasch, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
	S2.2.2 Se realiza un procedimiento operativo después de la estimación Rasch por cada área.	F.2.2.2.1 Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
G2.3 Los ítems del ExIES son seleccionados con base en su comportamiento en aplicaciones previas.	S2.3.1 Los ítems son probados antes de ser seleccionados para integrarlos en alguna forma.	F2.3.1.1 Pilotaje y evaluación de los resultados, Manual Técnico 2022- 2 (Pedroza Zúñiga et al., 2022)

Desarrollo de respaldos de la inferencia de evaluación

G2.1. La administración del ExIES se realiza siguiendo altos estándares para garantizar la estandarización e imparcialidad, alineados con estándares reconocidos internacionalmente

S2.1.1 El personal del equipo de evaluación está formado para asegurar la administración del examen según las pautas establecidas. El personal encargado de administrar el examen está capacitado conforme a lineamientos específicos, enfatizando la prevención de sesgos y errores. ExIES proporciona manuales específicos para aplicadores y supervisores, acompañados de presentaciones de capacitación y una Guía del Sustentante (Pedroza Zúñiga et al., 2023l). Asimismo, la Tabla 30 resume los aspectos documentados que expresan los aspectos sobre preparación del personal con el fin de reducir la variabilidad en la administración, fortaleciendo la validez argumental del examen.

Tabla 30.
Aspectos y evidencias documentales del ExIES

Aspecto	Evidencia documental	Descripción
Capacitación clara y específica del personal (aplicadores y supervisores)	Manual del Aplicador y Manual del Supervisor incluyen descripciones detalladas sobre las funciones y procedimientos que debe seguir el personal antes, durante y después de la aplicación.	Se establecen lineamientos precisos sobre cómo debe actuar el personal en cada fase de la administración del examen, eliminando ambigüedades.
Asignación clara de roles y responsabilidades	En los manuales se describen claramente las responsabilidades tanto de los aplicadores como de los supervisores, lo que garantiza una estandarización del proceso.	Se propone un proceso asignando tareas específicas a cada rol, lo que reduce la posibilidad de errores o confusión durante la aplicación.
Manejo estandarizado de materiales	El Manual del Supervisor detalla cómo manejar los materiales, desde la recepción hasta la devolución, asegurando un control estricto durante todo el proceso.	Al garantizar un manejo estricto de los materiales, se protege la integridad del examen y se asegura que todas las sedes sigan el mismo procedimiento.
Supervisión constante durante la aplicación	La presentación de capacitación y los manuales enfatizan la necesidad de que el personal supervise de manera continua y esté atento a cualquier irregularidad.	Existe un proceso de capacitación donde se procura la imparcialidad en la aplicación mediante una supervisión continua, lo que minimiza las posibles desviaciones del protocolo.
Protocolos de seguridad para la protección de los materiales	Los protocolos de seguridad descritos en el Manual del Aplicador y el Manual del Supervisor aseguran la protección de los materiales del examen y el cumplimiento de las normas.	Proveen medidas específicas para evitar incidentes que puedan comprometer la imparcialidad del examen, como el uso indebido de los materiales.

Nota. Elaboración propia basada en *Manual del aplicador del ExIES* (Pedroza Zúñiga et al.,2023i), el *Manual del supervisor del ExIES* (Pedroza Zúñiga et al.,2023j) y la *Presentación de capacitación para aplicadores y supervisores del ExIES* (Pedroza Zúñiga et al.,2023k).

S.2.1.2 Los candidatos cuentan con materiales de preparación y práctica para familiarizarse con las condiciones del examen. El segundo supuesto hace referencia a la disponibilidad de materiales adecuados que permitan a los candidatos prepararse efectivamente para el ExIES. Este aspecto es crucial, ya que dota a los sustentantes de herramientas esenciales para comprender claramente el formato del examen, identificar las áreas evaluadas y conocer los tipos de preguntas que enfrentarán.

La Guía del Sustentante (Pedroza Zúñiga et al., 2023l) cumple un papel central, pues proporciona una descripción exhaustiva de los contenidos evaluados, acompañada por ejemplos representativos de preguntas y recomendaciones estratégicas sobre cómo administrar adecuadamente el tiempo durante la prueba. Al ofrecer una guía detallada, el ExIES facilita la estandarización e imparcialidad en la administración del examen; véase la Tabla 31, que sintetiza los aspectos fundamentales y las evidencias documentales disponibles para respaldar este supuesto: descripción de las áreas evaluadas, ejemplos de preguntas, consejos estratégicos para la toma del examen y recomendaciones sobre el día del examen.

Tabla 31.
Aspectos y evidencias documentales sobre la suposición 2.1.2

Aspecto defendido	Evidencia documental	Descripción
Familiarización con las condiciones del examen	Guía del sustentante	Describe las áreas evaluadas, ejemplos de preguntas, y consejos para la toma del examen, lo que permite a los candidatos practicar.
Ejemplos de preguntas para la práctica	Guía del sustentante	Proporciona ejemplos de preguntas de las áreas de Lectura, Lengua Escrita y Matemáticas, lo que ayuda a los sustentantes a practicar.
Recomendaciones para el día del examen	Guía del sustentante	Detalla qué llevar, qué no llevar, y cómo manejar el examen, lo que minimiza errores y aumenta la confianza de los candidatos.
Consejos sobre la administración del tiempo y la lectura de instrucciones	Guía del sustentante	Ofrece estrategias para administrar el tiempo y leer con atención, maximizando las posibilidades de éxito en el examen.

Nota. Elaboración propia basada en la *Guía del Sustentante del ExIES* (Pedroza Zúñiga et al., 2023l).

S.2.1.3 Se emplean procedimientos de seguridad para el manejo del examen durante la aplicación. El tercer supuesto enfatiza la relevancia de contar con protocolos y procedimientos de seguridad que garanticen la integridad del examen y prevengan accesos o manipulaciones no autorizadas, ya que cualquier falla en este ámbito puede afectar la validez y la equidad de los resultados. Los manuales del ExIES incluyen la capacitación de supervisores para identificar y actuar ante posibles violaciones, fortaleciendo la confianza en los resultados obtenidos. Estos manuales parten, tanto de los estándares internacionales (AERA et al., 2014) como de literatura especializada (Haladyna & Rodríguez, 2013), que coinciden en que la protección de los materiales y la vigilancia constante durante la aplicación son esenciales para prevenir el fraude y asegurar la imparcialidad del proceso.

El ExIES tiene procedimientos de seguridad documentados, así como manuales del aplicador y supervisor, para garantizar la integridad del proceso de aplicación del examen. Tal como se resume en la Tabla 32, estos procedimientos incluyen la revisión previa de las sedes, el control de acceso a los materiales desde su recepción hasta su devolución, protocolos claros para detectar y prevenir fraudes, la vigilancia constante por parte de supervisores durante toda la administración y la protección post-examen mediante la devolución y resguardo estandarizado de los materiales. Estas acciones, junto con la capacitación homogénea del personal y la estandarización de procedimientos en todas las sedes, buscan evitar accesos no autorizados, minimizar riesgos de manipulación y asegurar condiciones justas para todos los sustentantes, en cumplimiento con los lineamientos institucionales (Pedroza Zúñiga et al., 2023i, 2023j).

Tabla 32.
Aspectos y evidencias documentales sobre la suposición 2.1.3

Aspecto defendido	Evidencia documental	Descripción y argumentación
Control de acceso a los materiales del examen	Manual del aplicador y Manual del supervisor; Reporte de aplicación, sección 2.2 y 2.4	Se detalla el manejo seguro desde la reproducción, empaquetado, traslado, resguardo en sedes, y conteo de materiales. El resguardo incluye espacios exclusivos y etiquetado de seguridad (p. 9-10). Así, se evita acceso no autorizado en cada fase y se protege la validez de la prueba, como exige AERA, APA y NCME (2014).
Protocolos para detección y prevención de fraudes	Manual del aplicador y Manual del supervisor; Reporte de aplicación, sección 3.1 y 4.5	Se implementan medidas como la revisión de dispositivos electrónicos, reglas claras para los sustentantes, y reportes de incidencias. El personal es capacitado para detectar y actuar ante intentos de fraude (p. 13, 18). Esto asegura imparcialidad y protección ante manipulaciones.
Monitoreo constante durante la aplicación	Manual del supervisor; Reporte de aplicación, sección 3.1 y 4.1	Los supervisores vigilan continuamente el proceso y a los sustentantes. Se describen funciones, rutas de comunicación y procedimientos de registro de incidentes. Se reportaron casos de intento de fraude que fueron detectados y documentados (p. 12-13, 18).
Protección posexamen	Manual del aplicador y Manual del supervisor; Reporte de aplicación, sección 2.5	Tras la aplicación, los materiales se recuperan, cuentan y resguardan bajo protocolos estandarizados, firmando acuses de recibido y asegurando que no haya manipulación posterior (p. 11).

Capacitación y homologación de procedimientos	Manuales; reporte de aplicación, secciones 3.2 y 3.3	Todo el personal involucrado recibe capacitación específica sobre protocolos de seguridad, manejo de materiales y actuación ante incidencias, incluyendo cursos y materiales de apoyo (p. 12-13).
Mecanismos de estandarización	Manuales y reporte, sección 4.2	Todas las sedes aplican los mismos procedimientos, horarios y reglas para garantizar igualdad de condiciones y minimizar riesgos de fuga o manipulación de información (p. 16).

Nota. Elaboración propia basada en *Reporte de aplicación del examen de ingreso a la educación superior (ExIES) 2023-1* (Pedroza et al., 2023z), y verificado según el *Manual del aplicador del ExIES* (Pedroza Zúñiga et al., 2023i), el *Manual del supervisor del ExIES* (Pedroza Zúñiga et al., 2023j).

Además de cumplir con los protocolos descritos de seguridad, se cuenta con el Reporte de aplicación del Examen de Ingreso a la Educación Superior (ExIES) 2023-1 (Pedroza et al., 2023r). Este abarca desde la planeación y gestión de recursos (financieros, materiales y humanos), pasando por la reproducción, resguardo y entrega segura de los materiales, hasta la capacitación del personal, la ejecución de la prueba y el análisis de incidencias ocurridas durante el proceso.

Según la Tabla 33, entre las más relevantes se encuentran casos de fraude y faltas graves, como intentos de robo del instrumento, uso no autorizado de dispositivos electrónicos y toma de fotografías del examen, que derivaron en la cancelación de la prueba para ciertos sustentantes. Asimismo, se presentaron numerosos errores de registro y logística, incluyendo aspirantes no registrados correctamente, firmas y datos faltantes o duplicados, así como problemas en el llenado de hojas de respuesta y errores en la distribución de materiales.

Tabla 33.
Principales incidencias según su categoría en la aplicación del ExIES (2023-1)

Categoría	Incidencia	Número/Descripción
Fraudes y faltas graves	Cancelación de la prueba por fraude, intento de robo, uso de celular, fotos del examen	9 casos identificados
Errores de registro	Sustentantes sin versión de prueba, fichas duplicadas, hojas de respuesta en blanco	31 sin versión, 2 duplicadas, 3 en blanco
Conductas inapropiadas	Omisión de instrucciones, inicio irregular de prueba, fotos sin autorización	Casos aislados reportados
Problemas logísticos	Daños materiales (hojas, cajas), devoluciones incompletas, problemas de autenticación	Reportes diversos
Intentos de soborno	Intento de soborno a aplicador	1 caso reportado
Errores materiales	Cuadernillos con errores de impresión	30 cuadernillos (~0.1 %)
Otros	Aplicadores con doble/triple lista, familiares en sedes, errores menores diversos	Casos varios

Nota. Elaboración propia basado en el *Reporte de aplicación del examen de ingreso a la educación superior (ExIES) 2023-1* (Pedroza et al., 2023r).

S2.1.4 Se proporcionan instrucciones claras a los candidatos sobre posibles consecuencias de deshonestidad durante el examen. El objetivo de este supuesto es prevenir comportamientos que puedan comprometer la integridad del proceso de evaluación, así como proteger la imparcialidad y la validez de los resultados del examen. Las instrucciones relativas a la deshonestidad académica en el ExIES están claramente establecidas en la Guía del Sustentante, así como en los manuales del aplicador y supervisor (Pedroza Zúñiga et al., 2023i, 2023j, 2023l). Según se detalla en la Tabla 34, estos documentos explican de manera precisa qué conductas se consideran deshonestas, tales como el uso de dispositivos electrónicos o copiar durante el examen, y las consecuencias de incurrir en ellas, que incluyen la expulsión inmediata y la anulación de los resultados, e incluso la posibilidad de ser vetado de futuras aplicaciones.

Tabla 34.
Evidencias documentales sobre comportamientos deshonestos en la aplicación

Aspecto Defendido	Evidencia documental	Descripción
Definición de comportamientos deshonestos	Guía del sustentante, Manual del aplicador y supervisor	Claramente detalla las acciones que constituyen deshonestidad, como el uso de dispositivos electrónicos o la copia entre sustentantes.
Consecuencias de la deshonestidad	Guía del sustentante, Manual del aplicador y supervisor	Se especifica que la participación en conductas deshonestas llevará a la expulsión inmediata y la anulación de los resultados.
Procedimientos para la detección de deshonestidad	Manual del supervisor	Los supervisores están capacitados para identificar comportamientos sospechosos y reportar incidentes siguiendo protocolos específicos.
Recordatorios previos al examen	Manual del aplicador y supervisor	Antes de la aplicación del examen, los aplicadores y supervisores deben recordar a los sustentantes las consecuencias de la deshonestidad.

Nota. Elaboración propia con base en la *Guía del sustentante del ExIES* (Pedroza Zúñiga et al., 2023l), el *Manual del aplicador del ExIES* (2023i) y el *Manual del supervisor del ExIES* (2023j).

La sección de Comportamientos Prohibidos de la *Guía del Sustentante del ExIES* (Pedroza Zúñiga et al., 2024, p. 24) presenta varias fortalezas, especialmente en términos de claridad y especificidad al definir qué conductas son inaceptables durante el examen. Por ejemplo, se prohíbe explícitamente el uso de dispositivos electrónicos y la divulgación del contenido del examen, lo cual es reforzado con advertencias sobre la cancelación del examen y la posible intervención de las autoridades, lo que subraya la gravedad de estas infracciones. .

S2.1.5 Se lleva a cabo un proceso estandarizado que permite las mismas condiciones de aplicación. El objetivo de este supuesto es garantizar que todos los sustentantes presenten el ExIES bajo condiciones uniformes, independientemente de la sede o del momento en que lo realicen. Los manuales del aplicador y del supervisor (Pedroza Zúñiga

et al., 2023i, 2023j), véase la Tabla 35, describen de manera detallada los pasos a seguir para asegurar que la administración del examen sea uniforme en todas las sedes. Esto incluye instrucciones específicas sobre la preparación del aula, la distribución y recolección de materiales y el control del tiempo durante el examen.

Tabla 35.
Evidencias documentales sobre condiciones de aplicación e imparcialidad (S2.1.5)

Aspecto defendido	Evidencia documental	Descripción
Preparación de las condiciones del aula	Manual del aplicador y manual del supervisor	Los manuales proporcionan instrucciones sobre cómo preparar las aulas de manera uniforme, asegurando que las condiciones ambientales y físicas sean adecuadas.
Entrega y recolección de materiales	Manual del supervisor	Se detallan los procedimientos para la distribución y recolección de los materiales del examen de manera estandarizada en todas las sedes.
Control del tiempo y monitoreo constante	Manual del aplicador y manual del supervisor	Se asegura que el tiempo asignado sea el mismo para todos los sustentantes, monitoreando su cumplimiento de manera rigurosa.
Protocolos de seguridad para mantener la imparcialidad	Manual del supervisor	Los protocolos de seguridad garantizan que todos los sustentantes enfrenten las mismas condiciones sin interrupciones ni distracciones durante el examen.

Nota. Elaboración propia basada en *Manual del aplicador del ExIES* y el *Manual del supervisor del ExIES* (Pedroza Zúñiga et al.,2023i, 2023j).

Al final de la aplicación se realiza un análisis y reporte de seguimiento (Pedroza et al., 2023z) para optimizar el proceso de aplicación del ExIES, enfocándose en fortalecer tanto la logística como la gestión de recursos humanos y materiales. Entre las principales recomendaciones (véase Tabla 36) se encuentran el perfeccionamiento de los protocolos ante imprevistos, la mejora en la calidad y manejo de los materiales, y el diseño de capacitaciones más claras y prácticas para el personal.

Tabla 36.
Síntesis de propuestas de mejora para el proceso de aplicación del ExIES 2023-1.

Área de mejora	Propuesta
Planificación y logística	Mayor rigor ante imprevistos y elaboración de protocolos claros; más tiempo entre turnos de aplicación; tiempo adicional para conteo de materiales.
Gestión de materiales	Uso de cajas más resistentes para el traslado; considerar nuevos proveedores para la reproducción de materiales.
Capacitación del personal	Mejorar el curso de capacitación para aplicadores, incluir materiales audiovisuales y establecer mecanismos de evaluación de desempeño.
Gestión de recursos humanos	Asegurar personal calificado en todas las sedes; incorporar más supervisores y personal de apoyo para actividades logísticas.
Comunicación y estandarización	Sincronizar instrucciones y procesos entre equipos responsables (ExIES y DSEyGE) para reducir confusión.
Participación de familiares	Establecer estrategias para limitar el acceso de familiares a las áreas de aplicación y mejorar el control de flujo de personas.
Inclusión tecnológica y audiovisual	Desarrollar videos instructivos para el llenado de hojas de respuesta y reforzar la comunicación con sustentantes.

Nota. Elaboración propia basada en *Reporte de aplicación del examen de Ingreso a la Educación Superior (ExIES) 2023-1* (Pedroza et al., 2023r).

G2.2 Los puntajes del ExIES son obtenidos mediante procedimientos que representan el grado de dominio de los sustentantes en cada una de las habilidades.

S2.2.1 La técnica de Rasch permite estimar la habilidad de los sustentantes basada en sus respuestas. En el ExIES se emplea la técnica Rasch para estimar la habilidad de los sustentantes (Pedroza Zúñiga et al., 2024a), —y, como se menciona en su Reporte Técnico— dicha técnica es reconocida por ser un modelo logístico unidimensional que asume que la probabilidad de una respuesta correcta está determinada por la diferencia entre la habilidad del sustentante y la dificultad del ítem (Wright y Stone, 1979; Tristán-López, 1998). Según Bond y Fox (2015), este modelo proporciona estimaciones invariantes de la habilidad del sustentante y la dificultad del ítem, lo que significa que las estimaciones son consistentes

y pueden compararse entre diferentes poblaciones y diferentes ítems de manera justa. En la Tabla 37 se muestra la descripción de las evidencias psicométricas para esta suposición.

Tabla 37.
Descripción de evidencias psicométricas

Aspecto defendido	Descripción
Estimación precisa de habilidad mediante Rasch	Describe cómo el uso del modelo Rasch permite estimar de forma confiable la habilidad de los evaluados, proporcionando una escala común de medición y diferenciando múltiples niveles de desempeño.
Ajuste de los ítems (*Infit, Outfit*)	Presenta los valores de *Infit* y *Outfit* para cada ítem, mostrando que la mayoría se ubica dentro de rangos considerados aceptables. Esto indica un ajuste adecuado de los ítems al modelo Rasch y una medición coherente.
Interpretación de la discriminación de los ítems	Expone la capacidad discriminativa de los ítems al evidenciar que la mayoría presentan índices de discriminación adecuados. Esto demuestra que los ítems diferencian eficazmente a quienes tienen altos y bajos niveles de la habilidad medida.

Nota. Elaboración propia basada en *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte Técnico* (Pedroza Zúñiga et al., 2024a).

La Tabla 54 resume, por área, información descriptiva y del objetivo de la prueba (Pedroza Zúñiga et al., 2024a): número de ítems y sujetos, tasa media de acierto (p) —que en el contexto de ítems dicotómicos varía entre 0 y 1 y cuya cercanía a 0.5 suele indicar buen ajuste del nivel de dificultad al conjunto de sustentantes— y el conteo de ítems por tramos de p (rango de dificultad). Estos conteos permiten evaluar el ajuste entre la dificultad de los ítems y la habilidad de la muestra (por ejemplo, una concentración de ítems en tramos muy altos o muy bajos indicaría mal targeting y pérdida de información en ciertos rangos de habilidad). La media de dificultad ($\bar{p}$) ofrece una visión agregada: valores más bajos implican ítems más difíciles en promedio.

Tabla 38.
Resumen de parámetros psicométricos por área del ExIES 2023-1

Parámetro	Lectura	Lengua Escrita	Matemáticas
N.º de ítems	200	200	188
N.º de sujetos	28 205	28 205	28 205
Tasa de acierto promedio (%)	54.5	49.9	37.4
Dificultad media ($\bar{p}$)	0.51	0.54	0.58
Rango de dificultad			
[0.0 – 0.1)	0	0	0
[0.1 – 0.2)	0	0	0
[0.2 – 0.3)	1	0	0
[0.3 – 0.4)	17	4	1
[0.4 – 0.5)	66	40	8
[0.5 – 0.6)	91	114	107
[0.6 – 0.7)	24	39	71
[0.7 – 0.8)	1	3	1
[0.8 – 0.9)	0	0	0
[0.9 – 1.0]	0	0	0

Nota. Reimpreso de *examen de ingreso a la educación superior (ExIES) 2023-1*: Reporte técnico (Pedroza Zúñiga et al., 2024a, p. 26).

En la Tabla 39 se enlistan los ítems con estadísticas de ajuste (Infit/ Outfit MNSQ) fuera del rango considerado aceptable para este estudio (0.70–1.30). En el marco Rasch, los índices MSQ > 1 indican que un ítem es menos predecible de lo esperado por el modelo (mayor ruido o multidimensionalidad residual), mientras que MSQ < 1 sugiere respuestas demasiado predecibles (posible redundancia o sobreajuste). Los valores t (Zstd) complementan la interpretación estadística de MSQ: t entre −2 y 2 se considera razonablemente consistente con el modelo; valores fuera de ese intervalo señalan significación estadística del desajuste. Ítems con MSQ y/o t fuera de los límites deben investigarse cualitativa y cuantitativamente (se deben revisar, p. ej., enunciado, distractores, clave, contenido cultural/lingüístico o condiciones de administración) y, dependiendo del hallazgo, reformularse, pilotearse o descartarse.

En el área de Lectura, los ítems fuera de rango pertenecen mayormente a las subversiones 1, 2, 3, 5, 6, 8 y 9. Generalmente, estos ítems presentan Infit u Outfit MNSQ superiores a 1.30, lo cual indica cierta desviación respecto del comportamiento ideal esperado por el modelo Rasch. En el caso de Lengua Escrita, la Tabla 39 muestra que solo unas cuantas subversiones (2, 3, 5, 6 y 9) exhiben ítems con Infit/Outfit MNSQ fuera de los criterios, y, al igual que en Lectura, la mayoría de estos ítems mantienen índices de discriminación por encima de 0.20. Esto significa que, si bien su ajuste al modelo podría mejorarse, siguen cumpliendo con un mínimo requerido para discriminar entre diferentes niveles de desempeño. Por su parte, en Matemáticas solo la subversión 1 presenta un ítem (el ítem 57) fuera de rango (Infit/Outfit 0.70–1.30). Este hallazgo refuerza la idea de que, en general, la mayoría de los ítems del área de matemáticas, se ajustan adecuadamente al modelo Rasch, y aquellos pocos que no lo hacen podrían revisarse para futuros ajustes o descartes.

Tabla 39.
Ítems con valores por área fuera del rango aceptable (0.70–1.30).

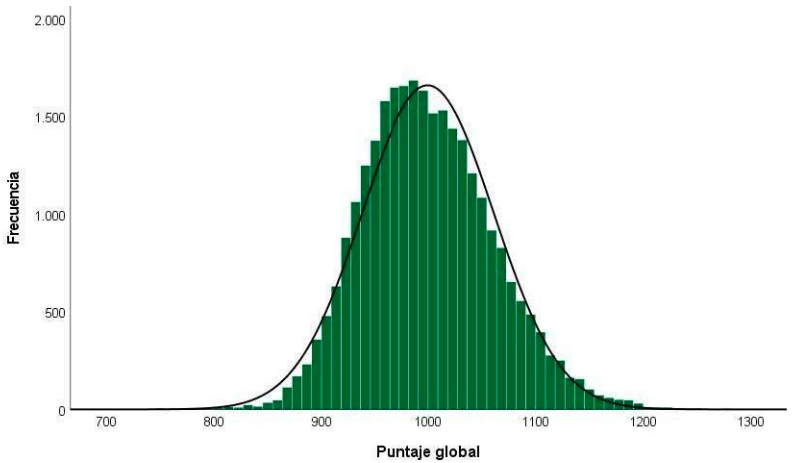
Área	Subversión	Ítem	Infit MNSQ	Outfit MNSQ
Lectura	1	25	1.25	1.36
	2	25	1.25	1.37
	3	25	1.21	1.34
	5	25	1.23	1.34
	5	45	1.21	1.41
	6	13	1.11	1.34
	6	46	1.15	1.65
	8	13	1.11	1.31
	9	13	1.12	1.32
Lengua Escrita	2	46	1.11	1.45
	3	40	1.13	1.37
	5	43	1.15	1.34
	6	47	1.13	1.41
	9	43	1.10	1.32
Matemáticas	1	57	1.22	1.31

Nota. Elaboración propia basada en *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte Técnico* (Pedroza Zúñiga et al., 2024a).

S2.2.2 Se realiza un procedimiento operativo después de la estimación Rasch por cada área. Con el fin de que los puntajes comunicados representen de manera válida el nivel de los sustentantes y sean comparables entre aplicaciones, el ExIES aplica un procedimiento operativo posterior a la estimación Rasch por cada área (Comprensión Lectora, Lengua Escrita y Matemáticas), tal como se documenta en el Reporte técnico (Pedroza Zúñiga et al., 2024a). En este sentido, según César Gómez (comunicación personal, 11 de agosto de 2025) —técnico del ExIES—, se realiza:

a. Depuración previa de reactivos: Se excluyen del cómputo los ítems que muestran mal funcionamiento (p. ej., correlación punto-biserial negativa o desajustes fuera de criterios), debido a que degradan la consistencia interna y agregan ruido a la medición; antes de retirarlos, se realiza revisión experta de clave y codificación, y toda decisión se documenta en la base histórica.
b. Exclusión de ítems en pilotaje: Los ítems piloto se usan para análisis de calidad, pero no computan en el puntaje operativo.
c. Escalamiento lineal: Las estimaciones Rasch por área se transforman en Winsteps fijando media 1000 y desviación estándar 100 (parámetros UIMEAN=1000 y USCALE=100), preservando el orden y las distancias relativas de la métrica latente y facilitando la interpretabilidad y comparabilidad intercohortes.
d. Reglas de reporte: Por criterios de comunicación institucional, cada área se reporta en el rango 700–1300; los valores fuera de este intervalo se truncan (700 por debajo; 1300 por arriba), sin modificar la estimación de base. La mayoría de los sustentantes se concentra en torno a la media y dentro de unas cuantas desviaciones estándar; esta distribución se ilustra en la Figura 13.
e. Puntaje global: Se obtiene como el promedio aritmético de los tres puntajes de área ya escalados y ajustados, y es esta escala la que se comunica para la selección de estudiantes.

Figura 13.
Frecuencia de la puntuación del ExIES



Nota. Reimpreso de *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a, p. 26). N = 28,205 (convocatorias 2023-2 y 2024-1). La línea vertical negra indica la media (1000); las líneas punteadas muestran ±1 desviación estándar (900 y 1100); las líneas rojas indican las reglas de truncamiento reportadas (700 y 1300).

Este procedimiento alinea la métrica de medición con una escala operativa estable, dejando trazabilidad de depuración, escalamiento y reporte en los reportes técnicos y en la base de mantenimiento del banco de ítems (en el reporte se cita a AERA et al., 2014; Bond y Fox, 2015; Kolen y Brennan, 2014; Tavakol y Dennick, 2011; Haladyna y Rodríguez, 2013).

G2.3. Los ítems del ExIES son seleccionados con base en su comportamiento en aplicaciones previas

S.2.3.1 Los ítems son probados antes de ser seleccionados para integrarlos en alguna forma. En el ExIES se emplea un proceso de pilotaje que permite someter los ítems a una evaluación previa antes de su inclusión en el examen operativo. El pilotaje de ítems, que se publicó en el manual técnico del 2022-2 (Pedroza Zúñiga et al., 2022), consistió en administrar los ítems a 2,210 estudiantes (de los tres campus de la UABC) para evaluar su comportamiento en condiciones de examen. El análisis deta-

llado de estas métricas permitió determinar si los ítems cumplían con los criterios necesarios para ser seleccionados en el examen operativo, que sería el 2023-1. Los ítems que no cumplieron con estos estándares son revisados y, en algunos casos, descartados para asegurar la integridad de la prueba; este procedimiento se realiza de forma semestral, dando un seguimiento continuo; sin embargo, no se explicita en los documentos consultados (Pedroza Zúñiga et al., 2022; Pedroza Zúñiga et al., 2024a).

Por último, una práctica recomendada en la gestión de instrumentos evaluativos es asegurar su actualización y mejora continua. En el caso del ExIES, la revisión sistemática de los parámetros Rasch y el análisis de distractores para todos los ítems de las distintas subversiones (tanto evaluativas como piloto) permitió identificar un porcentaje de reactivos que presentan áreas de oportunidad (véase Figura 14): un 22 % en lectura, 19.5 % en lengua escrita y 10.6 % en matemáticas (Pedroza Zúñiga et al., 2024a).

Evaluación de la inferencia de evaluación

La valoración global de la inferencia de Evaluación, presentada en la Tabla 40, alcanzó un desempeño de 86.9 %, lo que refleja un nivel alto de cumplimiento de los Estándares en los distintos componentes de la administración de la prueba. Este resultado se deriva del análisis de tres garantías.

Tabla 40.
Evaluación de la inferencia de evaluación

Suposición	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S.2.1.1 Formación del personal	El personal encargado de la administración del examen debe estar capacitado y formado según las pautas establecidas para asegurar estandarización e imparcialidad (Estándar 6.1, 6.5).	4	3	4	11 (91.7 %, Alta)
S2.1.2 Materiales de preparación	Los sustentantes deben contar con materiales de preparación para familiarizarse con las condiciones del examen, asegurando que tengan el mismo acceso a la información necesaria (Estándar 8.1).	3	4	3	10 (83.3 %, Moderada)
S2.1.3 Seguridad en el examen	Los procedimientos de seguridad deben estar implementados para asegurar la protección de los contenidos del examen y la integridad del proceso (Estándar 6.4, 7.2, 9.3).	4	3	3	10 (83.3 %, Moderada)
S2.1.4 Instrucciones sobre deshonestidad	Se proporcionan instrucciones claras a los sustentantes sobre las consecuencias de la deshonestidad académica, asegurando que entiendan las reglas del examen (Estándar 8.2).	3	4	3	10 (83.3 %, Moderada)

S2.1.5 Estandarización de condiciones	Las condiciones de administración del examen deben ser homogéneas para todos los sustentantes, garantizando la igualdad de oportunidades (Estándar 6.1, 6.2).	3	3	4	10 (83.3 %, Moderada)
S2.2.1 Técnica de Rasch	Los puntajes deben ser obtenidos utilizando la técnica de Rasch, que permite estimar la habilidad de los sustentantes con base en sus respuestas y eliminar el efecto de las características del ítem (Estándar 5.1, 5.2, técnicas de estimación apropiadas).	3	4	3	10 (100 %, Moderada)
S2.3.1 Pruebas piloto de ítems	Cada ítem se somete a un pilotaje para analizar su dificultad y ajuste psicométrico, validando que los ítems realmente midan el desempeño buscado antes de formar parte del examen operativo (Estándar 4.1).	3	4	3	10 (83.3 %, Moderada)
	Global	24 (85.7 %, Moderada)	25 (89.2 %, Alta)	24 (85.7 %, Moderada)	73 / 84 (86.9 %, Moderada)

En términos operativos, la inferencia de evaluación se sostuvo cuando existieron manuales, protocolos y evidencia de control de aplicación. Para replicarla, se recomendó estandarizar instrucciones, tiempos y condiciones; capacitar a aplicadores y supervisores; registrar incidencias y acciones correctivas; y documentar adaptaciones razonables para sustentantes con necesidades específicas. En el caso ExIES, estos productos permitieron defender que las variaciones de procedimiento se controlaron; en otros contextos, su ausencia suele traducirse en dudas sobre la justicia de los puntajes.

Capítulo 5

Inferencia de generalización

La inferencia de Generalización examinó la consistencia de las puntuaciones: si un puntaje reflejó, de manera suficientemente estable, el rendimiento del sustentante y no fluctuaciones atribuibles al muestreo de ítems, a la forma aplicada o al error de medición. En el EBA, esta inferencia suele sostenerse con evidencia de confiabilidad, precisión y comparabilidad entre formas (AERA et al., 2014; Chapelle, 2021; Kane, 2013).

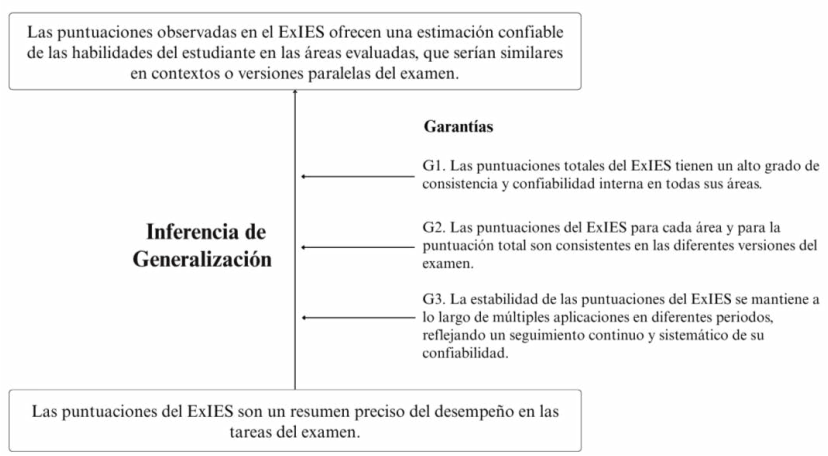
Cuando la prueba se aplicó en modalidades múltiples (por ejemplo, distintas fechas o versiones), también se revisaron procedimientos de equiparación, TRI o indicadores de funcionamiento por forma. En la revisión del ExIES 2023-2, la pregunta guía fue: ¿en qué medida las puntuaciones fueron consistentes entre áreas y formas? Las Tablas 39 a 45 reúnen los respaldos psicométricos disponibles y sus reservas.

Dado el propósito de guía, se presentan indicadores y decisiones clave (p. ej., consistencia interna, estabilidad de la escala, DIF y análisis Rasch/TRI) sin desplegar todas las salidas por ítem o por forma. En aplicaciones institucionales, estos anexos suelen concentrarse en reportes técnicos para su consulta especializada. Como extensiones posibles, la inferencia de generalización puede fortalecerse con estudios de teoría de la generalizabilidad, diseños de replicación (test-retest o formas paralelas), análisis de error estándar a lo largo de la escala y procedimientos de equiparación cuando existen múltiples formas (Brennan, 2006; Bond y Fox, 2015; Cronbach et al., 1972; Kolen y Brennan, 2014).

Definición de las garantías, supuestos y fuentes de la inferencia de generalización

La conclusión de generalización, véase la Figura 15, en conjunto con sus garantías, busca asegurar que los puntajes obtenidos sean consistentes y confiables en diversas versiones y aplicaciones del examen.

Figura 15.
Argumento de la inferencia de generalización como parte de la validez del argumento



Por otro lado, la Tabla 41 integra, de manera estructurada, las garantías esenciales, los supuestos y las fuentes de evidencia que respaldan la estabilidad, consistencia y comparabilidad de las puntuaciones obtenidas en las distintas áreas y aplicaciones del examen. Las fuentes de datos, en general, son claramente cuantitativos, es decir, análisis de confiabilidad, parámetros psicométricos, como equiparación. Con esto en cuenta, a continuación, se presentan estos resultados.

Tabla 41.
Estructura argumentativa para la inferencia de generalización del ExIES

Conclusión de generalización	Las puntuaciones observadas en el ExIES ofrecen una estimación confiable de las habilidades del estudiante en las áreas evaluadas, que serían similares en contextos o versiones paralelas del examen	
Garantía	Suposiciones	Fuentes de datos
G3.1 Las puntuaciones totales del ExIES tienen un alto grado de consistencia y confiabilidad interna en todas sus áreas.	S.3.1.1 El coeficiente de confiabilidad promedio para cada área del ExIES se mantiene en un rango aceptable.	F3.1.1 Confiabilidad (Alfa de Cronbach), Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
G3.2 Las puntuaciones del ExIES para cada área y para la puntuación total son consistentes en las diferentes versiones del examen.	S3.2.1 Las formas de los exámenes tienen la misma dificultad para todos los sustentantes.	F3.2.1.1 Parámetros psicométricos R, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) F3.2.1.2 Confiabilidad (Alfa de Cronbach), Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) F3.2.1.3 Equiparación, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
G3.3 La estabilidad de las puntuaciones del ExIES se mantiene a lo largo de múltiples aplicaciones en diferentes periodos, reflejando un seguimiento continuo y sistemático de su confiabilidad.	S3.3.1 La estabilidad de la confiabilidad del ExIES se revisa en cada nuevo periodo de aplicación, verificando la reproducibilidad de los coeficientes de consistencia interna y la comparabilidad de los puntajes obtenidos.	F3.3.1.1 Confiabilidad (Alfa de Cronbach), Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)

Desarrollo de respaldos de la inferencia de generalización

G3.1 Las puntuaciones totales del ExIES tienen un alto grado de consistencia y confiabilidad interna en todas sus áreas.

S3.1.1 El coeficiente de confiabilidad promedio para cada área del ExIES se mantiene en un rango aceptable. En la Tabla 42 se presentan los coeficientes de confiabilidad de Alfa de Cronbach para las áreas de Lectura, Lengua Escrita, Matemáticas, y el total de las subversiones en las Formas A y B; a partir de la población participante de 28,205 aspirantes. El área de Lectura muestra un coeficiente promedio de .73, lo cual está dentro del rango aceptable (.70) según Nunnally y Bernstein (1994). Lengua Escrita alcanza un promedio de .77, mientras que Matemáticas presenta una confiabilidad más alta con .82, lo que refleja una mejor consistencia interna en esta última área.

Tabla 42.
Coeficientes de confiabilidad por área y forma del ExIES 2023-1

	Lectura	**Lengua escrita**	**Matemáticas**	**Total**
No de ítems	36	36	50	121
Alfa de Cronbach				
Forma A	.74	.77	.84	.89
Forma B	.72	.77	.81	.87
Global	.73	.77	.82	0.88

Nota. Adaptado de la Tabla 9 del *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a, p. 27).

En cuanto al Alfa de Cronbach total, el promedio es de 0.88, indicando un nivel elevado de confiabilidad en la prueba. Estos resultados muestran que las puntuaciones totales del ExIES mantienen un alto grado de consistencia y confiabilidad interna.

G3.2 Las puntuaciones del ExIES para cada área y para la puntuación total son consistentes en las diferentes versiones del examen

S3.2.1 Las formas de los exámenes tienen la misma dificultad para todos los sustentantes. Esta suposición revisa si las formas mantienen la misma

dificultad. Para verificar esta afirmación, se llevó a cabo un proceso de equiparación utilizando una calibración concurrente de ítems ancla y diferenciadores entre las formas A y B del ExIES. La Tabla 43 presenta la estructura de la base de datos utilizada para este proceso, donde los ítems ancla (AB) están presentes en ambas formas, mientras que los ítems diferenciadores son exclusivos de cada forma. Este enfoque garantiza que las formas del examen puedan ser comparadas y ajustadas en términos de dificultad.

Tabla 43.
Estructura de la base de datos para la calibración concurrente en el ExIES 2023-1

	Ítems ancla					**Ítems de la Forma A**					**Ítems de la Forma B**				
Ficha	AB1	AB2	AB3	AB4	AB5	A6	A7	A8	A9	A10	B6	B7	B8	B9	B10
1	x	x	x	x	x	x	x	x	x	x					
2	x	x	x	x	x	x	x	x	x	x					
3	x	x	x	x	x	x	x	x	x	x					
4	x	x	x	x	x	x	x	x	x	x					
5	x	x	x	x	x	x	x	x	x	x					
6	x	x	x	x	x						x	x	x	x	x
7	x	x	x	x	x						x	x	x	x	x
8	x	x	x	x	x						x	x	x	x	x
9	x	x	x	x	x						x	x	x	x	x
10	x	x	x	x	x						x	x	x	x	x

Nota. Adaptado de la Tabla 10 del *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a). AB indica ítem ancla presente en ambas formas (A y B); A y B indican ítems diferenciadores exclusivos de cada forma.

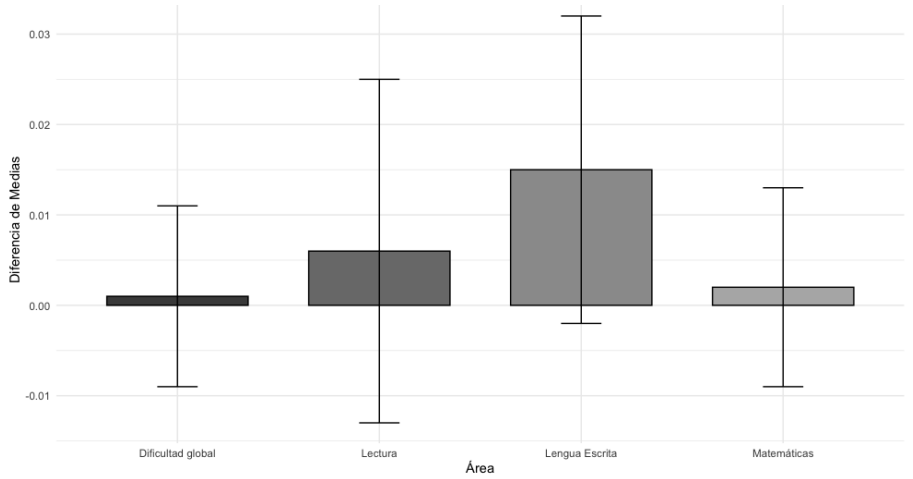
Posteriormente, se realizó una prueba *t* para muestras independientes sobre la dificultad de las formas A y B, cuyos resultados se muestran en la Tabla 44 y en la Figura 16, que corresponden al ExIES 2023-1. Los valores de p en las áreas de Lectura ($p = 0.734$), lengua escrita ($p = 0.380$), y matemáticas ($p = 0.862$), así como en la dificultad global ($p = 0.864$), indican que no existen diferencias significativas entre las formas del examen en términos de dificultad. Estos resultados respaldan la suposición de que las diferentes versiones del ExIES son equivalentes en dificultad, lo que permite que las puntuaciones de los sustentantes sean comparables independientemente de la versión que presenten.

Tabla 44.
Resultados de la prueba t para comparar dificultades entre formas A y B

Área	p valor	Diferencia de medias	Diferencia de error estándar
Lectura	0.734	0.006	0.019
Lengua escrita	0.380	0.015	0.017
Matemáticas	0.862	0.002	0.011
Dificultad global	0.864	0.001	0.010

Nota. Adaptado de la Tabla 11 del *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a, p. 28).

Figura 16.
Diferencias de medias y errores estándar por área entre formas A y B 2023-1.



Nota. Elaboración propia con base en los resultados de la prueba t reportados en el *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a, p. 28).

La Tabla 45 presenta los resultados de la prueba *t* para muestras independientes del ExIES 2023-2, comparando los puntajes obtenidos en los componentes evaluativos de las formas A y B del ExIES en tres áreas (Lectura, Lengua Escrita y Matemáticas), además del puntaje global. En el área de Lectura, se observa un *p* valor de 0.688, lo que indica que no existen diferencias significativas entre los puntajes de las formas A y B, con una diferencia de medias de 0.363. De manera similar, el área de

Matemáticas presenta un *p* valor de 0.730, también indicando la ausencia de diferencias significativas con una diferencia de medias de 0.343. Sin embargo, en el área de Lengua Escrita, el *p* valor es de 0.045, lo que sugiere una diferencia estadísticamente significativa entre las formas A y B, con una diferencia de medias de 1.925. A nivel global, no se observan diferencias significativas entre las formas, ya que el *p* valor es de 0.580 y la diferencia de medias es de 0.406. No obstante, en el *Reporte Técnico 2023-1* se menciona que no se encontraron diferencias significativas.

Tabla 45.
Resultados de la prueba t para comparar los puntajes entre las formas A y B 2023-2

Área	p valor	Diferencia de medias	Diferencia de error estándar
Lectura	.688	.363	.905
Lengua escrita	.045	1.925	.961
Matemáticas	.730	.343	.993
Puntaje global	.580	.406	.732

Nota. Adaptado de la Tabla 12 del *examen de ingreso a la educación superior (ExIES) 2023-2: Reporte técnico* (Pedroza Zúñiga et al.,2024b, p. 29).

A pesar de que la equiparación fue exitosa según la prueba *t*, existen reservas de validez relacionadas con la variabilidad en los ítems ancla y las condiciones de aplicación, lo cual puede afectar la comparabilidad entre versiones.

G3.3 La estabilidad de las puntuaciones del ExIES se mantiene a lo largo de múltiples aplicaciones en diferentes periodos, reflejando un seguimiento continuo y sistemático de su confiabilidad

S3.3.1 La estabilidad de la confiabilidad del ExIES se revisa en cada nuevo periodo de aplicación, verificando la reproducibilidad de los coeficientes de consistencia interna y la comparabilidad de los puntajes obtenidos. La confiabilidad del ExIES ha sido monitoreada de manera constante a lo largo de sus diferentes aplicaciones, lo que asegura que las puntuaciones obtenidas con la prueba mantengan un alto grado de

estabilidad y precisión en el tiempo. En sintonía con los criterios de Nunnally y Bernstein (1994) y Tavakol y Dennick (2011), un valor de Alfa de Cronbach que supere .70 se considera aceptable en el ámbito educativo y psicológico, mientras que un coeficiente por encima de .80 ofrece aún mayor confianza en la precisión de la medición.

Los resultados presentados en la Tabla 46 confirman que, tanto en la aplicación 2023-1 (Formas A y B) como en el pilotaje 2023-2 (Forma A), la confiabilidad global del ExIES es sistemáticamente alta, siendo que se descartó la Forma B en 2023-2. Para 2023-1, con una población de 28,205 aspirantes, los coeficientes de Alfa de Cronbach en la Forma A (.89 en total, .74 en Lectura, .77 en Lengua Escrita, .84 en Matemáticas) y en la Forma B (.87 en total, .72 en Lectura, .77 en Lengua Escrita, .81 en Matemáticas) revelan una solidez notable en la consistencia interna de la prueba. Por su parte, la aplicación de 2023-2, que contó con 2,291 participantes, registró valores también estables (.87 en total, .74 en Lectura, .76 en Lengua Escrita, .83 en Matemáticas), lo que respalda la fiabilidad de los puntajes emitidos bajo condiciones de aplicación diferentes. Estos hallazgos ratifican la garantía G3.3 referente a la consistencia y confiabilidad interna de las puntuaciones del ExIES en sus distintas aplicaciones.

Tabla 46.
Seguimiento de los coeficientes de confiabilidad por aplicación 2023-1 y 2023-2

Área	Aplicación 2023-1 (Forma A)	Aplicación 2023-1 (Forma B)	Aplicación 2023-2 (Forma A)
Lectura	.74	.72	.74
Lengua Escrita	.77	.77	.76
Matemáticas	.84	.81	.83
Global	.89	.87	.87

Nota. Elaboración propia basada en *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* y *Examen de ingreso a la educación superior (ExIES) 2023-2: Reporte técnico* (Pedroza Zúñiga et al., 2024a, 2024b).

En suma, los coeficientes de Alfa de Cronbach superiores a .70 (y en varios casos por encima de .80) confirman la confiabilidad de la prueba a través de diferentes formatos y momentos de aplicación. Dichos resul-

tados evidencian que los eventuales cambios de población, ajustes en la composición de los ítems o variaciones en la logística de aplicación no comprometen la consistencia de la medición, lo que refuerza la validez del examen al garantizar que las puntuaciones de los sustentantes reflejan con precisión sus diferencias en habilidad.

Evaluación de la inferencia de generalización

Según la evaluación realizada, en la Tabla 47 se puede observar que, de manera global, los puntajes indican un puntaje Alto con 86.66 %. Este desempeño se explica a partir del análisis de las tres garantías consideradas.

Para replicar la inferencia de generalización, se sugirió tratar la consistencia de puntajes como un producto de rutina: reportar confiabilidad por área y forma, describir calibraciones TRI, justificar la selección de ítems ancla y documentar procedimientos de equiparación. Además, se recomendó monitorear indicadores entre ciclos (por ejemplo, distribución de puntajes y funcionamiento de ítems) para detectar cambios no deseados. En el ExIES, estos análisis aportaron evidencia sólida; como posibilidad de mejora, se consideró ampliar la comunicación de estos resultados a públicos no especialistas.

Tabla 47.
Evaluación de la inferencia de generalización

Supuesto	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S3.1.1 Coeficiente de confiabilidad aceptable por área	Se revisa la consistencia interna de cada área (Lectura, Lengua Escrita, Matemáticas) mediante Alfa de Cronbach u otras métricas, garantizando valores adecuados en cada dominio (Estándar 2.2, 4.1).	4	3	3	10 (83.3 %, Moderada)
S3.2.1 Las formas del examen tienen la misma dificultad	Se equiparan las versiones del ExIES (Forma A, Forma B) para asegurar que no existan diferencias significativas en su complejidad, de modo que los puntajes sean comparables (Estándar 44.10, 5.2, 6.2).	4	4	3	11 (91.7 %, Alta)
S3.3.1 La confiabilidad se revisa continuamente para verificar estabilidad	Se monitorean en cada periodo los coeficientes de Alfa de Cronbach y otros índices para mantener su reproducibilidad y asegurar que los cambios poblacionales o logísticos no afecten la medición (Estándar 2.1, 2.4, 2.9).	4	3	4	11 (91.7 %, Alta)
	Global	12 (100 %, Alta)	10 (83.3 %, Moderada)	10 (83.3 %, Moderada)	32/36 (88.88 %, Alta)

Capítulo 6

Inferencia de explicación

La inferencia de Explicación se centró en el significado sustantivo del puntaje: si las puntuaciones pudieron interpretarse como indicadores del constructo previsto y si la estructura interna del examen fue coherente con el modelo teórico. En el EBA, este paso articuló evidencia sobre dimensionalidad, procesos de respuesta y relaciones internas entre áreas (AERA et al., 2014; Chapelle, 2021; Messick, 1989). En la revisión del ExIES 2023-2, esta inferencia se sustentó con análisis de estructura interna (AFC), correlaciones entre áreas y estudios de funcionamiento diferencial de ítems (DIF), con el propósito de identificar si el examen operó de forma equivalente entre grupos.

Definición de las garantías, supuestos y fuentes de la inferencia de explicación

Con base en lo anterior, esta inferencia se estructura en tres frentes (véase la Figura 17). Mientras que la Tabla 48 sintetiza las garantías centrales, los supuestos asociados y las fuentes de evidencia que respaldan la correspondencia entre la estructura teórica y los datos empíricos, la comparabilidad con otros instrumentos y la ausencia de sesgos relevantes; que guardan relación con las tres técnicas mencionadas.

Figura 17.
Argumento de la inferencia de explicación como parte de la validez del argumento

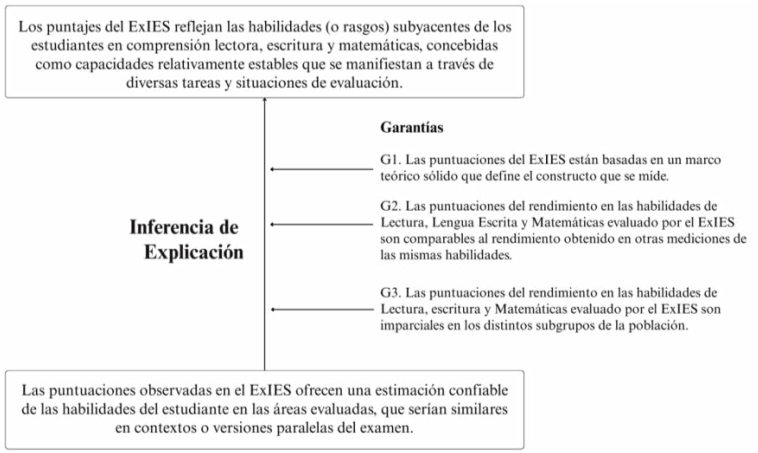


Tabla 48.
Estructura argumentativa para la inferencia de Explicación del ExIES

Conclusión de explicación	Los puntajes del ExIES reflejan las habilidades (o rasgos) subyacentes de los estudiantes en lectura, escritura y matemáticas, concebidas como capacidades relativamente estables que se manifiestan a través de diversas tareas y situaciones de evaluación	
Garantía	**Suposiciones**	**Fuentes de datos**
G4.1. Las puntuaciones del ExIES están basadas en un marco teórico sólido que define el constructo que se mide.	S4.1.1 El número de factores refleja la estructura esperada. S4.1.2 Las correlaciones del puntaje global del ExIES y cada una de sus áreas son positivas y altas. S4.1.3 La estructura factorial corresponde a la estructura teórica.	F4.1.1.1-3 Resultados del AFC que respaldan la estructura teórica de los factores, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
G4.2. Las puntuaciones del rendimiento en las habilidades de Lectura, Lengua Escrita y Matemáticas evaluado por el ExIES son comparables al rendimiento obtenido en otras mediciones de las mismas habilidades.	S4.2.1 Las correlaciones entre los puntajes del ExIES y el EXANI II son positivas. S4.2.2 Las correlaciones entre puntajes de los mismos dominios entre el ExIES y el EXANI II son positivos y fuertes.	F4.2.1.1, 4.2.2.1 Resultados de regresión lineal con los puntajes de EXANI II de los informes particulares y generales del ExIES (Pedroza Zúñiga y Gómez Monárrez, 2025a, 2025b)
G4.3. Las puntuaciones del rendimiento en las habilidades de lectura, escritura y matemáticas evaluado por el ExIES son imparciales en los distintos subgrupos de la población.	S4.3.1 Los ítems tienen la misma dificultad para todos los sustentantes.	F4.3.1.1 Resultados del análisis diferencial del Funcionamiento de los ítems (DIF) por sexo (Pedroza Zúñiga et al., 2025c)

Como se observa en la Tabla 48, en la tercera garantía (S4.3.1) conviene precisar la ubicación del análisis DIF. Siguiendo a Chapelle (2021), el DIF pertenece a la inferencia de explicación porque verifica si los ítems representan el mismo constructo en distintos subgrupos —por ejemplo,

si el puntaje tiene el mismo significado para mujeres y hombres—, es decir, si no hay varianza irrelevante que distorsione la interpretación del constructo (Chapelle, 2021). La ausencia de DIF a nivel global aporta evidencia de consistencia útil para la inferencia de generalización (estabilidad/confiabilidad entre formas y administraciones); sin embargo, los patrones por terciles —en particular, DIF moderado o severo en los extremos de la habilidad— constituyen señales diagnósticas de fuentes de varianza irrelevante (p. ej., contenido, lenguaje, formato o interacción ítem por habilidad) que afectan directamente la interpretación de lo que miden las puntuaciones (Kane, 2013).

En consecuencia, reportar y discutir el DIF en la inferencia de Explicación permite: (a) explicitar y justificar las limitaciones del Argumento de Validez en determinados tramos de habilidad; (b) acotar con precisión el alcance del reclamo de imparcialidad; y (c) sustentar decisiones de revisión o retiro de ítems. Paralelamente, la ausencia de DIF global puede mencionarse como evidencia complementaria para la generalización, pero no sustituye el análisis fino requerido para respaldar la interpretación del constructo (AERA et al., 2014; Chapelle, 2021; Kane, 2013); en este caso solo se ubica en esta inferencia.

Desarrollo de respaldos de la inferencia de explicación

G4.1. Las puntuaciones del ExIES están basadas en un marco teórico sólido que define el constructo que se mide

Para responder a esta garantía, cada uno de los supuestos contempla que la fuente de datos fueron los resultados del AFC, aplicado a las respuestas de 28 205 aspirantes que presentaron el examen en el ciclo 2023-1. Para ello, el analista de datos del ExIES primero depuró la base, eliminando registros con valores perdidos o inconsistentes y verificando la correspondencia entre respuestas y variables demográficas. Después especificó, en R 4.3.2 con el paquete lavaan, un modelo de tres factores —lectura, lengua escrita y matemáticas— haciendo que cada reactivo cargara solo en su dominio teórico.

El ajuste global se juzgó con los índices CFI, TLI y RMSEA, considerando satisfactorio un CFI y TLI iguales o superiores a .90 y un RMSEA

no mayor a .05, de acuerdo con los criterios de Hu y Bentler (1999); se aceptaron leves desviaciones dada la naturaleza de alto impacto de la prueba y el gran tamaño muestral. Por último, se interpretaron los pesos estandarizados (Std.all), señalando como evidencias sólidas las cargas superiores a .35.

S.4.1.1 El número de factores refleja la estructura esperada. Para corroborar esta estructura, se realizó un AFC en el que se modelaron tres factores, cada uno correspondiente a una de las áreas evaluadas. Los resultados obtenidos (véase Tabla 49) indican que, a pesar de que los índices de ajuste como el CFI y el TLI se encuentran por debajo de los valores ideales (≥.90), el RMSEA presenta valores aceptables (<.05), lo que sugiere que la estructura subyacente es coherente con la propuesta teórica.

Tabla 49.
Índices de ajuste en el análisis factorial confirmatorio por forma y área.

	Factores	N variables	p valor	CFI	TLI	RMSEA
Forma A	Lectura	36	<.001	.724	.719	.013
	Lengua escrita	36				
	Matemáticas	50				
Forma B	Lectura	36	<.001	.719	.714	.014
	Lengua escrita	36				
	Matemáticas	50				

Nota. Adaptado de sección 12.7 de *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a).

S.4.1.2 Las correlaciones del puntaje global del ExIES y cada una de sus áreas son positivas y altas. Según la Tabla 50, los resultados del AFC muestran una fuerte correlación positiva entre Lectura y Lengua Escrita, lo que respalda la coherencia del dominio verbal; no obstante, las correlaciones estandarizadas entre Matemáticas y las áreas verbales son negativas en esta muestra (≈ −0.47 a −0.51), por lo que S.4.1.2 queda parcialmente sostenido: apoyado para las dimensiones verbales, no para la relación positiva esperada con Matemáticas. Pese a esto, los coeficientes estadísticamente significativos ($p < .001$) señalan que las tres

áreas no son independientes, sino que comparten una varianza relevante, aunque inversa en el caso de la dimensión cuantitativa. Esto podría deberse, por ejemplo, a aquellos estudiantes con un fuerte dominio de las habilidades verbales —reflejado en puntajes altos en lectura y lengua escrita— que, sin embargo, presentan un desempeño relativamente menor en Matemáticas.

Tabla 50.
Covarianzas y correlaciones latentes entre factores en el AFC por forma

Forma	Covarianza	Estimación	Std.Err	z-value	p	Std.all
A	Factor1~~					
	Factor2	0.764	0.01	76.315	<.001	0.764
	Factor3	−0.508	0.012	−41.958	<.001	−0.508
	Factor2~~					
	Factor3	−0.472	0.013	−37.572	<.001	−0.472
B	Factor1~~					
	Factor2	0.755	0.01	78.81	<.001	0.755
	Factor3	−0.502	0.013	−39.293	<.001	−0.502
	Factor2~~					
	Factor3	−0.475	0.013	−37.718	<.001	−0.475

Nota. Adaptado de *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a). Factor 1 = Lectura; Factor 2 = Lengua Escrita; Factor 3 = Matemáticas. Los valores en Std.all indican la correlación estandarizada entre los factores.

S.4.1.3 La estructura factorial corresponde a la estructura teórica. El AFC fue utilizado para evaluar si los ítems de cada área (Lectura, Lengua Escrita y Matemáticas) están alineados con los factores esperados (Pedroza Zúñiga et al., 2024a). Se construyeron modelos factoriales separados para las dos formas del examen (Forma A y Forma B), considerando tres factores correspondientes a las tres áreas evaluadas. Cada modelo incluyó 36 ítems para los factores de Lectura y Lengua Escrita, y 50 ítems para el factor de Matemáticas.

En este sentido, el AFC de las Formas A y B muestra que, en ambos casos, predomina el grupo de ítems con cargas débiles, especialmente en Matemáticas (hasta 64 % en la Forma B); véase Tabla 51.

Tabla 51.
AFC de tres factores: cargas por forma y factor (Std.all)

Forma	Factor	No. de ítems	Fuerte	Moderado	Débil	Ítems negativos	No significativos ($p \geq 0.05$)
Forma A	Factor1	36	8	5	23	15	1
	Factor2	36	4	11	21	22	2
	Factor3	50	8	13	29	21	5
Forma B	Factor1	36	7	10	19	20	7
	Factor2	36	4	17	15	20	0
	Factor3	50	5	13	32	18	4

Nota. Elaboración propia a partir de los resultados del AFC del *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a). Factor 1 = Lectura; Factor 2 = Lengua Escrita; Factor 3 = Matemáticas. Los puntos de corte son ≥.35 fuerte; .20–.29 moderado; <.20 débil. Los valores en Std.all indican la correlación estandarizada entre los factores.

G4.2. Las puntuaciones del rendimiento en las habilidades de lectura, lengua escrita y matemáticas evaluadas por el ExIES son comparables al rendimiento obtenido en otras mediciones de las mismas habilidades.

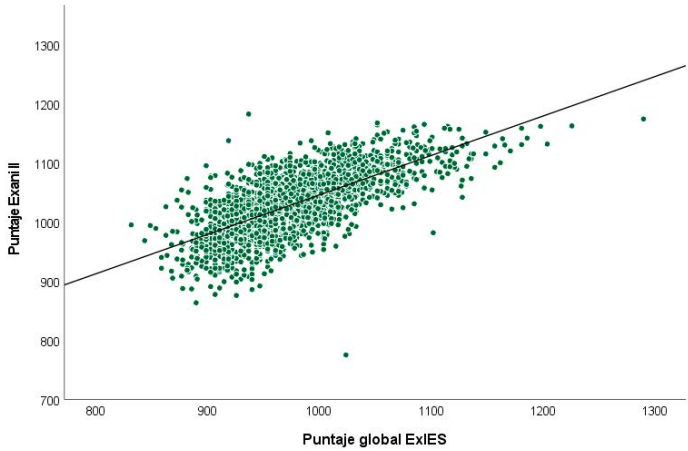
Cada uno de los supuestos que se proponen para esta garantía se alinean con el análisis de validez concurrente, entre quienes aplicaron el ExIES y también el EXANI-II en el pilotaje de 2022-2. Para valorar la validez concurrente, se calcularon correlaciones de Pearson entre los puntajes globales y por área del ExIES y del EXANI-II – donde hace referencia a Morales et al. (2015) –, de 1,937 aspirantes. Así se llevó a cabo por parte del analista de datos del ExIES:

- Emparejamiento de registros: Se aseguraron correspondencias precisas entre los registros individuales de los sustentantes en ambas pruebas y las bases de datos académicas.
- Cálculo de correlaciones: Se empleó la función cor.test de R para obtener coeficientes de Pearson (r), con interpretación según Schober et al. (2018): r ≥ .70 indica correlación fuerte, r = .50–.69 moderada y r = .30–.49 baja pero relevante para contextos educativos.

- Análisis adicional: Se exploró la relación entre los puntajes por área (lectura, lengua escrita, matemáticas) y las calificaciones en asignaturas correspondientes, para evaluar el potencial predictivo de los puntajes de admisión sobre el rendimiento universitario temprano.

S4.2.1 Las correlaciones entre los puntajes globales del ExIES y el EXANI II son positivas. Como se observa en la Figura 18 y la Tabla 52, se detallan los coeficientes de correlación de Pearson, que muestran una asociación fuerte entre el puntaje global del ExIES y el puntaje global del EXANI-II ($r=.76, p<.001$), así como entre los promedios de las áreas base del ExIES (lectura, lengua escrita y matemáticas).

Figura 18.
Diagrama de dispersión entre puntajes del ExIES y áreas base del EXANI II



Nota. Reimpreso de *examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (Pedroza Zúñiga et al., 2024a, p. 31, Figura 6).

Ambas correlaciones son estadísticamente significativas ($p<.001$), lo que refuerza la validez concurrente del ExIES al compararse con un instrumento ampliamente utilizado como el EXANI-II. Este hallazgo indica que, a pesar de la diferencia en la denominación y el enfoque de cada examen, las competencias subyacentes que evalúan guardan una relación sólida, sostenida en el rendimiento de los sustentantes.

Tabla 52.
Correlación entre puntajes del ExIES y del EXANI II por promedio y global

		ExIES	
		Pearson	p valor
EXANI II	Puntaje global	.76	<.001

Nota. Adaptado de *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico (Pedroza* Zúñiga et al., 2024a, p. 30, Tabla 14).

S4.2.2 Las correlaciones entre puntajes de los mismos dominios entre el ExIES y el EXANI II son positivos y fuertes. Para examinar el supuesto de consistencia convergente entre los dominios evaluados por el ExIES y el EXANI II (Pedroza Zúñiga y Gómez Monárrez, 2025a, 2025b), se analizaron las correlaciones entre las calificaciones universitarias en áreas afines (Lectura, Lengua Escrita y Matemáticas) correspondientes al pilotaje 2022-2. Ambos instrumentos fueron aplicados a todos los aspirantes del pilotaje (1937). Se realizó una conversión a la regresión lineal elaborada. Y, como se observa en la Tabla 53, las correlaciones entre los puntajes de los mismos dominios del ExIES y del EXANI II son positivas y de magnitud moderada a fuerte: Lectura r = .528, Lengua Escrita r = .487, Matemáticas r = .340 (todas p < .001). Estos resultados respaldan la consistencia convergente entre ambos instrumentos en dominios homólogos.

Tabla 53.
Correlaciones entre calificaciones y puntajes del ExIES, EXANI

Dominio	**r (ExIES, EXANI II)**	**p**
Lectura	0.528	< .001
Lengua escrita	0.487	< .001
Matemáticas	0.340	< .001

Nota. Elaboración propia basada en datos adaptados de *estudio de validez concurrente con los puntajes de EXANI II* (Pedroza Zúñiga y Gómez Monárrez, 2024a). El informe no publicó estas correlaciones entre instrumentos; se recuperaron indirectamente a partir de las correlaciones cero-orden con calificaciones y los coeficientes β estandarizados del modelo de regresión conjunta reportado para el pilotaje 2022-2. Todas las variables están estandarizadas; N = 1,937; todas las p < .001.

G4.3. Las puntuaciones del rendimiento en las habilidades de Lectura, Lengua Escrita y Matemáticas evaluado por el ExIES son imparciales en los distintos subgrupos de la población

En este caso se llevó a cabo el análisis DIF con una muestra de 2,288 aspirantes correspondientes al periodo 2023-2 del ExIES. Del total, el 49.4 % se identificó como mujer, el 49.4 % como hombre y el 1.1 % no especificó sexo; para los análisis comparativos por sexo, únicamente se consideraron registros con datos válidos en esta variable (Pedroza Zúñiga y Gómez Monárrez, 2025c). Esta muestra se consideró suficiente para la detección de diferencias estadísticamente significativas y robustas entre subgrupos.

Este análisis se aplicó para identificar sesgos potenciales por sexo en los ítems del ExIES (Pedroza Zúñiga y Gómez Monárrez, 2025c). El procedimiento fue exhaustivo y combinó métodos clásicos y modernos, siguiendo criterios de robustez estadística y buenas prácticas internacionales (García et al., 2016; Zieky, 1993):

- Preprocesamiento: Exclusión de registros sin especificación de sexo o con respuestas omitidas excesivas, clasificación de los sustentantes según grupo de comparación (hombres vs. mujeres).
- Método Mantel-Haenszel (MH): Se aplicó el procedimiento MH para comparar la probabilidad de respuesta correcta entre grupos, controlando el nivel global de habilidad. El estadístico Odds Ratio (OR) y su logaritmo (logOR) se usaron como estimadores de magnitud de DIF, aplicando los puntos de corte de Zieky (1993): $|\text{logOR}| < .43$ (DIF leve), $.43 \leq |\text{logOR}| < .64$ (moderado), $|\text{logOR}| \geq .64$ (severo).
- Modelo de dos parámetros logísticos (2PL): Para explorar el DIF no uniforme, se utilizó el paquete mirt en R, estimando el parámetro de habilidad (θ) de cada sustentante y dividiendo la muestra en terciles. Se recalculó logOR en cada tercil para identificar variaciones en el DIF según el nivel de habilidad (Tate, 2004).
- Visualización y síntesis: Los resultados se graficaron y se sistematizaron en tablas, especificando los ítems con DIF moderado/severo y el grupo favorecido.

S4.3.1 Los ítems tienen la misma dificultad para todos los sustentantes. En la Tabla 54 se muestran, de forma global, los resultados del análisis del DIF por sexo en las tres áreas del ExIES: lectura, lengua escrita y matemáticas, considerando sus dos formas de aplicación (A y C). De un total de 244 ítems evaluados, el 84 % de los ítems no presentó DIF estadísticamente significativo ni superó los umbrales de magnitud establecidos. El 16 % restante mostró diferencias en el desempeño entre sexos, distribuidas de manera casi equitativa a favor de hombres (8.19 %) y de mujeres (7.78 %). No obstante, al considerar los criterios propuestos por Zieky (1993), que categorizan el log (OR) en rangos de DIF leve ($|\log (OR)| < 0.43$), moderado ($|\log (OR)| \geq 0.43$) y severo ($|\log (OR)| \geq 0.64$), se advierte que la mayoría de estos ítems se ubican dentro del rango de DIF leve. Esto indica que, si bien existen ítems con diferencias entre grupos, dichas diferencias no alcanzan un tamaño del efecto que comprometa sustancialmente la imparcialidad de la prueba en términos globales. A continuación, se detallarán los hallazgos específicos por área y forma.

Tabla 54.
Número de ítems con DIF por sexo según área y forma del ExIES 2023-2

Área	Forma	Total	No DIF	DIF a favor de H	DIF a favor de M
Comprensión	A	36	28	3	5
Comprensión	C	36	30	3	3
Lengua escrita	A	36	34	1	1
Lengua escrita	C	36	27	4	5
Matemáticas	A	50	42	5	3
Matemáticas	C	50	44	4	2
Total	–	244	205 (84 %)	20 (8.19 %)	19 (7.78 %)

Nota. Elaboración propia con datos obtenidos *de Funcionamiento Diferencial del ítem (DIF): Examen de Ingreso a la Educación Superior (ExIES) 2023-2* (Pedroza Zúñiga y Gómez Monárrez, 2025c).

Análisis DIF por terciles. El desglose por terciles revela focos de sesgo concretos que no aparecen en el análisis global. En el caso de Lectura (36 ítems por forma), el tercil bajo mostró 4 ítems moderados (11 %); el medio otros 4 (3 moderados, 1 severo; 11 %); y el alto 10 (28 %; 7 moderados, 3 severos). Aquí sobresalen los ítems 9 y 10 de la Forma C, con DIF severo a favor de mujeres en los niveles medio y alto.

Sobre Lengua Escrita (36 ítems por forma), en la Forma A, el tercil inferior concentró 4 reactivos con DIF (11 %; 3 moderados, 1 severo) y el tercil superior 8 (22 %; 5 moderados, 3 severos). En la Forma C se repite el patrón: 4 ítems en el tercil bajo (11 %; 3 moderados, 1 S), 3 en el medio (8 %; 1 moderados, 2 severos) y 7 en el alto (19 %; 5 moderados, 2 severos). Además, identifica 27 reactivos únicos con DIF moderado/severo, el 37.5 % de los 72 ítems de esta área; no obstante, solo 9 (12.5 %) alcanzan la categoría de severo.

Y, en Matemáticas (50 ítems por forma), para la Forma A se detectaron 8 reactivos en el tercil inferior (16 %; 7 moderados, 1 severo), 8 en el medio (16 %; todos moderados) y 9 en el superior (18 %; 6 moderados, 3 severos). En la Forma C, la incidencia crece en el tercil alto: 4 ítems en el bajo (8 %; todos moderados), 7 en el medio (14 %; todos moderados) y 12 en el alto (24 %; 7 moderados, 5 severos). Cinco reactivos (23, 24, 26, 37 y 45) presentan DIF salto y requieren una revisión adicional.

Síntesis. En conjunto, 59 de los 244 reactivos (24 %) exhibieron DIF moderado o severo en al menos un tercil y 15 (6 %) alcanzaron la categoría severa. Aunque el banco mantiene una imparcialidad aceptable, estos puntos severos confirman que reactivos neutros en el promedio global pueden volverse parciales en extremos de habilidad; por ello se recomienda revisar de inmediato los 15 ítems con efectos severos recurrentes y continuar monitoreando los porcentajes críticos (≥ 20 % por tercil) en futuras aplicaciones.

Aunque el análisis global de DIF indica que el 84 % de los ítems no presenta sesgo por sexo, el análisis por terciles revela que 59 de 244 ítems (24 %) muestran DIF moderado o severo en al menos un tercil y 15 ítems (6 %) alcanzan severidad recurrente. Estos hallazgos muestran que ítems que parecen neutros en la muestra global pueden comportarse de forma parcial en extremos de habilidad, lo cual condiciona la

interpretación de los puntajes: los puntajes reflejan rasgos estables y equivalentes entre sustentantes; es plausible para la mayoría, pero no puede asumirse sin reservas cuando se toman decisiones centradas en los percentiles extremos.

Evaluación de la inferencia de explicación

Según la Tabla 55, el resultado global de la inferencia de explicación fue de 80.95 %, ya que sus puntos a mejorar son en claridad y plausibilidad de la evidencia, es decir, expresar en los reportes y documentos el procedimiento y la selección teórica con el fin de interpretar de forma adecuada los puntajes.

Tabla 55.
Evaluación de la inferencia de Explicación.

Supuesto	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S4.1.1 Número de factores	La estructura dimensional del ExIES (tres áreas) se corresponde parcialmente con la propuesta teórica; CFI/TLI < .90, RMSEA aceptable (Estándar 1.13, 1.14).	3	3	3	9 (75 %, Moderada)
S4.1.2 Correlaciones positivas y altas	Las correlaciones entre el puntaje global y las áreas individuales del ExIES evidencian coherencia interna (Estándar 1.14, 1.15).	4	3	3	10 (83.33 %, Moderada)

S4.1.3 Cargas factoriales por área	Los ítems presentan cargas más altas en el factor teórico esperado, respaldando la validez interna del examen (Estándar 1.13, 1.16).	3	3	3	9 (75 %, Moderada)
S4.2.1 Concurrencia con otras mediciones	El ExIES conserva coherencia con pruebas como el EXANI-II y otras evaluaciones de habilidades similares (Estándar 1.19, 1.20).	3	4	3	10 (83.33 %, Moderada)
S4.2.2 Concordancia de dominios	Correlaciones fuertes entre dominios semejantes en distintos instrumentos, p. ej. Lectura vs. Comprensión Lectora (Estándar 1.20, 1.21).	3	4	3	10 (83.33 %, Moderada)
S4.3.1 Imparcialidad en subgrupos	Mantiene la misma dificultad para sustentantes de subgrupos diversos, sin indicios de sesgo sistémico (Estándar 3.2, 3.6, 3.7).	4	3	3	10 (83 %, Moderada)
	Global	23 (82.14 %, Moderada)	24 (85.71 %, Moderada)	21 (75 %, Moderada)	68 / 84 (80.95 %, Moderada)

En la inferencia de Explicación, la posibilidad más útil fue conectar estadística y teoría: no bastó con que el examen “ajustara” a un modelo, sino que se interpretó qué significó ese ajuste para el constructo y para la

equidad. Para replicarla, se recomendó documentar la estructura interna (dimensionalidad), revisar relaciones entre áreas y realizar análisis de imparcialidad (como DIF) por subgrupos pertinentes. Cuando se detectaron señales de sesgo o de estructura inesperada, el EBA permitió traducir hallazgos en decisiones concretas (revisar reactivos, ajustar especificaciones o redefinir interpretaciones).

Capítulo 7

Inferencia de extrapolación

La inferencia de extrapolación preguntó si el puntaje del examen se extendió más allá del contexto de aplicación: si el desempeño observado en la prueba se relacionó con desempeño posterior o concurrente en el dominio objetivo (Cronbach y Meehl, 1955; Kane, 2013; Chapelle, 2021). En exámenes de admisión, suele traducirse en evidencia predictiva respecto a trayectorias académicas iniciales o rendimiento en los primeros semestres. En la revisión del ExIES 2023-2, esta inferencia se apoyó en modelos predictivos y en el análisis de relaciones con variables externas disponibles, con el fin de estimar la fuerza y los límites de la extrapolación; aquí solo se coloca una síntesis del estudio realizado.

Definición de las garantías, supuestos y fuentes de la inferencia de extrapolación

En la Figura 19 se presenta la conclusión sobre que las habilidades evaluadas por el ExIES predicen adecuadamente un rendimiento futuro, en este caso el primer año de universidad; por lo que este argumento es muy concreto. Como se mencionó en el *Diseño Metodológico*, para esta inferencia, sí hubo un estudio propio y ajeno al equipo del ExIES. Por otro lado, la Tabla 56 muestra la garantía, supuesto y fuentes de datos que respaldan la inferencia de extrapolación del ExIES.

Figura 19.
Argumento de la inferencia de extrapolación como parte de la validez del argumento

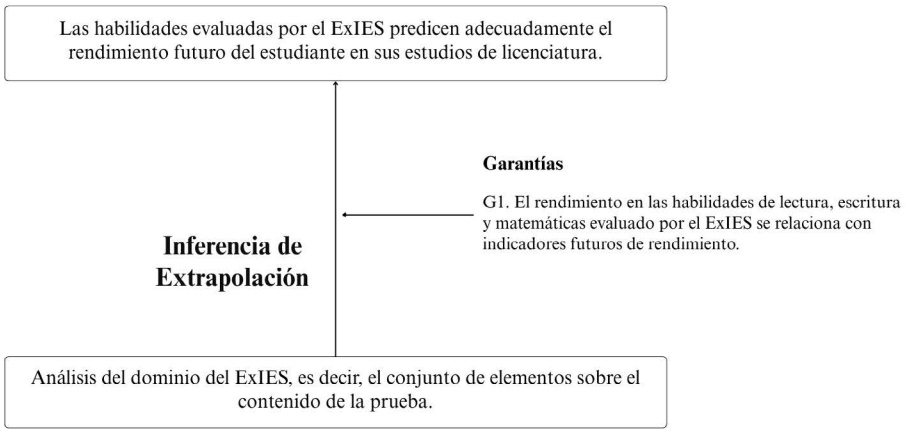


Tabla 56.
Estructura argumentativa para la inferencia de extrapolación del ExIES

Conclusión de extrapolación	Las habilidades evaluadas por el ExIES predicen adecuadamente el rendimiento futuro del estudiante en sus estudios de licenciatura	
Garantía	**Suposiciones**	**Fuentes de datos**
G5.1. El rendimiento en las habilidades de Lectura, escritura y Matemáticas evaluado por el ExIES se relaciona con indicadores futuros de rendimiento.	S5.1.1 Las puntuaciones del ExIES se correlacionan positivamente con indicadores de desempeño académico en la universidad para comprender y utilizar la Lengua Escrita, las Matemáticas y su habilidad en Lectura, y otros indicadores relacionados con estas áreas.	F5.1.1.1 Estudio de validez predictiva (con Machine Learning) del puntaje y los promedios de calificación de los estudiantes durante su trayecto formativo (Elaboración propia) F5.1.1.2 Base de datos de promedios por alumno de Bachillerato (EMS BC, 2024) F5.1.1.3 Base de datos del ExIES (Pedroza Zúñiga et al.,2024) F5.1.1.4 Base de datos del promedio del primer y segundo semestre de universidad (UABC, 2024)

Desarrollo de respaldos de la inferencia de extrapolación

G5.1 El rendimiento en las habilidades de lectura, escritura y matemáticas evaluado por el ExIES se relaciona con indicadores futuros de rendimiento

El objetivo de este estudio fue comprobar la eficacia de modelos de regresión apoyados en aprendizaje automático (machine learning, en inglés) para anticipar el desempeño académico durante el primer año universitario, empleando como variables predictoras las calificaciones del ExIES y el promedio bachillerato. Se fijó como criterio alcanzar un coeficiente de determinación $R^2 \geq 0.15$, lo que implicaría explicar al menos el 15 % de la variabilidad del rendimiento académico con los predictores seleccionados, según los antecedentes revisados.

El estudio analizó los 10 184 registros correspondientes a la totalidad de aspirantes que presentaron el ExIES en el ciclo 2023-1, lo que permitió trabajar con la cohorte completa y evitar sesgos de muestreo. Para evaluar la capacidad predictiva del ExIES sobre el rendimiento universitario de primer año (Año1) se obtuvieron los puntajes del ExIES, los promedios de bachillerato y los resultados de su primer año universitario. Tras un preprocesamiento exhaustivo —imputación de faltantes, detección de atípicos y estandarización z— se ajustaron siete modelos (regresión lineal, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost y un MLP), todos implementados en Python con pandas, numpy y scikit-learn y validados mediante cross-validation 5-fold con una partición 90%/10% entrenamiento-prueba. El desempeño se juzgó con R^2, MSE y RMSE, mientras que la colinealidad se identificó mediante VIF < 3, concluyendo que los parámetros por defecto ofrecían el mejor balance entre sencillez y ajuste (Burkov, 2019; Gerón, 2019).

S5.1.1 Las puntuaciones del ExIES se correlacionan positivamente con indicadores de desempeño académico en la universidad para comprender y utilizar la Lengua Escrita, las Matemáticas y su habilidad en Lectura, y otros indicadores relacionados con estas áreas. Los resultados del estudio sobre confirman la suposición 5.1.1, ya que las puntuaciones del ExIES muestran correlaciones positivas y significativas con indicadores de desempeño académico en la universidad, como el promedio del primer año de licenciatura, sobre todo cuando se utilizan otras variables como el promedio bachillerato.

Los modelos de regresión lineal, ridge y otros métodos predictivos evaluados respaldan esta relación, según los modelos utilizados (véase Tabla 57). El modelo ridge (básico) arrojó un desempeño casi idéntico al de la regresión lineal básica (R^2 = 0.221388, RMSE = 0.192069), confirmando que las puntuaciones del ExIES, junto con el promedio bachillerato, son predictores clave del rendimiento académico. El modelo ridge puede considerarse como una opción preferible ante colinealidad y varianza del muestreo (Gerón, 2019).

Tabla 57.
Desempeño comparativo de modelos predictivos sobre conjunto de prueba.

Modelo	R2 (prueba)	RMSE (prueba)
Regresión lineal (básico)	0.221561	0.192047
Ridge (básico)	0.221388	0.192069
Lasso (básico)	-0.000074	0.217677
Random Forest (básico)	0.133060	0.202670
Gradient Boosting (básico)	0.212081	0.193213
Red Neuronal (MLP) (básico)	0.191233	0.195753
XGBoost (básico)	0.096929	0.206851

Nota. Los valores en negritas representan el mejor desempeño en cada columna. R^2 indica la proporción de varianza explicada por el modelo; valores más altos representan mejor ajuste. RMSE (Root Mean Square Error) refleja el error medio cuadrático de predicción; valores más bajos indican mayor precisión. En estos modelos no hubo ajustes de hiperparámetros ya que demostraron mayor estabilidad.

Por otra parte, el modelo de regresión lineal (básico), que incluye tanto las puntuaciones de las áreas evaluadas por el ExIES por separado (lectura, matemáticas y lengua escrita) como el promedio bachillerato, obtuvo un coeficiente de determinación (R^2) de 0.2215 en el conjunto de prueba, indicando que estas variables explican el 22.1 % de la variabilidad en el rendimiento académico. Este resultado se ve reforzado por un RMSE bajo de 0.192 (19.2 %).

En contraste con la solidez de la regresión lineal y ridge, los modelos más complejos evidenciaron limitaciones inherentes a su configuración básica: lasso casi no explica varianza ($R^2 \approx 0$) porque la penalización L1 suprimió coeficientes relevantes; el random forest mejora ligeramente ($R^2 = 0.13$), pero, con pocos árboles y profundidad predeterminada, infra ajusta; el Gradient Boosting se aproxima a los lineales ($R^2 = 0.21$), aunque sin afinar la tasa de aprendizaje ni el número de iteraciones no logra superarlos; la red neuronal MLP ($R^2 = 0.19$) requiere optimizar capas y regularización para captar patrones más complejos; y XGBoost ($R^2 = 0.10$) queda rezagado porque su potencial depende del ajuste de hiperparámetros, pero que al momento no se encontró un mejor ajuste.

Estos resultados confirman que la relación entre las subpuntuaciones del ExIES y el rendimiento universitario es esencialmente lineal y que, sin otros tipos de optimización de hiperparámetros, los algoritmos de mayor varianza aportan poco valor añadido (Géron, 2019; Hastie, Tibshirani, & Friedman, 2009).

Comparado con enfoques más complejos como random forest o gradient boosting, el modelo ridge demostró un desempeño competitivo con un R^2 de 0.213 en el conjunto de prueba, posicionándose como una opción más simple y estable. Esto es consistente con investigaciones que subrayan la eficacia de los modelos regularizados en escenarios donde predominan relaciones lineales entre las variables predictoras, como señalan Montgomery et al. (2012), y Breiman (2001). Al tratarse de un algoritmo perteneciente a la familia de aprendizaje automatizado, el modelo ridge también integra prácticas avanzadas que optimizan su desempeño en contextos educativos, donde se manejan grandes volúmenes de datos o variables correlacionadas. Este enfoque, alineado con los avances en análisis predictivo, permite diseñar estrategias académicas más eficientes y basadas en evidencia.

Evaluación de la inferencia de extrapolación

La evidencia utilizada para esta inferencia comprendió el estudio de validez predictiva con modelos de regresión y validación cruzada, la base del ExIES 2023-1, los promedios de bachillerato y los promedios universitarios del primer año; los insumos son identificables y trazables, y los criterios de evaluación (R^2, RMSE) se reportan de forma explícita, cumpliendo con los Estándares 1.19 y 12.13 sobre documentación de relaciones criterio y error de predicción (AERA et al., 2014). Además, el argumento actual es claro porque describe con precisión los datos usados (ExIES, promedios de bachillerato y promedios universitarios), el tamaño de muestra amplio (N = 10 184), los modelos aplicados (OLS, Ridge y otros) y las métricas de evaluación (R^2, RMSE) junto con validación cruzada 5-fold; todo eso facilita reproducir y evaluar el hallazgo (Hastie et al., 2009; Géron, 2019; AERA et al., 2014). La concordancia entre OLS y Ridge ($R^2 \approx .21$–$.22$) aporta coherencia interna: distintos métodos lineales llegan a conclusiones parecidas, lo que hace más creíble la inferencia.

Tabla 58.
Evaluación de la inferencia de extrapolación según el supuesto S5.1.1

Supuesto	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S5.1.1 Correlación con rendimiento académico futuro	Las puntuaciones del ExIES muestran una asociación positiva y significativa con indicadores del primer año universitario, reflejando la continuidad entre el desempeño en la prueba y la habilidad real (Estándar 1.19, 1.20, 5.1, 12.13)	4	3	4	11 (91.7 %, Alta)
	Global	4 (100 %, Alta)	3 (87.5 %, Moderada)	4 (100 %, Alta)	11 / 12 (91.66 %, Alta)

Como observaciones, aunado a la validación cruzada ya usada, es recomendable: (a) realizar análisis de sensibilidad (p. ej.. cambiar particiones, usar Bootstrap) para ver si R^2 y RMSE se mantienen; (b) probar calibración (gráficos predichos vs. observado) e intervalos de confianza para predicciones; (c) hacer validación externa si es posible (otra cohorte o universidad); y (d) desagregar resultados por subgrupos y revisar medidas de equidad (DIF y estabilidad por sexo, escuela, región). Estas pruebas fortalecen que el patrón observado no sea artefacto de una decisión de modelado o de la muestra (Hastie, Tibshirani, & Friedman, 2009; Magis et al., 2010).

Capítulo 8

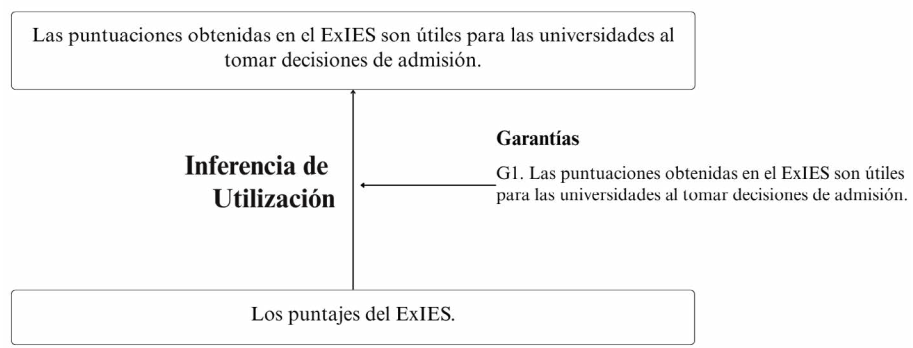
Inferencia de utilización e implicación de consecuencias

Las inferencias de utilización e implicación de consecuencias conectaron el puntaje con decisiones reales: cómo se usaron los resultados (reglas de decisión, cortes, ponderaciones) y qué efectos produjeron en aspirantes, programas e institución. En exámenes de alto impacto, este tramo fue especialmente sensible, porque combinó evidencia técnica con criterios de política educativa, equidad y responsabilidad pública (AERA et al., 2014; Messick, 1989; Newton y Shaw, 2014; Shepard, 2016). En la revisión del ExIES 2023-2, el análisis se centró en (a) la utilidad del puntaje para el propósito declarado de selección/admisión y (b) la identificación de consecuencias previstas y no previstas.

Definición de las garantías, supuestos y fuentes de la inferencia de utilización

Con base en estas perspectivas, ante la ausencia de encuestas o acercamiento a la población y comunidad educativa, en esta inferencia se revisan normativas institucionales, guías operativas y reportes técnicos del ExIES, para determinar si los puntajes efectivamente guían procesos de admisión acordes con los principios éticos, técnicos y administrativos esperados; como lo propone Chapelle (2021) en este tipo de casos, con el fin de realizar reflexiones y ponerlo a la vista de los desarrolladores, es decir, del equipo del ExIES. La Figura 20 representa gráficamente el proceso de utilización del ExIES.

Figura 20.
Argumento de la inferencia de utilización como parte de la validez del argumento



La Tabla 59, por su parte, sintetiza las garantías y supuestos clave, respaldando que los puntajes obtenidos son realmente útiles para las decisiones de admisión universitaria y reflejan fielmente las habilidades requeridas según sus fuentes de datos disponibles, que posteriormente son evaluadas.

Tabla 59.
Estructura argumentativa para la inferencia de Utilización del ExIES

Conclusión	Las puntuaciones obtenidas en el ExIES son útiles para las universidades al tomar decisiones de admisión	
Garantía	**Suposiciones**	**Fuentes de datos**
G6.1. Las puntuaciones ExIES son útiles para ayudar en las decisiones de admisión en instituciones de educación superior, en este caso de la UABC, reflejando la habilidad del examinado en comprensión y uso de la lengua escrita, matemáticas y lectura.	S.6.1.1 La orientación proporcionada por el desarrollador del ExIES es utilizada por las instituciones, en este caso de la UABC, para establecer sus propios criterios de admisión y ubicación; en el caso de la UABC, por el orden de prelación. S.6.1.2 Las investigaciones empíricas muestran que es efectivo utilizar los puntajes del ExIES para la selección de los sustentantes.	F6.1.1.1 Ley Orgánica de la UABC (2010) F6.1.1.2 Estatuto General de la UABC (2019) F6.1.1.3 Documentos oficiales de admisión descritos en el Estatuto Escolar de la UABC, como los Artículos 16, 18 y 24, (UABC, 2021) F6.1.2.1 Reportes Técnicos (Pedroza et al., 2024a, 2024b) F6.1.2.2 Guía del sustentante (Pedroza Zúñiga et al., 2023l)

Desarrollo de respaldos de la inferencia de evaluación

G6.1. Las puntuaciones del ExIES son útiles para ayudar en las decisiones de admisión en instituciones de educación superior, en este caso de la UABC, reflejando la habilidad del examinado en comprensión y uso de la lengua escrita, matemáticas y lectura

S6.1.1 La orientación proporcionada por el desarrollador del ExIES es utilizada por las instituciones, en este caso de la UABC, para establecer sus propios criterios de admisión y ubicación; en el caso de la UABC por

el orden de prelación. Como se mencionó, no se cuenta con evidencias directas —por ejemplo, encuestas, entrevistas o registros de consulta a la comunidad educativa— que confirmen cómo y con qué alcance se aplican orientaciones para establecer sus propios criterios de admisión y ubicación; aunque se transmite una orientación de los puntajes: "1) campus, 2) ficha,3) puntaje en Lectura, 4) puntaje en Lengua Escrita, 5) puntaje en Matemáticas, 6) puntaje global, 7)nombre, 8) apellido paterno y 9) apellido materno del aspirante" (Pedroza Zúñiga et al., 2024a, 2024b). Aun así, pueden identificarse cuatro elementos documentales que sustentan la necesidad y viabilidad de investigaciones posteriores: (1) el marco legal y de autonomía universitario expresado en la Ley Orgánica de la UABC (UABC, 2010), Art. 3°, fracc. I; (2) el mandato del Estatuto General (UABC, 2019), Título Quinto, Art. 171, sobre la sujeción de aspirantes al proceso de selección determinado por la Universidad; (3) las disposiciones del Estatuto Escolar (UABC, 2021), que definen el examen de selección y el procedimiento de orden de prelación (Art. 3º, fracc. XXIII; Art. 24); y (4) la operatividad y orientación técnica del ExIES (p. ej. puntos de corte; publicaciones en admisiones.uabc.mx; Guía del Sustentante y Reportes Técnicos semestrales), que evidencian mecanismos institucionales para la aplicación y divulgación del instrumento.

S6.1.2 Las investigaciones empíricas muestran que es efectivo utilizar los puntajes del ExIES para la selección de los sustentantes. Las investigaciones disponibles para este supuesto son principalmente internas y de carácter psicométrico, y muestran indicios —pero no una verificación concluyente— de que los puntajes del ExIES pueden contribuir a la selección de sustentantes: por ejemplo, los coeficientes de consistencia interna, según las áreas, son buenos e indican calidad técnica del instrumento (Pedroza Zúñiga et al., 2024a).

No obstante, no se cuenta con estudios independientes ni con evidencia longitudinal o de resultados de decisión (p. ej., seguimiento del desempeño académico de admitidos vs. rechazados, auditorías de las decisiones por prelación, entrevistas a tomadores de decisión) que prueben de forma directa la eficacia del ExIES como herramienta de selección.

Evaluación de la inferencia de utilización

Como se mencionó, hacen falta pruebas directas de cómo las áreas de admisión aplican en la práctica las orientaciones del ExIES (por ejemplo, actas, encuestas o entrevistas). Sí existen normas que permiten usar puntajes y ordenar por prelación (UABC, 2010, 2019, 2021, 2025). Pero los Estándares solicitan evidencias que vinculen lo que se interpreta del puntaje con lo que se hace con él y con sus efectos (AERA et al., 2014). Para fortalecer este supuesto se requieren estudios adicionales: análisis predictivo longitudinal, estudios de consecuencias, auditorías de uso institucional, encuestas y entrevistas a responsables de admisión y a sustentantes. Bajo el EBA, esa conexión debe demostrarse y no asumirse (Kane, 2013; Chapelle, 2021). Por ello, en la Tabla 60 se muestra que esta inferencia se valoró con un puntaje global de 58.3%.

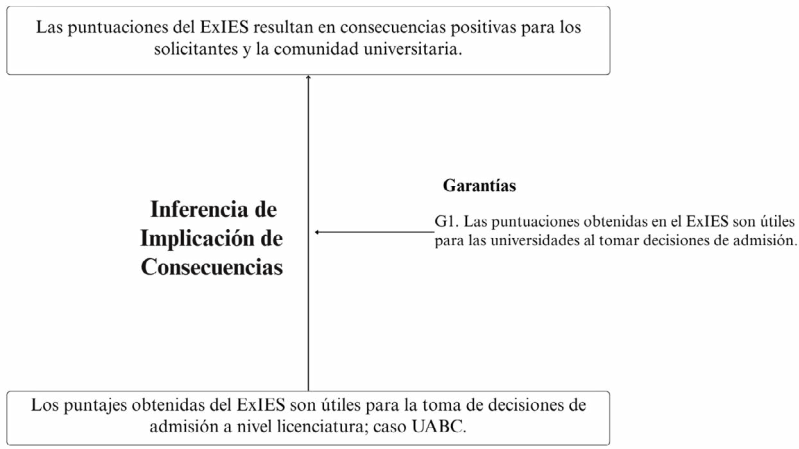
Tabla 60.
Evaluación de la inferencia de Utilización.

Supuesto	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S6.1.1 Se orienta a la UABC a establecer sus propios criterios de admisión.	Las instituciones de educación superior utilizan las puntuaciones del ExIES como un criterio fiable para tomar decisiones de admisión (Estándar 12.2).	2	2	2	6 (50 %, Baja)
S6.1.2 Es efectivo utilizar los puntajes del ExIES para la selección de sustentantes.	Se aplica correctamente la orientación del ExIES para interpretar los resultados y ubicar a los sustentantes en su respectivo orden de prelación. Además, los puntajes sirven para detectar fortalezas y debilidades que orienten cursos remediales o de regularización. (Estándar 12.1, 11.4).	2	3	3	8 (66.66 %, Moderada)
	Global	4 (50 %, Baja)	5 (62.5 % Baja)	5 (62.5 %, Baja)	14 (58.3 %, Baja)

Definición de las garantías, supuestos y fuentes de la inferencia de implicación de consecuencias

La Figura 21 define la conclusión de la implicación de consecuencias del ExIES, para resolver cómo las garantías clave aseguran que los estudiantes admitidos mediante el uso del examen son seleccionados con base en un criterio de orden de prelación justo y equitativo.

Figura 21.
Argumento de Implicación de Consecuencias como parte de la validez del argumento



En este sentido, el propósito es comprobar que la aplicación del examen produce resultados imparciales y positivos para la comunidad académica, asegurando que los aspirantes admitidos cumplan con los niveles de dominio requeridos y que exista un correlato entre el puntaje obtenido y su potencial de éxito universitario (véase la Tabla 61).

Tabla 61.
Estructura argumentativa para la inferencia de implicación de vonsecuencias

Conclusión	Las puntuaciones del ExIES resultan en consecuencias positivas para los solicitantes y la comunidad universitaria.	
Garantía	**Suposición**	**Fuentes de datos**
G7.1. Los estudiantes con los puntajes adecuados, según el orden de prelación, obtienen admisión a la universidad, y aquellos sin los puntajes adecuados, según la prelación, no son admitidos.	S.7.1.1 Existe una responsabilidad conjunta entre el desarrollador del ExIES y el personal universitario encargado de las decisiones de admisión para garantizar el orden de prelación.	F7.1.1.1 Ley Orgánica de la UABC (2010) F7.1.1.2 Estatuto General de la UABC (2019) F7.1.1.3 Procedimientos de prelación definidos en el Estatuto Escolar de la UABC, Artículo 24 (UABC, 2021) F7.1.1.4 Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a)
	S7.1.2 Los puntajes del ExIES permiten clasificar de manera imparcial y útil a los aspirantes para asignar prelación de ingreso universitario.	F7.1.2.1 Resultados del análisis del DIF por sexo (Pedroza Zúñiga & Gómez Monárrez, 2025c) F7.1.2.2 Resultados de la inferencia de Extrapolación.

Desarrollo de respaldos de la inferencia de implicación de consecuencias

G7.1. Los estudiantes con los puntajes adecuados, según el orden de prelación, obtienen admisión a la universidad, y aquellos sin los puntajes adecuados, según la prelación, no son admitidos

S7.1.1 Existe una responsabilidad conjunta entre el desarrollador del ExIES y el personal universitario encargado de las decisiones de admisión para garantizar el orden de prelación. Para este supuesto, no se cuenta con estudios o actas documentadas que evalúen de manera directa

y conjunta cómo el ExIES influye en el orden de prelación y en las consecuencias derivadas de las decisiones de admisión. El Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) describe indicadores globales de rendimiento, como la media de los puntajes y los coeficientes de confiabilidad, pero no detalla un proceso de retroalimentación formal donde la UABC y los diseñadores del examen se reúnan para adaptar o validar las políticas de admisión. Esta ausencia de documentación no implica que no exista colaboración, pero sí indica que, por ahora, el supuesto permanece sin evidencia de consecuencias de forma directa y requiere formalizar esquemas de supervisión y actualización.

Si bien el Estatuto Escolar (UABC, 2021) indica la importancia de los méritos académicos y el uso de los puntajes como parte de la selección de ingreso por méritos, es necesario establecer mecanismos para evaluar si, con el paso de las generaciones, los aspirantes admitidos efectivamente cumplen o superan las expectativas de desempeño dentro de la universidad (Messick, 1989). Dicho de otro modo, el vínculo entre el uso del ExIES y beneficios institucionales observables permanece como una hipótesis razonable que aún debe verificarse mediante estudios de consecuencias.

S7.1.2 Los puntajes del ExIES permiten clasificar de manera imparcial y útil a los aspirantes para asignar prelación de ingreso universitario. En este supuesto, lo central para sostener esta inferencia no es solo la evidencia psicométrica acerca de los puntajes, sino también la existencia y transparencia de reglas de decisión que describan quién hace qué con los puntajes y las condiciones de su aplicación. No obstante, la evidencia técnica reunida hasta ahora aporta señales favorables. De los 244 ítems analizados en el Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a), el 84 % carece DIF significativo y el 16 % restante se distribuye casi equitativamente entre hombres (8.19%) y mujeres (7.78 %), lo que indica que las diferencias en los puntajes responden a la habilidad y no a sesgos de género, en concordancia con el estándar 7.0 de los Estándares (AERA et al., 2014). Asimismo, los resultados de confiabilidad del ExIES respaldan la precisión de la medida, y el estudio predictivo (validez predictiva) con una muestra de la cohorte 2023-1 (N = 10 184) mostró que las subpuntuaciones de Lectura, Matemáticas y Lengua Escrita, combinadas con el promedio de bachillerato, explican aproximadamente el 22 % de la

varianza del promedio universitario de primer año (R^2 = .22; RMSE = .19) mediante modelos de regresión lineal y Ridge. Estos hallazgos son indicios técnicos útiles (imparcialidad, confiabilidad, validez predictiva), pero no constituyen por sí mismos evidencia sobre las consecuencias del uso del ExIES en la toma de decisiones de admisión.

En consecuencia, por ahora no es posible afirmar —con base en evidencia sobre consecuencias— que las decisiones de admisión derivadas del ExIES generen beneficios o perjuicios específicos;

Evaluación de la inferencia de Implicación de Consecuencias

Los reportes técnicos del ExIES muestran evidencia psicométrica (p. ej., confiabilidad, análisis y resultados por área) (Pedroza Zúñiga et al., 2024a); además, hay análisis de DIF por sexo que revisan posible sesgo en ítems (Pedroza Zúñiga y Gómez Monárrez, 2025c) y estudios que sugieren predicción del desempeño cuando se combina el puntaje con el promedio de bachillerato (Pedroza Zúñiga y Gómez Monárrez, 2025b). Aun así, estas evidencias técnicas no prueban por sí mismas efectos justos o beneficios concretos en las decisiones (AERA et al., 2014; Sackett, Borneman, y Connelly, 2008). Se necesita conectar decisiones de admisión con resultados observables como rendimiento y retención en el primer año para valorar consecuencias (Kuh et al., 2008; Tinto, 1993). Así, el 50% que se refleja en la Tabla 62, refleja, en lenguaje simple, un apoyo técnico parcial pero condicional: hay bases para considerar que el examen funciona técnicamente, pero no hay pruebas suficientes para asegurar que su uso produce las consecuencias deseadas sin realizar las investigaciones y auditorías recomendadas (Kuh et al., 2008; Tinto, 1993).

En este caso se puede proceder con: (a) un seguimiento longitudinal de cohortes admitidas para observar calificaciones y permanencia (Kuh et al., 2008; Tinto, 1993); (b) auditorías periódicas de DIF con reglas claras para ajustar o reemplazar ítems en caso de sesgo (Magis et al., 2010); (c) informes públicos anuales que expliquen cortes, prelación y excepciones, en línea con la transparencia y uso responsable exigidos por los Estándares (AERA et al., 2014); y (d) un comité conjunto UABC–

desarrollador que supervise estas tareas según buenas prácticas de gobernanza técnica (Eignor, 2013). Con ello se habilita un ciclo de mejora continua de validez y consecuencias, cuidando equidad y rendición de cuentas (Sackett et al., 2008; Zumbo y Chan, 2014).

Tabla 62.
Evaluación de la Inferencia de implicación de consecuencias

Supuesto	Descripción	Claridad (1–4)	Coherencia (1–4)	Plausibilidad (1–4)	Puntaje global (3–12)
S 7 . 1 . 1 Responsabilidad conjunta y prelación justa	Los estudiantes con los puntajes adecuados, según el orden de prelación, obtienen admisión a la universidad (Estándar 6.10).	2	2	2	6 (50 %, Baja)
S 7 . 1 . 2 Los puntajes del ExIES permiten clasificar de manera imparcial y útil.	Los puntajes del ExIES permiten clasificar de manera imparcial y útil. La población objetivo y la representatividad muestral se documentan rigurosamente (Estándar 13.1), y las diferencias de desempeño entre grupos se reportan con el contexto necesario y advertencias sobre usos indebidos (Estándar 13.6).	2	2	2	6 (50 %, Baja)
	Global	6 (50 %, Baja)	6 (50 %, Baja)	6 (50 %, Baja)	12 / 24 (50 %, Baja)

Reunidos los hallazgos de las siete inferencias—definición de dominio, evaluación, generalización, explicación, extrapolación, utilización e implicación de consecuencias—y considerando que todas reportan niveles de cumplimiento entre baja, moderado y alto, resulta procedente avanzar a la valoración global del ExIES.

Capítulo 9

Cierre del caso ExIES y orientaciones para aplicar el EBA

Síntesis del argumento de validez del ExIES

El cierre del EBA consistió en expresar el Argumento de Validez: articular la relación entre evidencia, respaldos y recomendaciones en un juicio global. Para ello se integraron las valoraciones por inferencia en la Escala EBA y se consideraron los criterios de claridad, coherencia y plausibilidad (Kane, 2013).

El argumento se fortaleció principalmente por las inferencias con interpretación alta: Generalización (88.8 %) y extrapolación (91.7 %). A su vez, Evaluación (86.9 %) y Explicación (84.8 %) se sostuvieron en niveles moderados a altos, aportando coherencia a la interpretación de las puntuaciones.

Tabla 63.
Resultados según los criterios EBA para la valoración del Argumento Global

Inferencia	Claridad	Coherencia	Plausibilidad	Global	Puntaje	Interpretación
Definición de Dominio	37/44 (84.1 %)	37/44 (84.1 %)	38/44 (86.4 %)	112/ 132 (84.8 %)	10.1	Moderada
Evaluación	24/28 (85.7 %)	25/28 (89.2 %)	24/28 (85.7 %)	73 / 84 (86.9 %)	10.4	Moderada
Generalización	12/12 (100 %)	10/12 (83.3 %)	10/12 (83.3 %)	32/36 (88.8 %)	10.6	Alta
Explicación	20/28 (83.3 %)	20/24 (83.3 %)	18/24 (75 %)	58 / 72 (80.5 %)	9.6	Moderada
Extrapolación	4/4 (100 %)	3/4 (87.5 %)	4/4 (100 %)	11 / 12 (91.7 %)	11	Alta
Utilización	4/8 (50 %)	5/8 (62.5 %)	5/8 (62.5 %)	14/24 (58.3 %)	7	Baja
Implicación de Consecuencias	4/8 (50 %)	4/8 (50 %)	4/8 (50 %)	12/ 24 (50 %)	6	Baja
Global	105/132 (79.5 %)	105/132 (85.4 %)	102/132 (77.2 %)	312/396 (78.7 %)	9.4	Moderada

En contraste, las áreas con mayor margen de mejora se concentraron en Definición de Dominio (84.8 %, moderada) y, especialmente, en Utilización (58.3 %, baja) e Implicación de Consecuencias (50 %, baja). Estos resultados señalaron que el reto no se ubicó solo en medir, sino en documentar mejor el uso institucional y el seguimiento de efectos derivados de las decisiones (AERA et al., 2014). En la Tabla 63 se integra la síntesis final del caso ExIES, que condensa la valoración global del argumento de validez y las orientaciones para transferir el procedimiento a otros contextos (Chapelle, 2021; Kane, 2013).

Aprendizajes y recomendaciones para aplicar el EBA

Con base en los resultados del AIU y la Escala EBA, la evidencia reunida respaldó en términos generales la interpretación y el uso de los puntajes del ExIES como apoyo para seleccionar aspirantes a licenciatura en la UABC en 2023-2. Para dimensionar este juicio, conviene recordar que las pruebas transitan por procesos de validación largos y continuos (AERA et al., 2014; Kane, 2013; Markus y Borsboom, 2013), y el ExIES es una prueba joven: transitó del pilotaje 2022-2 a su primera aplicación operativa en 2023-1 (28 205 aspirantes) y mantiene su vinculación curricular con el MCCEMS (SEP, 2008a, 2008b). En contraste, programas consolidados como el SAT o el ACT disponen de décadas de documentación técnica continua (manuales, estudios de equiparación y de validez predictiva) y amplias fuentes de evidencia accesibles públicamente (College Board, 2023a; ACT, 2024).

En América Latina, los programas nacionales suelen publicar matrices/especificaciones, informes de resultados y materiales operativos que aportan cobertura parcial de las fuentes de evidencia que indican los estándares (AERA et al., 2014); al mismo tiempo, rara vez articulan públicamente un AIU organizado por inferencias, como propone el EBA (INEP, 2023; DEMRE, 2021; ICFES, 2024a; Lavery et al., 2020). En el caso de Ceneval (México), por mencionar algún examen nacional, la disponibilidad pública de documentación técnica completa o de un AIU explícito es limitada. Por lo que iniciar con este tipo de declaraciones nos lleva al camino de la transparencia y mejora continua.

Por lo anterior, los resultados de la presente investigación abonan en dos direcciones: 1) en la interpretación del uso de los puntajes del ExIES a través de sus evidencias (partiendo de sus propias fuentes de datos) y siendo evaluados por los tres criterios establecidos (claridad, coherencia y plausibilidad) según su AIU, de forma pública y transparente; y 2) hacer pública, de forma pragmática, la metodología seguida para el desarrollo del AIU a partir del EBA, exponiendo inferencias, supuestos y las evidencias iniciales que las sostienen. Es así como, gracias a los criterios establecidos (autoevaluación), se puede decir que el ExIES mantuvo resultados altos (extrapolación y generalización), moderados a altos (evaluación, definición de dominio y explicación), pero también bajos, como lo fue el caso de la inferencia de utilización e implicación de consecuencias.

Este resultado se interpretó a la luz de un principio clave: la validez se construyó como un proceso continuo y acumulativo, no como un veredicto definitivo (AERA et al., 2014; Kane, 2013). En un examen joven como el ExIES, la prioridad no solo fue "tener evidencia", sino conservar trazabilidad, actualizar reportes por ciclo y transparentar el razonamiento para distintos públicos.

Los patrones encontrados coincidieron con revisiones sistemáticas previas: la literatura suele documentar con mayor detalle la confiabilidad y la relación con criterios externos (generalización y extrapolación), mientras que utilización y consecuencias aparecen menos, pese a ser críticas en decisiones de alto impacto (Dursun y Li, 2021; Lavery et al., 2020). En el caso ExIES, las inferencias con valoración más alta fueron extrapolación (91.7 %) y generalización (88.8 %). La evaluación (86.9 %) y la explicación (84.8 %) se sostuvieron en niveles moderados a altos. Definición de dominio alcanzó un nivel moderado (84.8 %), mientras que Utilización (58.3 %) e Implicación de Consecuencias (50 %) quedaron como puntos de mejora prioritaria.

A nivel operativo, estos resultados sugirieron una ruta de mejora: mantener y publicar la trazabilidad dominio-especificaciones-ítems; reforzar el reporte de aplicación con incidencias, adaptaciones y auditorías de procedimiento; sostener la comparabilidad entre formas con reportes psicométricos consistentes; ampliar la evidencia de estructura interna

y DIF por subgrupos; e institucionalizar estudios de seguimiento y de impacto para el uso de puntajes.

En particular, para utilización e implicación de consecuencias, se recomendó documentar de forma pública las reglas de decisión (por ejemplo, ponderaciones, umbrales y criterios de asignación), y generar un sistema de monitoreo que vinculara resultados de admisión con desempeño y permanencia. Sin estos productos, el argumento tendió a depender de supuestos difíciles de defender ante comunidades educativas (AERA et al., 2014).

Como aporte editorial y metodológico, estos resultados integraron tablas por inferencia para hacer visible el razonamiento y facilitar su replicación. Este formato dialogó con propuestas recientes que emplearon matrices tipo Toulmin o esquemas equivalentes para evitar que la evidencia quedara dispersa (Choi, 2021, 2022; Fechter et al., 2021).

En conjunto, la revisión permitió pasar de "tener reportes" a "tener un argumento": un hilo lógico que conectó diseño, aplicación, resultados y decisiones. Bajo esta lógica, el ExIES contó con una base técnica relevante para sostener su uso en 2023-2, y al mismo tiempo mostró con claridad qué evidencias debían producirse para fortalecer el examen en convocatorias futuras (Chapelle, 2021; Kane, 2013).

Un aprendizaje adicional fue que el EBA no es exclusivo de los exámenes de admisión. Aunque aquí se aplicó a un instrumento sumativo y de alto impacto, la misma lógica puede sostener decisiones en evaluación formativa: interpretar desempeños para retroalimentar, ajustar la enseñanza y acordar metas de mejora. En este escenario, el AIU suele enfatizar el uso pedagógico de la información y la comunicación de criterios, más que la selección o certificación (Brookhart, 2013; Shepard, 2006).

En términos operativos, una adaptación formativa consiste en redactar un AIU breve (por ejemplo, para una rúbrica o una evaluación de unidad) y revisar, al menos, cuatro inferencias: (1) definición de dominio, para alinear tareas y criterios con los aprendizajes esperados; (2) evaluación, para asegurar instrucciones, condiciones y criterios de calificación consistentes; (3) explicación, para comprobar que el patrón de respuestas se interpreta de manera coherente con la habilidad o desempeño que se busca desarrollar; y (4) utilización/consecuencias, para vigilar que la retroalimentación y las decisiones derivadas promuevan oportunidades

de aprendizaje sin generar efectos inequitativos o desmotivadores (AERA et al., 2014; Kane, 2013; Shepard, 2016).

Alcances y límites de la evidencia

Como en cualquier revisión documental, las limitaciones se relacionaron con la disponibilidad y la actualización de fuentes. El ExIES se encontraba en evolución; por ello, algunos documentos cambiaron por ciclo y no siempre existieron versiones consolidadas o accesibles de todas las evidencias relevantes para cada inferencia.

Además, la evaluación se apoyó en la evidencia disponible para 2023-1 y 2023-2 y priorizó ciertos análisis en función del acceso a bases de datos institucionales. Estudios complementarios —como análisis cualitativos de procesos de respuesta o seguimientos longitudinales de consecuencias— no se desarrollaron por restricciones de tiempo, acceso o prioridad institucional.

Finalmente, la Escala EBA se utilizó como herramienta de autoevaluación para traducir evidencia en un juicio comunicable. Sus porcentajes funcionaron como referencia, pero no sustituyeron el juicio experto ni el contraste con nueva evidencia. Estas limitaciones delinearon una agenda de mejora continua más que un cierre definitivo. En conjunto, esta valoración ofreció un punto de referencia para la mejora continua del ExIES: consolidó lo ya documentado en las inferencias con evidencia alta y orientó prioridades en aquellas donde la evidencia fue baja o incipiente, en particular utilización e implicación de consecuencias.

Epílogo

El EBA como práctica pública de evaluación

Más allá del caso ExIES, el EBA puede leerse como una respuesta práctica a una tensión conocida en los exámenes de alto impacto: la necesidad de decisiones eficientes y comparables, frente al riesgo de inequidad, opacidad o reducción del aprendizaje a un solo número. La crítica a las pruebas de admisión ha señalado, entre otros puntos, sus efectos en trayectorias educativas, la reproducción de desigualdades y la presión institucional por "rendir" en términos de puntajes (Bennett, 2015; French et al., 2024; Sackett et al., 2008; UNESCO, 2021). En ese escenario, el aporte del EBA no consiste en prometer neutralidad, sino en volver discutibles y revisables las decisiones: explicitar supuestos, mostrar evidencia, reconocer reservas y monitorear consecuencias cuando los puntajes se usan para asignar oportunidades (AERA et al., 2014; Kane, 2013; Messick, 1989).

Esta lógica también resulta fértil para usos educativos más extensos, como la evaluación formativa. Cuando el propósito es apoyar el aprendizaje, el AIU se redefine: los puntajes o niveles no buscan seleccionar, sino retroalimentar, ajustar la enseñanza y orientar apoyos. Bajo esa finalidad, el EBA ayuda a que docentes y equipos definan criterios, reúnan evidencia pertinente y revisen la consistencia de sus inferencias sin convertir la evaluación en un trámite. La advertencia central es que los criterios funcionan como instrumentos de autoevaluación sujetos a revisión: deben calibrarse en contexto, con atención a la equidad y a la interpretación compartida entre evaluadores (Brookhart, 2013; Popham, 2008; Sambell et al., 2012; Shepard, 2006).

Finalmente, el EBA recuerda que la validez no se agota en la técnica: implica una ética de la verdad pública. Validar no equivale a declarar certezas finales; equivale a construir afirmaciones que puedan resistir la crítica, incorporar evidencia nueva y reconocer sus límites. Desde un horizonte falibilista, la objetividad se vuelve un logro argumentativo más que una propiedad "natural" del instrumento (Peirce, 1878;

Rorty y Habermas, 2012; Toulmin, 2003). En términos institucionales, documentar inferencias y reservas también es una forma de memoria: preserva el razonamiento que hizo interpretable un puntaje y evita que la organización dependa de recuerdos fragmentarios o de prácticas heredadas sin justificación (Ricoeur, 2004).

Referencias

Abu Dabrh, A. M., Waller, T. A., Bonacci, R. P., Nawaz, A. J., Keith, J. J., Agarwal, A., ... Angstman, K. B. (2020). Professionalism and inter-communication skills (ICS): A multi-site validity study assessing proficiency in core competencies and milestones in medical learners. *BMC Medical Education, 20*(1), Article 2290. https://doi.org/10.1186/s12909-020-02290-3

ACT, Inc. (2023). *ACT national profile report: Graduating class of 2023.* ACT, Inc.

ACT, Inc. (2024). *ACT technical manual 2024.* ACT, Inc.

American Educational Research Association [AERA], & National Council on Measurement in Education [NCME]. (1955). *Technical recommendations for achievement tests.*

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

American Psychological Association [APA]. (1954). *Technical recommendations for psychological tests and diagnostic techniques*.

Andersen, N. B., O'Neill, L., Gormsen, L. K., Hvidberg, L., & Morcke, A. M. (2014). A validation study of the psychometric properties of the Groningen Reflection Ability Scale. *BMC Medical Education, 14*(1), Article 214. https://doi.org/10.1186/1472-6920-14-214

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *Taxonomía del aprendizaje, la enseñanza y la evaluación: La revisión de los objetivos de la educación de Bloom* (Edición en español). Pearson Educación.

Asociación Nacional de Universidades e Instituciones de Educación Superior [ANUIES]. (2024). *Anuario estadístico de la población escolar en educación superior 2023–2024* (Versión 1.2, última actualización: 26 de septiembre de 2024). https://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new Computer Adaptive Test of Size and Strength (CATSS): Development and validation. *Language Assessment Quarterly, 16*(4), 418–437. https://doi.org/10.1080/15434303.2019.1649409

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive domain*. David McKay Company.

Bond, T., & Fox, C. M (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). Information Age Publishing.

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

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brennan, R. L. (2006). Generalizability theory. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 203–234). Praeger.

Brijmohan, A., Khan, G. A., Orpwood, G., Sandford Brown, E., & Childs, R. A. (2018). Collaboration between content experts and assessment specialists: Using a validity argument framework to develop a college mathematics assessment. *Canadian Journal of Education, 41*(2), 584–600. https://journals.sfu.ca/cje/index.php/cje-rce/article/view/3239

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD.

Burkov, A. (2019). *The hundred-page machine learning book.* Andriy Burkov.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Carrillo-Ávalos, B. A., Leenen, I., Trejo-Mejía, J. A., & Sánchez-Mendiola, M. (2024). Evidencias de validez del proceso de admisión a una escuela de medicina en México. *Investigación en Educación Médica, 13*(50), 37–55. https://doi.org/10.22201/fm.20075057e.2024.50.23546

Caso, J., & Díaz, C. D. (2016). *Guía para la Evaluación de Ítems del Nuevo Examen de Selección de aspirantes a ingresar a la Universidad Autónoma de Baja.* Instituto de Investigación y Desarrollo Educativo-Universidad Autónoma de Baja California.

Caso, J., Díaz, C. D., Castro-Morera, M., & Martínez-Arias, M. R. (2017). *Manual técnico del Examen de Ingreso a la Educación Superior (ExIES).* Universidad Autónoma de Baja California.

Ceneval. (2022). *Informe anual de resultados 2021*. Centro Nacional de Evaluación para la Educación Superior. https://ceneval.edu.mx/wp-content/uploads/2022/06/Ceneval-Informe-Anual-de-Resultados-2021.pdf

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment.* SAGE. https://doi.org/10.4135/9781071878811

Choi, Y. (2021). *What interpretations can we make from scores on graphic-prompt writing (GPW) tasks? An argument-based approach to test validation.* Assessing Writing, 48, Article 100523. https://doi.org/10.1016/j.asw.2021.100523

Choi, Y. (2022). Validity of score interpretations on an online English placement writing test. *Language Testing in Asia, 12*(42). https://doi.org/10.1186/s40468-022-00187-0

College Board. (2023a). *SAT Suite of Assessments Annual Report 2023.* https://reports.collegeboard.org

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

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* Cengage Learning.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* John Wiley.

Cronbach, L. J., Shavelson, R. J., & Webb, N. M. (2004). Generalizability theory: 1973–2003. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404264844

Cureton, E. E. (1951). Validity. En E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.

Durson, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). En C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing* (pp. 45–70). Cambridge University Press. https://doi.org/10.1017/9781108669849.005

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). En C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing* (pp. 45–70). Cambridge University Press. https://doi.org/10.1017/9781108669849.005

Eignor, D. R. (2013). The standards for educational and psychological testing. En K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology (pp. 245–250). American Psychological Association. https://doi.org/10.1037/14047-013

Fechter, T., Dai, T., Cromley, J. G., Nelson, F. E., Van Boekel, M., & Du, Y. (2021). Developing a validity argument for an inference-making and reasoning measure for use in higher education. *Frontiers in Education, 6,* Article 727539. https://doi.org/10.3389/feduc.2021.727539

French, M., Juárez, C., & Stone, A. (2024). The role of high-stakes testing in higher education admissions: Global perspectives. J*ournal of Educational Assessment, 22*(1), 45–63. https://doi.org/10.1007/s10734-023-01148-z

García, A. M., Martínez, F., & Cordero, G. (2016). Análisis del funcionamiento diferencial de los ítems del Excale de Matemáticas para tercero de secundaria. *Investigación, 21*(71), 1191–1210.

García, A., Martínez, F., Cordero, G. y Caso, J. (2017). Evolución del concepto de validez en la medición educativa. En E. Luna y G. Cordero (Coords.), *Contribuciones a la evaluación educativa desde la formación doctoral* (pp. 15-46). UdeG/UABC.

García, M. (2016). Evidencias de validez predictiva en exámenes de ingreso a la educación superior: Comparación entre PAA y EXANI II. *Revista Latinoamericana de Medición y Evaluación Educativa, 11*(2), 15–29.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media.

Gotch, C. M., & French, B. F. (2020). A validation trajectory for the Washington Assessment of Risks and Needs of Students. *Educational Assessment, 25*(1), 65–82. https://doi.org/10.1080/10627197.2019.1702462

Gutiérrez, M. J. (2024). Uma breve história dos testes de alto impacto e seus possíveis futuros. *Estudos em Avaliação Educacional, 35*, e11050. https://doi.org/10.18222/eae.v35.11050

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology, 64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (3rd ed.). Routledge.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. En D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hatala, R., Gutman, J., Lineberry, M., et al. (2019). How well is each learner learning? Validity investigation of a learning curve-based assessment approach for ECG interpretation. *Advances in Health Sciences Education, 24*(1), 45–63. https://doi.org/10.1007/s10459-018-9846-x

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), T*est Validity* (pp. 129–145). Lawrence Erlbaum Associates.

House, E. R. (1980). Evaluating with validity. SAGE.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

ICFES. (2024a). *Informe nacional de resultados del examen Saber 11º – 2022.* Instituto Colombiano para la Evaluación de la Educación. https://icfes.gov.co

ICFES. (2024b, noviembre). ¿Qué se entiende por confiabilidad y validez en el contexto de la medición con instrumentos? *Boletín Saber al Detalle*, Edición *16*, 1–11. https://icfes.gov.co

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*(2), 276–299. https://doi.org/10.1177/02655322221112364

Jones, M. G., & Ennes, M. (2018). *High-stakes testing. En Oxford Bibliographies.* https://doi.org/10.1093/obo/9780199756810-0200

Jornet, J., González-Such, J., & Suárez, J. M. (2010). Validación de los procesos de determinación de estándares de interpretación para pruebas de rendimiento educativo. *Estudios Sobre Educación, 19*, 11–29. https://doi.org/10.15581/004.19.4578

Kane, M. (2006). Content-related validity evidence in test development. En S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Kane, M. (2011). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. En S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2ª ed., pp. 64–80). Routledge.

Kane, M. (2020). Validity studies commentary. *Educational Assessment, 25*(1), 83–89. https://doi.org/10.1080/10627197.2019.1702465

Kane, M. T. (1990). An argument-based approach to validation (ACT Research Report Series, Report No. ACT-RR-90-13). *American College Testing Program.* https://eric.ed.gov/?id=ED336428

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

Kane, M., & Bridgeman, B. (2021). The evolution of the concept of validity. En B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (1.ª ed., pp. 174–195). Routledge. https://doi.org/10.4324/9780367815318

Koizumi, R., In'nami, Y., Asano, K., & Agawa, T. (2016). Validity evidence of Criterion® for assessing L2 writing proficiency in a Japanese university context. *Language Testing in Asia, 6*, Article 5. https://doi.org/10.1186/s40468-016-0027-7

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3ª ed.). Springer.

Koselleck, R. (2000). *Los estratos del tiempo: Estudios sobre la historia.* Paidós Ibérica.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on college grades and persistence. *The Journal of Higher Education, 79*(5), 540–563. https://doi.org/10.1080/00221546.2008.11772116

Lane, S., Raymond, R., & Haladyna, T. (2016). *Validation of score meaning for the next generation of assessments*. Routledge.

Lavery, M., Bostic, J., Kruse, L., Krupa, E., & Carney, M. (2020). Argumentation surrounding argument-based validation: A systematic review of validation methodology in peer-reviewed articles. *Educational Measurement: Issues and Practice, 40*(1), 22–33. https://doi.org/10.1111/emip.12378

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lee, E. (2020). Evaluating test consequences based on ESL students' perceptions: An appraisal analysis. *Studies in Applied Linguistics & TESOL, 20*(1), 1–22. https://doi.org/10.7916/salt.v20i1.3394

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., ... Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. *Implementation Science, 13*(Suppl 1), 2. https://doi.org/10.1186/s13012-017-0688-3

Li, S. (2018). Developing a test of L2 Chinese pragmatic comprehension ability. *Language Testing in Asia, 8*, (3). https://doi.org/10.1186/s40468-018-0054-7

Lissitz, R. (2009) *The Concept of Validity. Revisions, New Directions, and Applications.* Charlotte, NC: Information Age Publishing, Inc. 263 pages. ISBN 978-1-60752-227-0

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental tests scores. Reading.* Addison-Wesley.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Marcinek, T., Jakobsen, A., & Partová, E. (2023). Using MKT measures for cross-national comparisons of teacher knowledge: Case of Slovakia and Norway. *Journal of Mathematics Teacher Education, 26*(3), 303–333. https://doi.org/10.1007/s10857-021-09530-3

Marini, J. P., Westrick, P. A., Young, L., Ng, H., & Shaw, E. J. (2023, abril). *Digital SAT® pilot predictive validity study – A first look.* College Board.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.* Routledge.

Mattos, P., Stieg, R., Barcelos, M., & Santos, W. dos. (2024). Evaluaciones nacionales a gran escala y acceso a la educación superior: perspectivas en países de América y Europa. *Contextos: Estudios de Humanidades y Ciencias Sociales*, *54*, 1–25. https://revistas.umce.cl/index.php/contextos/article/view/2660

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41–55. https://doi.org/10.1016/j.asw.2017.12.003

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3ª ed., pp. 13–103). Macmillan.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and Assessment in Teaching* (10ª ed.). Pearson Education.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: interdisciplinary Research and Perspectives, 1*, 3-62.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5ª ed.). Wiley. https://doi.org/10.1002/9781118532843

Morales, R., Barrera, A., & Garnett, E. (2015). *Validez predictiva y concurrente del EXANI-II en la Universidad Autónoma del Estado de México. En Memorias del X Congreso Nacional de Investigación Educativa: Sujetos de la educación.* Consejo Mexicano de Investigación Educativa (COMIE). https://www.comie.org.mx/congreso/memoriaelectronica/v10/pdf/area_tematica_16/ponencias/0701-F.pdf

Newton, P., & Shaw, S. (2014). *Validity in educational & psychological assessment.* SAGE.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3.a ed.). McGraw-Hill.

Pedroza Zúñiga, L. H. & Gómez Monárrez, C. (2025a). *Informe particular ExIES vs EXANI vs Promedio* [Manuscrito no publicado].

Pedroza Zúñiga, L. H. & Gómez Monárrez, C. (2025b). *Informe general ExIES vs EXANI vs Promedio* [Manuscrito no publicado].

Pedroza Zúñiga, L. H. & Gómez Monárrez, C. (2025c). Funcionamiento *Diferencial del ítem (DIF): Examen de Ingreso a la Educación Superior (ExIES) 2023-2* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., & Gutiérrez Zavala, A. P. (2023n). *Especificaciones de Lectura.* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., & Ruiz Mendoza, K. K. (2023o). *Especificaciones de Lengua Escrita*. [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023a). *Manual para el desarrollo de reactivos: Lectura* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023b). *Manual para el desarrollo de reactivos: Lengua Escrita* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023c). *Manual para el desarrollo de reactivos: Matemáticas* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023d). *Manual para el jueceo de reactivos: Lectura* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023e). *Manual para el jueceo de reactivos: Lengua escrita* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023f). *Manual para el jueceo de reactivos: Matemáticas* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023h). *Presentación de las capacitaciones para el desarrollo de ítems ExIES* [Diapositivas no publicadas].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023i). *Manual del aplicador del ExIES* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023j). *Manual del supervisor del ExIES* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023k). P*resentación de capacitación para aplicadores y supervisores del ExIES* [Diapositivas no publicadas].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023l). *Guía del sustentante del ExIES* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Gómez Monárrez, C., Orozco Vergara, M. A., Ruiz Mendoza, K. K., & Gutiérrez Zavala, A. P. (2023m). *Protocolos para incidencias en caso de siniestro o emergencia durante la aplicación del ExIES* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., García Aldaco, S. A., Orozco Vergara, M. A. & Gómez Monárrez, C., Verdugo Olachea, J. (2023p). *Especificaciones de Matemáticas.* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. & Solís del Moral, S. S. (2024a). E*xamen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico.* Instituto de Investigación y Desarrollo Educativo.

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. & Solís del Moral, S. S. (2024b). *Examen de ingreso a la educación superior (ExIES) 2023-2: Reporte técnico*. Instituto de Investigación y Desarrollo Educativo.

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. A., & Solís del Moral, S. S. (2023s). *ExIES Base de datos completa de Resultados Rasch y estadísticas ítem–forma* [Base de datos no publicada]. Instituto de Investigación y Desarrollo Educativo, Universidad Autónoma de Baja California.

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. A., & Solís del Moral, S. S. (2024c). *Base de datos de organización de ítems, histórico del ExIES: control de ítems NDC-especificación-contenido* [Manuscrito no publicado].

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. A., & Solís del Moral, S. S. (2023q). *Base de datos del jueceo de ítems del ExIES: Lengua escrita, Lectura y Matemáticas* [Conjunto de datos no publicado].

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. A., & Solís del Moral, S. S. (2023r). *Reporte de aplicación del Examen de Ingreso a la Educación Superior (ExIES) 2023-1*. [Manuscrito no publicado].

Pedroza Zúñiga, L. H., Gómez Monárrez, C., García Aldaco, S. A., Orozco Vergara, M. A., & Vargas Ceseña, A. N. (2022). *Examen de ingreso a la educación superior (ExIES) 2022-2: Manual Técnico* [Manual técnico]. Instituto de Investigación y Desarrollo Educativo, Universidad Autónoma de Baja California.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly, 13*, 470–482.

Popham, W. J. (2008). *Transformative assessment*. ASCD.

Rafatbakhsh, E., & Ahmadi, A. (2022). The Argument-Based Validation of a Large-Scale High-Stakes Vocabulary Test. *Practical Assessment, Research, and Evaluation, 27*. https://scholarworks.umass.edu/pare/vol27/iss1/28

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Ricoeur, P. (2004). *Memory, history, forgetting.* University of Chicago Press. (Trabajo original publicado en 2000).

Rorty & Habermas (2012). *Sobre la verdad: ¿validez o justificación? Amorrortu.*

Ruiz Mendoza, K., Pedroza Zúñiga, L., & López García, A. (2025). Validation of tests using an argument-based approach: a review based on the PRISMA model. *Sapienza, International Journal of Interdisciplinary Studies, 6*(4). https://doi.org/10.51798/sijis.v6i4.1177

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215–227. https://doi.org/10.1037/0003-066X.63.4.215

Sambell, K., McDowell, L., & Montgomery, C. (2012). *Assessment for Learning in Higher Education* (1st ed.). Routledge. https://doi.org/10.4324/9780203818268

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Secretaría de Educación Pública [SEP]. (2008a, 27 de junio). Acuerdo número 442 por el que se establecen los Lineamientos [... título exacto del acuerdo ...]. *Diario Oficial de la Federación*. https://educacionmediasuperior.sep.gob.mx/work/models/sems/Resource/11435/1/images/5_1_acuerdo_numero_442_establece_snb.pdf

Secretaría de Educación Pública [SEP]. (2008b, 27 de junio). Acuerdo número 444 por el que se expiden los Lineamientos [... título exacto del acuerdo ...]. *Diario Oficial de la Federación.* https://educacionmediasuperior.sep.gob.mx/work/models/sems/Resource/11435/1/images/5_2_acuerdo_444_competencias_mcc_snb.pdf

Shepard, L. (2006). *Classroom assessment. En R. L. Brennan (Ed.), Educational measurement* (4ª ed., pp. 623–646). Praeger.

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

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299–321. https://doi.org/10.1207/s15326977ea0504_2

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*(2-3), 108–131. https://doi.org/10.1080/10627190802394255

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation-based assessment of clinical competence. *Advances in Health Sciences Education, 23*(2), 323–338. https://doi.org/10.1007/s10459-017-9800-3

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). University of Chicago Press.

Toulmin, S. E. (2003). *The uses of argument* (ed. actualizada). Cambridge University Press. (Trabajo original publicado en 1958).

UNESCO. (2021). *Learning assessment and high-stakes exams.* IIEP Learning Portal. https://learningportal.iiep.unesco.org/en/library/learning-assessment-and-high-stakes-exams

Universidad Autónoma de Baja California (2024). *Base de datos del promedio del primer y segundo semestre de universidad*. [Conjunto de datos no publicado].

Universidad Autónoma de Baja California. (2010). Ley Orgánica de la Universidad Autónoma de Baja California. *Periódico Oficial del Estado de Baja California.* https://sriagral.uabc.mx/Externos/AbogadoGeneral/Reglamentos/Leyes/01_LEY_ORGANICA_UABC_reforma_2010.pdf

Universidad Autónoma de Baja California. (2019). *Estatuto General de la Universidad Autónoma de Baja California.* https://sriagral.uabc.mx/Externos/AbogadoGeneral/Reglamentos/Leyes/02_EstatutoGeneralUABC_19-11-2019.pdf

Universidad Autónoma de Baja California. (2021). *Estatuto Escolar de la Universidad Autónoma de Baja California* (Edición especial No. 460). Gaceta UABC. https://sriagral.uabc.mx/externos/abogadogeneral/Reglamentos/Estatutos/03_EstatutoEscolarUABC_Reforma_May_202021.pdf#:~:text=XXIII,aspirantes%20para%20el%20nuevo%20ingreso

Watson, G. (2002). *The Modern Mind: An Intellectual History of the 20th Century.* Harper.

Yan, X., & Staples, S. (2019). Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE. *Language Testing, 36*(1), 1–26. https://doi.org/10.1177/0265532219876226

Zhu, W. (2001). Book Review. *Measurement in Physical Education and Exercise Science*, *5*(4), 251–254. https://doi.org/10.1207/S15327841MPEE0504_05

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. En P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates.

Zumbo, B. D., & Chan, E. K. H. (2014). Validity and validation in social, behavioral, and health sciences. *Springer*. https://doi.org/10.1007/978-3-319-07794-9

Estudios económicos y sociales. Tomo I
Se terminó de imprimir en marzo de 2026
en los talleres de Astra Ediciones
Av. Acueducto No. 829
Colonia Santa Margarita, C. P. 45140
Zapopan, Jalisco, México.
33 38 34 82 36
E-mail: edicion@astraeditorial.com.mx
www.astraeditorialshop.com
Impresión digital con interiores en papel bond de 75 g.
El tiraje consta de 300 ejemplares

De la teoría a la práctica:
Implementación del Enfoque Basado en Argumentos en el Examen de Ingreso a la Educación Superior (ExIES)

Este libro-guía presenta una ruta de trabajo para aplicar el Enfoque Basado en Argumentos (EBA) en el proceso de validación de un Examen de Alto Impacto. Se parte de la idea de que la validez se sostiene en la interpretación y el uso de los puntajes, y no en el instrumento en sí mismo. Como caso ilustrativo, se revisa el Examen de Ingreso a la Educación Superior (ExIES), con énfasis en la aplicación 2023-2 de la Universidad Autónoma de Baja California (UABC). En esta ruta se recupera y organiza la documentación técnica (especificaciones, manuales y reportes psicométricos), normas institucionales y evidencias cuantitativas; lo cual permite la redacción del Argumento de Interpretación y Uso (AIU), para después evaluar el Argumento de Validez como sus siete inferencias encadenadas, desde Definición de Dominio hasta Utilización e Implicación de Consecuencias que parte de la teoría del proceso de validación, según el EBA, desde Michael Kane y Carol Chapelle.

Karla Karina Ruiz Mendoza

ISBN: 979-13-88142-59-8
9 791388 142598

Consulta y descarga

astra editorial