UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

D O C T O R A D O

Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos

Tesis
Que para obtener el grado de

Doctora en Ciencias Educativas

P r e s e n t a

Karla Karina Ruiz Mendoza

Ensenada, Baja California, México
Octubre, 2025

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Doctorado en Ciencias Educativas

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos"

TESIS

Que para obtener el grado de

DOCTOR(A) EN CIENCIAS EDUCATIVAS

Presenta

Karla Karina Ruiz Mendoza

APROBADO POR:

Dr. Luis Horacio Pedroza Zúñiga
Director de tesis

Dra. Alma Yadhira López García
Sinodal

Dra. Edna Luna Serrano
Sinodal

Dr. Carlos David Díaz López
Sinodal

Dr. Adán Moisés García Medina
Sinodal

DCE
DOCTORADO EN CIENCIAS EDUCATIVAS

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

Ensenada, B.C., a 7 de octubre de 2025

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctora en Ciencias Educativas

Dra. Rubí Surema Peniche Cetzal
Coordinadora de Investigación y Posgrado
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. KARLA KARINA RUIZ MENDOZA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo.

Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, con el trabajo titulado:

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos".

Esperando reciba el presente de conformidad, quedo de usted.

Atentamente

DR. LUIS HORACIO PEDROZA ZÚÑIGA

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

Ensenada, B.C., a 7 de octubre de 2025

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctora en Ciencias Educativas

Dra. Rubí Surema Peniche Cetzal
Coordinadora de Investigación y Posgrado
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. KARLA KARINA RUIZ MENDOZA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo.

Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, con el trabajo titulado:

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos".

Esperando reciba el presente de conformidad, quedo de usted.

Atentamente

DRA. ALMA YADHIRA LÓPEZ GARCÍA

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

Ensenada, B.C., a 7 de octubre de 2025

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctora en Ciencias Educativas

Dra. Rubí Surema Peniche Cetzal
Coordinadora de Investigación y Posgrado
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. KARLA KARINA RUIZ MENDOZA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo.

Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, con el trabajo titulado:

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos".

Esperando reciba el presente de conformidad, quedo de usted.

Atentamente

Edna Luna S.

DRA. EDNA LUNA SERRANO

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

Ensenada, B.C., a 7 de octubre de 2025

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctora en Ciencias Educativas

Dra. Rubí Surema Peniche Cetzal
Coordinadora de Investigación y Posgrado
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. KARLA KARINA RUIZ MENDOZA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo.

Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, con el trabajo titulado:

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos".

Esperando reciba el presente de conformidad, quedo de usted.

Atentamente

DR. CARLOS DAVID DÍAZ LÓPEZ

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
Instituto de Investigación y Desarrollo Educativo

Ensenada, B.C., a 7 de octubre de 2025

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctora en Ciencias Educativas

Dra. Rubí Surema Peniche Cetzal
Coordinadora de Investigación y Posgrado
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. KARLA KARINA RUIZ MENDOZA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo.

Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, con el trabajo titulado:

"Validación del Examen de Ingreso a la Educación Superior a través del Enfoque Basado en Argumentos".

Esperando reciba el presente de conformidad, quedo de usted.

Atentamente

DR. ADÁN MOISÉS GARCÍA MEDINA

Agradecimientos

Expreso mi agradecimiento a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por la beca de estudio otorgada, permitiendo la culminación de la presente investigación desarrollada en el marco de mis estudios de posgrado en el Instituto de Investigación y Desarrollo Educativo (IIDE) de la Universidad Autónoma de Baja California (UABC).

A la comunidad del IIDE-UABC, gracias por el rigor, la guía y la confianza. A la dirección del instituto, a la coordinación de posgrado, al personal académico, técnico y administrativo, y al equipo del ExIES (quienes apoyaron el acceso a datos y resguardo de la información): su profesionalismo sostuvo cada etapa del proyecto.

A quienes guiaron este proceso y me dieron diferentes tipos de aprendizajes, al Dr. Luis Horacio Pedroza Zúñiga, Dra. Alma Yadhira López García, Dra. Edna Luna Serrano, Dr. Carlos David Díaz López y Dr. Adán Moisés García Medina— les agradezco su lectura atenta y sus orientaciones exigentes y generosas. Sus comentarios fortalecieron la claridad, coherencia y plausibilidad de cada uno de los argumentos, así como a la solidez metodológica del trabajo, recordándome que el rigor, la ética y el contexto son inseparables de toda decisión de evaluación.

Con especial respeto y admiración, Dra. Guadalupe Tinajero, Dra. Graciela Cordero, y Dra. Rubi Peniche; quienes con sus palabras y actos alimentaron mi curiosidad como creatividad. Y, finalmente, mis compañeros de la línea de evaluación educativa: Perla Córdova, Karen Rivera y Ricardo González, por su compañía invaluable.

A mi intelectual favorita: *mamá.*

Resumen

La presente tesis evalúa la validez de la interpretación y el uso de los puntajes del Examen de Ingreso a la Educación Superior (ExIES 2023-1) como criterio de admisión a licenciatura en la Universidad Autónoma de Baja California. Se adopta el Enfoque Basado en Argumentos (EBA) de Kane y Chapelle para articular el Argumento de Interpretación y Uso (AIU) sustentado en siete inferencias: Definición de Dominio, Evaluación, Generalización, Explicación, Extrapolación, Utilización e Implicación de Consecuencias. El diseño metodológico es de carácter evaluativo y documental; integra 28 fuentes de datos del ExIES, bases psicométricas históricas, normativas institucionales y bases de datos de rendimiento académico, analizadas mediante la Escala EBA (claridad, coherencia y plausibilidad) y procedimientos estadísticos —entre ellos regresiones lineales y análisis DIF— para estimar confiabilidad, validez predictiva e imparcialidad. Los resultados muestran un Índice de Validez Global de 78.7%, interpretado como moderado, con puntuaciones altas en las inferencias de Definición de Dominio, Evaluación, Generalización, Explicación y Extrapolación (≥ 80 %) y bajas a moderadas en Utilización e Implicación de Consecuencias (58.3% y 50%). Se concluye que el ExIES, a pesar de ser una prueba joven, proporciona evidencia suficiente sobre la interpretación de los puntajes para decisiones de alto impacto. Para la mejora continua, se recomiendan ciclos iterativos que incluyan actualización de evidencias, revisión de criterios institucionales y estudios longitudinales de cohorte para optimizar utilidad y transparencia.

Palabras clave: ExIES; validez basada en argumentos; admisión universitaria; evaluación educativa; UABC.

Contenido

Índice de tablas

Índice de figuras

Listado de acrónimos

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*

Planteamiento del problema

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, denominamos 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 & 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], & 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 & Dang, 2021; Ihlenfeldt & 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 licenciatura, 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 los cuales se admitieron 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 a comparación de pruebas internacionales, e incluso nacionales, como el SAT o el ACT.

Es relevante aclarar que el uso de este instrumento tiene fundamento normativo explícito en la Ley Orgánica de la UABC (UABC, 2010), la cual otorga autonomía a la institución para organizarse y regirse conforme a sus propios fines e intereses. Asimismo, se sustenta en el Estatuto General de la UABC (2019), particularmente en el Título Quinto, Capítulo Único, artículo 171, que establece que los aspirantes deben someterse al proceso de selección determinado por la universidad. Además, el Estatuto Escolar (UABC, 2021) en su artículo 3º, fracción XXIII, define claramente el examen de selección como un instrumento conformado por una o más pruebas que determina la selección de aspirantes para el nuevo ingreso. Este mismo estatuto, en los artículos 15 al 28 del Título Segundo, Capítulo Primero, especifica los criterios, procedimientos y condiciones para la admisión universitaria, subrayando la importancia del examen en el proceso de selección, destacando aspectos clave como la transparencia y la equidad, establecidos explícitamente en el artículo 16 y complementados por el Comité de Equidad mencionado en el artículo 21.

Considerando que los EAI son utilizados para tomar decisiones educativas trascendentales de forma transparente —como la admisión universitaria o la certificación de competencias—, resulta indispensable garantizar que los resultados obtenidos sean apropiados, precisos y justificados. Así, según la AERA, la APA y el NCME (2014), la validez se configura como un concepto esencial para el desarrollo y evaluación de instrumentos. Desde esta perspectiva, el análisis de la validez de los exámenes de admisión se inserta en la evaluación educativa, disciplina interdisciplinaria que integra aportaciones de la psicometría, la psicología educativa y la pedagogía (Scriven, 1991).

Esta relevancia atribuida a la validez se ha consolidado a lo largo de su evolución conceptual, transitando desde la correspondencia lógica entre los contenidos del instrumento y el dominio evaluado —denominada validez de contenido—, así como la relación empírica entre los puntajes y un criterio externo —validez de criterio— (Cureton, 1951), hacia la integración de ambas perspectivas en la clásica tríada que añadió la validez de constructo como tercer elemento esencial (Cronbach & Meehl, 1955); para luego concebir a la validez como un constructo unitario que abarca múltiples fuentes de evidencia, incluidas las repercusiones sociales y éticas de la prueba (Messick, 1989); hasta llegar al Enfoque Basado en Argumentos (EBA) propuesto por Kane (2006, 2013), el cual amplió esta visión al incluir no solo los análisis estadísticos de la prueba, sino también las implicaciones éticas y sociales derivadas de su uso.

En la actualidad, en congruencia con esta perspectiva contemporánea y de acuerdo con los *Standards for Educational and Psychological Testing* (en adelante, Estándares), la validez consiste en la medida en que tanto la evidencia como la teoría fundamentan las interpretaciones que se hacen de los puntajes obtenidos en una prueba, en función de los usos específicos que se planean para dicha evaluación (AERA et al., 2014). Así, el eje de la validez se ubica en la interpretación de los puntajes de la prueba, a partir del uso al que se destinen, y no de la prueba en sí misma (AERA et al., 2014). Además, como apuntan la AERA et al. (2014), Kane (2013), y Markus y Borsboom (2013), el proceso de validación es un proceso largo y continuo, donde se van obteniendo evidencias a través del tiempo, es decir, siempre puede reunirse nueva información y, por tanto, los juicios se sostienen provisionalmente y deben actualizarse cuando cambian las poblaciones o los usos de la prueba.

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. En esta línea, 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 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).

El AUI se articula mediante una estructura lógica, en este caso el Modelo de Toulmin (1958/2003), conformada por inferencias sucesivas que enlazan puntajes observados con

interpretaciones específicas y usos propuestos (Kane, 2006, 2013). Cada inferencia requiere garantías y supuestos claramente establecidos, que aseguran la solidez y justificación del argumento. Las garantías son principios o declaraciones generales que validan la transición lógica entre puntajes, interpretaciones y decisiones; mientras que los supuestos constituyen condiciones necesarias, generalmente contextuales, para que dichas garantías se mantengan válidas (Kane, 2013; Chapelle et al., 2010). Así, a partir de fuentes de datos se desarrollan las evidencias necesarias para respaldar cada supuesto y, por ende, toda la cadena argumentativa, donde cada vez que se genera una nueva evidencia se debe someter nuevamente al contraste y revisión del AIU (Kane, 2013; Chapelle, 2021).

Aunque Kane definió ampliamente la estructura argumentativa general, son Chapelle et al. (2010) quienes enfatizaron explícitamente el papel fundamental que juegan los supuestos específicos asociados a cada garantía, subrayando así la importancia del contexto particular en la validez de las interpretaciones de los puntajes. Diversas Revisiones Sistemáticas de la Literatura (RSL) evidencian que, en la práctica, la implementación del EBA suele realizarse de manera incompleta, predominando la discusión teórica sobre la aplicación empírica (Cook et al., 2013; Dursun & Li, 2021; Hatala et al., 2015; Lavery et al., 2020). Esto genera una marcada tendencia a enfatizar indicadores psicométricos comunes, tales como la confiabilidad y las correlaciones con variables externas, mientras que aspectos cruciales como la equidad, el análisis del proceso de respuesta, la estructura interna detallada y las consecuencias del uso de las puntuaciones son frecuentemente relegados (Cook et al., 2013; Dursun & Li, 2021). Por tanto, para fortalecer las evidencias derivadas del EBA, es relevante equilibrar los aspectos técnicos tradicionales con un análisis exhaustivo de las implicaciones prácticas y éticas asociadas a la interpretación y uso de los resultados (Cook et al., 2013; Hatala et al., 2015).

De igual manera, se han documentado estudios sobre la validez a través del EBA desde enfoques tanto cualitativos como cuantitativos (Durson & Li, 2021; Hatala, 2019), aunque no siempre bajo una integración sistemática. Durson y Li (2021) señalan que la selección y justificación de las inferencias dentro del EBA con frecuencia carece de claridad, pues no se explica de manera transparente cómo se conectan con otras en el marco del argumento general, lo que dificulta la interpretación global. Aunado a ello, numerosos estudios no especifican los supuestos y garantías vinculados a cada inferencia, por lo que se desdibuja su contribución real a la validez (Durson & Li, 2021; Lavery et al., 2020). Asimismo, se describen problemas en la interpretación de las inferencias, ocasionando confusión y variabilidad en la forma en que los investigadores aplican el enfoque (Durson & Li, 2021).

A pesar de la falta de claridad, Durson y Li (2021) constatan un incremento progresivo en la adopción del EBA en trabajos publicados a partir del año 2000, con especial énfasis desde 2005. El surgimiento de este enfoque a partir de Kane (1992, 2006) y su expansión por parte de otros autores (p. ej., Bachman & Palmer, 2010; Chapelle et al., 2008) han contribuido a sistematizar la validación, al proponer el análisis de una cadena de inferencias específicas y examinar la solidez de las interpretaciones y usos que se desprenden de los puntajes.

Una investigación que ejemplifica lo anterior, en el terreno de la validación de EAI, es el de Sánchez Mendiola et al. (2020), orientado al examen de ingreso de la Universidad Nacional Autónoma de México (UNAM). Dicho trabajo tomó como base los marcos de Messick y Kane y reunió múltiples fuentes de evidencia (contenido, estructura interna, consecuencias, entre otras), sin embargo, no declara el AIU ni determina una evaluación específica, es decir, un Argumento de Validez. De forma similar, algunos estudios que se aproximan al EBA, como Fechter et al. (2021) y Carrillo-Ávalos et al. (2024), solo abarcan fases iniciales o descripciones parciales del

proceso, sin culminar en una evaluación explícita y global sobre la adecuación de los usos de los puntajes o definir claramente las evidencias acumuladas; lo mismo sucede en exámenes nacionales como el EXANI-II, del cual no se hacen públicas las evidencias. Ante esta situación, también se hace patente la necesidad de revisar más a fondo la literatura en EAI y así lograr robustecer la metodología de forma clara y coherente en torno al EBA; por ello, en los *Antecedentes* se profundiza a través de una RSL.

Asimismo, la falta de definiciones homogéneas respecto a las inferencias y sus supuestos complica el contraste de hallazgos y la elaboración de lineamientos de validación más unificados. Lavery et al. (2020) refieren que, a pesar del amplio consenso en torno a la importancia de la validación basada en argumentos, persisten debates sobre el empleo de lógicas formales o informales, la flexibilidad o rigidez del proceso.

Con el propósito de fortalecer la coherencia metodológica en el análisis del proceso de validación, Chapelle (2021) propone una versión ampliada y actualizada del EBA, estructurada en torno a siete inferencias lógicamente encadenadas: definición del dominio, evaluación, generalización, explicación, extrapolación, utilización y consecuencias. Esta formulación extiende la propuesta original de Kane (2004, 2006, 2013) y se orienta particularmente al análisis de exámenes de lenguas, como el TOEFL. El desarrollo detallado de este enfoque se presenta en el *Marco Teórico* del presente documento. Chapelle (2021) ofrece ejemplos, esquemas y tablas para ilustrar la estructura argumental y sugiere un enfoque mixto que podría contribuir a resolver los debates sobre qué tipo de evidencia priorizar según cada contexto; siendo, entonces, la mejor propuesta planteada desde EAI. Sin embargo, no se dispone aún de criterios consensuados para calificar la pertinencia de cada uso de los puntajes, ni se han identificado estudios que apliquen

el modelo en su totalidad, lo cual evidencia la demanda de propuestas adicionales (Lavery et al., 2020).

En el caso ExIES, desde sus inicios, se han desarrollado esfuerzos importantes orientados a sustentar técnicamente sus resultados, los cuales constituyen una base valiosa para continuar fortaleciendo la validez de la interpretación de sus puntajes. El Reporte Técnico 2023-1 (Pedroza et al., 2024a), por ejemplo, documenta diversas fuentes de evidencia conforme a los Estándares (AERA et al., 2014), incluyendo evidencia basada en la estructura interna, en relaciones con otras variables, y en aspectos de contenido y confiabilidad. Asimismo, se han incorporado estrategias complementarias como la elaboración de versiones paralelas, la especificación de niveles de demanda cognitiva y el establecimiento de lineamientos de aplicación y seguridad que favorecen la equidad y consistencia del proceso de evaluación.

En este contexto, resulta pertinente considerar enfoques complementarios que permitan integrar de manera más explícita las interpretaciones y usos derivados de los puntajes. El EBA, propuesto por Kane (2006, 2013, 2020), ampliado y profundizado por Chapelle (2021), ofrece una estructura sistemática para construir Argumentos de Validez que articulen las inferencias desde la prueba hasta las decisiones educativas que esta informa. Este marco puede enriquecer las prácticas actuales, al proporcionar una base lógica para evaluar la coherencia y adecuación de los usos propuestos de los resultados.

Aunque en el contexto mexicano existen antecedentes de aplicación del EBA —como el caso documentado en la Universidad Nacional Autónoma de México (Sánchez Mendiola et al., 2020)— se presentan oportunidades para avanzar en su implementación en otros exámenes de admisión. En el caso de la UABC, el proceso de la validación del ExIES presenta una oportunidad para robustecer la interpretación de sus puntajes con dicho enfoque.

En esta línea, la presente investigación propone validar al ExIES a partir del EBA, mediante la construcción de un Argumento de Validez y un AIU acorde con los Estándares (AERA et al., 2014) como de los criterios de Kane (2006, 2013, 2020) y las siete inferencias de Chapelle (2021). Para tal fin, se definen las preguntas y objetivos de investigación, que se detallan en la sección siguiente.

Preguntas y objetivos de investigación

Anteponiendo la explicación sobre la definición de los Estándares (AERA et al., 2014), como la propuesta ampliada de Chapelle (2021) respecto a Kane (2006), la pregunta general es:

> ¿En qué medida es válido el uso de los puntajes del ExIES como herramienta en el proceso de selección para la admisión a programas en la educación superior?

Para responder a esta pregunta, a continuación, se definen un conjunto de objetivos y preguntas de investigación encaminados a abordar el proceso de validación, así como las implicaciones y los usos de las puntuaciones del ExIES.

Objetivo general

El objetivo general de la presente investigación es: Evaluar la validez de la interpretación del uso de los puntajes del ExIES (2023-1) como instrumento de selección de aspirantes a programas de licenciatura en la UABC, a través del EBA, según los criterios de claridad, coherencia y plausibilidad de sus inferencias.

Así mismo, en la Tabla 1 se presentan los objetivos específicos y las preguntas de investigación correspondientes, concebidos para desglosar los distintos elementos que influyen en la validez de las puntuaciones y para proveer un análisis integral de este instrumento de evaluación. Todas las inferencias abordadas en esta investigación se enmarcan en el nivel taxonómico de evaluación, de acuerdo con la taxonomía de Anderson y Krathwohl (2001). Esta

elección se sustenta en el hecho de que toda inferencia sobre los resultados de una prueba requiere un juicio fundamentado acerca de la validez de la interpretación de dichos resultados para el uso propuesto.

Tabla 1

Objetivos y preguntas específicas de la investigación

Objetivo específico	Pregunta de investigación
1. Evaluar la congruencia entre el contenido del ExIES y el dominio definido en los planes curriculares y normativos.	¿En qué medida el contenido del ExIES representa de manera adecuada y suficiente el dominio o constructo que se pretende evaluar?
2. Evaluar los procedimientos de aplicación del ExIES y su alineación con estándares técnicos de evaluación.	¿En qué medida las condiciones de administración del ExIES garantizan una evaluación justa y estandarizada?
3. Evaluar la confiabilidad y estabilidad de las puntuaciones del ExIES.	¿En qué medida las puntuaciones del ExIES son consistentes y confiables entre distintas versiones y aplicaciones?
4. Evaluar la estructura factorial y las correlaciones entre áreas del ExIES con base en el modelo teórico subyacente.	¿En qué medida la estructura interna del ExIES se alinea y respalda el modelo teórico de habilidades que pretende medir?
5. Evaluar la evidencia predictiva del ExIES respecto al desempeño universitario.	¿En qué medida el puntaje del ExIES predice el desempeño académico de los estudiantes en el nivel superior?
6. Evaluar el uso de los puntajes del ExIES en los procesos de selección y admisión institucionales.	¿En qué medida los resultados del ExIES son utilizados de forma adecuada y útil en las decisiones institucionales de admisión?
7. Evaluar los efectos de las decisiones basadas en el ExIES desde una perspectiva de equidad y efectividad.	¿En qué medida las decisiones tomadas con base en los puntajes del ExIES generan consecuencias justas y efectivas para los aspirantes y la institución?

Justificación

La presente investigación es pertinente por el impacto que tienen los exámenes de admisión en las trayectorias educativas y profesionales de miles de aspirantes a la educación superior, así como en las políticas de acceso de las instituciones. En el caso de la UABC, como se mencionó, el ExIES se aplicó a 28,205 aspirantes en la convocatoria 2023-1 (Pedroza et al., 2024a), lo que demuestra su alta relevancia social e institucional; no solo con miras al pasado sino también para las aplicaciones posteriores. Tal como señala la UNESCO (2021), los procesos

de selección inciden en la equidad y la movilidad social, por lo que requieren un análisis riguroso.

En congruencia con el *Planteamiento del Problema* sobre la necesidad de garantizar procesos de admisión imparciales, pertinentes y contextualizados en la educación superior, esta investigación asume una concepción de validez que trasciende la confiabilidad psicométrica e incorpora dimensiones teóricas, empíricas y éticas, tal como lo establecen los Estándares (AERA et al., 2014) y lo subrayan Shepard (2016), Kane (2015, 2020) y Chapelle (2021). En este marco, la adopción del EBA, propuesto por Kane (2006, 2013, 2015, 2020), mediante la construcción de un AIU, permite organizar de forma lógica las inferencias entre puntajes, interpretaciones y decisiones, integrando sistemáticamente evidencias tanto cuantitativas como cualitativas (Chapelle, 2021). Desde esta perspectiva, la presente investigación se alinea con los debates actuales en evaluación educativa al promover un modelo de validación que responde a exigencias de equidad, claridad y coherencia (AERA et al., 2014). En particular, esta investigación aporta a la literatura especializada (Lavery et al., 2020) — sobre todo a la latinoamericana donde aún prevalece el enfoque psicométrico clásico— al proponer una metodología que fortalece la legitimidad del uso de los puntajes en la toma de decisiones a partir del uso de EAI.

Desde el plano metodológico, basados en Kane (2006, 2013, 2015), esta investigación propone criterios concretos para valorar la claridad, coherencia y plausibilidad de cada inferencia del Argumento de Validez. Esto representa un aporte relevante a la discusión sobre cómo implementar operativamente el EBA, un aspecto que ha sido identificado como una limitación en la literatura actual (Lavery et al., 2020; Hatala, 2019; Durson & Li, 2021). Para ello, a partir del modelo de Chapelle (2021) —con sus siete inferencias—, se presenta un diseño metodológico

que permite representar de forma visual y argumentativa las inferencias, evidencias, reservas y oportunidades de mejora, fortaleciendo el rigor y la sistematicidad del proceso de validación de cualquier EAI; e incluso como ejemplo para los procesos de validación de otro tipo de pruebas.

Finalmente, desde una perspectiva práctica, los resultados de esta investigación ofrecen un marco de referencia útil para los equipos técnicos responsables del ExIES, orientado a la mejora continua del instrumento, así como generar una base para continuar con la aplicación del EBA. Todo lo anterior, bajo criterios de imparcialidad y transparencia, como lo proponen los Estándares (AERA et al., 2014).

Contexto

Para situar el problema de investigación, esta sección de *Contexto* delimita el marco en el que se inserta el uso de EAI para el ingreso a la educación superior. Primero, se sintetiza la naturaleza y finalidad de los EAI y las tipologías de exámenes de admisión; después, se describe el escenario nacional de demanda y acceso a la ES, con énfasis en México. Enseguida, se caracteriza la UABC en su dimensión institucional y normativa y, finalmente, se describe la adopción del ExIES y su papel en la toma de decisiones de admisión. Este panorama provee las bases necesarias para avanzar hacia el *Marco Teórico*.

Exámenes de Alto Impacto en la educación superior

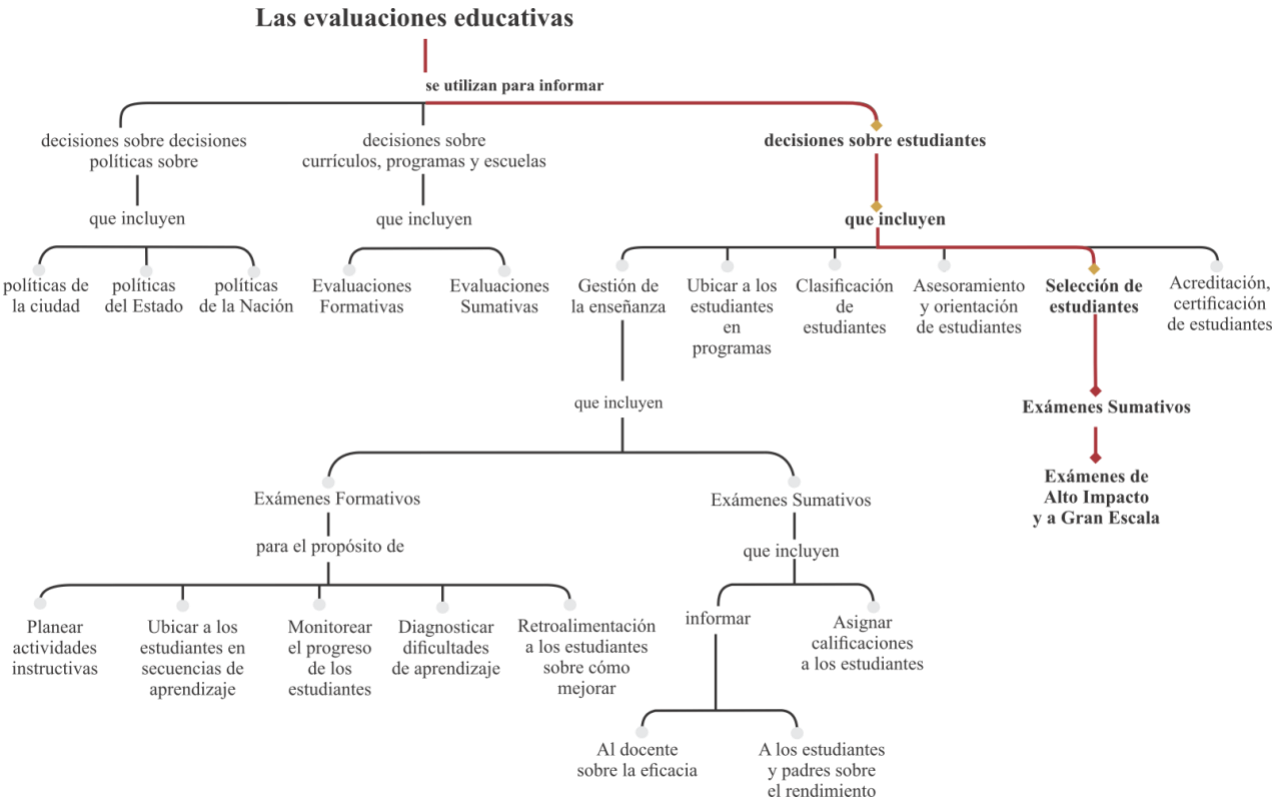
Los EAI se ubican dentro de la evaluación sumativa, sus resultados emiten un juicio final que influye de manera significativa en la trayectoria educativa o profesional de los sustentantes (Instituto Nacional para la Evaluación de la Educación [INEE], 2017; Jones & Ennes, 2018; UNESCO, 2021). A diferencia de las evaluaciones en el aula, enfocadas en la retroalimentación continua (Shepard, 2006), los EAI se aplican a grandes poblaciones en ámbitos regionales o

nacionales y se centran en la medición de habilidades o conocimientos para fines de selección o acreditación (Mattos et al., 2024).

Un ejemplo de EAI son precisamente los exámenes de ingreso a la universidad, cuyo objetivo es determinar el nivel de conocimientos y habilidades de los aspirantes para fines selectivos, reforzando así su carácter sumativo (Brookhart & Nitko, 2019). La Figura 1, adaptada de Brookhart y Nitko (2018), muestra la amplitud de propósitos de la evaluación en la toma de decisiones educativas, resaltando en rojo la ruta específica de los EAI.

Figura 1

Ejemplos de tipos de decisiones educativas donde se pueden usar las evaluaciones



Nota. Adaptado de *Educational Assessment of Students* (p. 8), por S. Brookhart y A. Nitko, 2018, Pearson. © 2019 Pearson. Se añadió el resaltado en rojo para identificar los exámenes.

Para garantizar un uso transparente y equitativo de los EAI, como se ha señalado previamente, resulta fundamental asegurar tanto su validez como su confiabilidad (Popham,

2008; AERA et al., 2014; Mattos et al., 2024). Estos instrumentos tienen implicaciones inmediatas, ya que determinan quién accede —y quién no— a la educación superior, lo cual puede generar efectos positivos, como incentivar la preparación académica; aunado a ello, los EAI tienen la ventaja significativa de identificar potencial académico en estudiantes con desventajas educativas, promoviendo equidad al enfocarse en habilidades cognitivas generales y no únicamente en conocimientos específicos, permitiendo diagnósticos precisos programas de apoyo integral que reducen la deserción y favorecen la integración social (Sackett et al., 2008; Soares, 2012).

Si bien los EAI mantienen aspectos positivos, también presentan desafíos importantes. Cliff y Montero (2010), desde hace más de diez años, especificaron su complejidad técnica, costos elevados, posibles sesgos culturales o lingüísticos, y la necesidad de recursos humanos capacitados, lo que ha apoyado un cambio continuo en la concepción de los exámenes; tanto como el concepto de validez que se revisará en el apartado siguiente. Asimismo, se ha hablado del aumento de la presión sobre los estudiantes y las instituciones; lo cual ha conllevado a distorsiones curriculares, prácticas de enseñanza centradas en la preparación para la prueba, y la proliferación de cursos comerciales sin evidencia sólida de efectividad, lo que refuerza la necesidad de una evaluación educativa más profesional, justa y contextualizada (Sánchez-Mendiola & Delgado-Maldonado, 2017; Furuta et al., 2021; Bennett, 2022). Para enfrentar estos retos, es esencial implementar enfoques científicos sólidos que combinen el desarrollo riguroso de instrumentos psicométricos con estrategias complementarias de apoyo académico, curricular, psicológico y financiero que garanticen un acompañamiento integral del estudiante durante toda su trayectoria educativa (Cliff & Montero, 2010).

Tipos de exámenes de ingreso universitario: nacionales e internacionales

Este tipo de exámenes varían según el país o el sistema educativo, pero suelen agruparse en categorías que abarcan desde pruebas de amplio alcance hasta exámenes específicos. En la Tabla 2 se presenta un panorama de estos tipos de exámenes de ingreso universitario, tomando en cuenta su enfoque temático y el énfasis en habilidades generales o específicas.

Tabla 2

Tipos de exámenes de ingreso a la universidad

Tipo de examen	Descripción	Ejemplos
Exámenes de amplio alcance	Diseñadas con estructura específica para evaluar habilidades y conocimientos generales, aplicadas a gran escala.	SAT, ACT, EXANI II y III, ENEM, EXHCOBA, EXCOBA, GenerEx
Exámenes de aptitud	Evalúan habilidades y aptitudes específicas relevantes para un área de estudio o carrera.	Examen de aptitud para ingeniería
Exámenes de conocimientos específicos	Enfocados en evaluar el conocimiento en un campo académico particular.	Examen de medicina, Examen de derecho
Pruebas de idiomas	Evaluaciones del dominio del idioma, especialmente para programas en otros idiomas.	TOEFL, IELTS
Exámenes prácticos o de habilidades específicas	Requieren demostrar habilidades necesarias para un área de estudio específica.	Presentaciones de arte, Pruebas de laboratorio
Entrevistas y ensayos	Evaluación del candidato mediante entrevistas personales o ensayos para demostrar interés y aptitudes adicionales.	Entrevista personal, Ensayos

Nota. EXANI = Examen Nacional de Ingreso; ENEM = *Exame Nacional do Ensino Médio*; EXCOHBA = Examen de Habilidades y Conocimientos Básicos; EXCOBA = Examen de Conocimientos Básicos; GenerEx = Generación de Exámenes. Los ejemplos incluidos tienen fines exclusivamente ilustrativos.

En el caso de México, los exámenes de ingreso a licenciatura suelen ser de opción múltiple y se emplean tanto en instituciones públicas como privadas. Entre los más usados están el Examen de Habilidades y Conocimientos Básicos (EXHCOBA), aplicado desde 1992 (Backhoff & Tirado, 1992), y administrado por Métrica Educativa desde el año 2013 (2023), y sus versiones Excoba y GenerEx. Además, el Centro Nacional de Evaluación para la Educación Superior (CENEVAL), fundado en 1994 para satisfacer la demanda de un instrumento confiable y estandarizado, creó los exámenes EXANI (Vidal, 2009). El EXANI-II (CENEVAL, 2023), es

uno de los más empleados en el país para el ingreso a estudios de nivel superior e incluye la evaluación de habilidades básicas, conocimientos disciplinares y un nivel de inglés aproximado al B1 del Marco Común Europeo de Referencia (MCER). El ExIES, a partir del 2023, forma parte de los exámenes de ingreso a la educación superior.

A nivel internacional, cada región emplea exámenes con metodologías y enfoques diversos, de acuerdo con las exigencias y prioridades de su sistema educativo. Entre los más conocidos se encuentran el SAT y el ACT en Estados Unidos (College Board, 2023b; ACT, 2023, 2024); el Bachillerato Internacional (International Baccalaureate, IB), fundado en 1968, y el Baccalauréat en Europa (International Baccalaureate Organization, 2021), desde 1968. Así como pruebas nacionales en América Latina, por ejemplo: el Exame Nacional do Ensino Médio (ENEM) en Brasil desde 1998 y reformado en 2009 (INEP, 2023), la Prueba de Selección Universitaria (PSU) en Chile desde 2003-2020; hoy PAES desde 2022 (DEMRE, 2020), el Ciclo Básico Común (CBC) en Argentina (implementado en 1985) y la prueba Saber 11 en Colombia (ICFES, 2021) desde 1968. Estas evaluaciones no solo miden la aptitud académica de manera objetiva, sino que también verifican la posesión de competencias básicas necesarias para afrontar con éxito la educación superior en cada contexto.

Educación superior en México: demanda, cobertura y acceso

En el panorama de la educación superior en México, la demanda de espacios en universidades públicas resuelta relevante y en constante crecimiento. Durante el ciclo escolar 2023-2024, según el Anuario Estadístico de la Población Escolar en Educación Superior (ANUIES, 2024), la matrícula de las universidades públicas federales fue de 548,987 estudiantes (véase Tabla 3).

Tabla 3

Matrícula total y porcentaje sobre el total: universidades públicas federales

No.	Nombre de la institución	Matrícula total	% sobre el total
1	Universidad Nacional Autónoma de México	229,141	41,7%
2	Instituto Politécnico Nacional	132,288	24,1%
3	Universidad Abierta y A Distancia De México	114,714	20,9%
4	Universidad Autónoma Metropolitana	54,266	9,9%
5	Universidad Pedagógica Nacional	7,777	1,4%
6	Universidad Autónoma Agraria Antonio Narro	5,575	1,0%
7	Universidad Autónoma Chapingo	5,226	1,0%
	Total	**548,987**	**100,0%**

Nota. Datos extraídos de *Anuario Educación Superior – Técnico Superior, Licenciatura y Posgrado 2023–2024* (versión 1.2), por ANUIES, 2024, en la sección "Anuarios estadísticos de educación superior", “Universidades públicas federales”.

Por otro lado, la matrícula total en instituciones estatales y públicas alcanzó 1,263,096 estudiantes (ANUIES, 2024). Dentro de este contexto, la Universidad de Guadalajara (UdeG) registra la matrícula más alta, seguida por la Universidad Autónoma de Nuevo León (UANL) y la Benemérita Universidad Autónoma de Puebla (BUAP). En sexto lugar se ubica la UABC, con 66,885 estudiantes, equivalente al 5.30% del total de la matrícula nacional de universidades públicas (véase Tabla 4).

Tabla 4

Las 10 universidades públicas con más matrícula en México (2023-2024)

N.º	Entidad federativa	Nombre de la institución	Matrícula total	% sobre el total
1	Jalisco	Universidad de Guadalajara	136,980	10.84%
2	Nuevo León	Universidad Autónoma de Nuevo León	128,716	10.19%
3	Puebla	Benemérita Universidad Autónoma de Puebla	91,722	7.26%
4	Sinaloa	Universidad Autónoma de Sinaloa	74,210	5.88%
5	Estado de México	Universidad Autónoma del Estado de México	68,343	5.41%
6	Baja California	Universidad Autónoma de Baja California	66,885	5.30%
7	Veracruz	Universidad Veracruzana	65,096	5.15%
8	Michoacán	Universidad Michoacana de San Nicolás de Hidalgo	39,697	3.14%
9	Chihuahua	Universidad Autónoma de Ciudad Juárez	36,765	2.91%
10	Sonora	Universidad de Sonora	35,694	2.83%

Nota. Datos extraídos de *Anuario Educación Superior – Técnico Superior, Licenciatura y Posgrado 2023–2024* (versión 1.2), por ANUIES, 2024, en la sección "Anuarios estadísticos de educación superior", “Universidades públicas estatales”.

Como se aprecia en la Tabla 5, cada año un gran número de aspirantes busca acceder a ES en instituciones públicas o privadas. Para el ciclo 2023-2024, se registraron 2,106,774 solicitudes de nuevo ingreso en Licenciatura Universitaria y Tecnológica, de las cuales se admitió al 54.20% (1,141,964). Las instituciones públicas concentraron 1,456,408 solicitudes y aceptaron a 625,847 estudiantes, alcanzando una tasa de 42.97%, en tanto que las privadas recibieron 650,366 solicitudes y admitieron a 516,117, lo que supone un 79.36%. Este contraste pone de manifiesto la disparidad en la oferta de espacios y la competencia por ingresar a la ES, asociada al tipo de sostenimiento institucional.

Tabla 5

Solicitudes de nuevo ingreso a licenciatura universitarias y tecnológicas

Tipo de sostenimiento	Solicitudes de nuevo ingreso	Nuevo ingreso	% Tasa de aceptación
Público	1,456,408	625,847	42.97%
Privado	650,366	516,117	79.36%
Total	2,106,774	1,141,964	54.20%

Nota. Datos extraídos de *Anuario Educación Superior – Técnico Superior, Licenciatura y Posgrado 2023–2024* (versión 1.2), por ANUIES, 2024, en la sección "Anuarios estadísticos de educación superior".

La UABC y el ExIES: Dimensión institucional y proceso de admisión

La UABC ha atendido anualmente un promedio de 18,989 estudiantes de nuevo ingreso en licenciatura, del ciclo 2017-2 al 2023-2, lo que pone en evidencia su relevancia en el noroeste de México y la extensa gama de áreas de formación que ofrece a la comunidad estudiantil. Durante el ciclo 2023-1, como se mencionó anteriormente, la UABC reportó un total de 66,715 alumnos, de los cuales 65,057 cursaban programas de licenciatura y 1,658 pertenecían a posgrado (CGSEGE, 2023). Para satisfacer esta demanda, la UABC dispone de 146 programas de licenciatura y 76 de posgrado (véase la Tabla 6), distribuidos en sus tres campus; Mexicali (con sus extensiones en Cd. Morelos y San Felipe), Tijuana (con la Unidad Tecate, la Unidad

Valle de las Palmas y el Centro de Estudios Rosarito) y Ensenada (El Sauzal, Valle Dorado y la Unidad San Quintín).

Tabla 6

Datos generales de la matrícula de la UABC

Número de matrícula escolar por nivel educativo (2023-1)	66,715
a. Licenciatura	65,057
b. Posgrado	1,658
Numero de programas educativos de licenciatura escolarizados	146
Número de programas educativos de posgrado	76
a. Investigación	32
b. Profesionalizante	44

Nota. Elaboración propia basada en *Reporte de matrícula y programas de la UABC, ciclo 2023-1*, por UABC-CGSEGE, 2023.

Por otro lado, la Tabla 7 presenta el histórico de nuevo ingreso a nivel licenciatura en la UABC para diversos ciclos académicos, con un desglose por campus y unidades, así como un promedio general en la última fila. En ciertos periodos, la institución ha registrado valores relativamente bajos (por ejemplo, 7,675 estudiantes en 2023-1), mientras que en otros ha superado los 19,000 (caso de 2023-2024). Estas fluctuaciones se asocian con factores como la capacidad de cada programa, el número de egresados de bachillerato y las políticas de ampliación o estabilización de la matrícula. De manera global, el promedio indica 18,989 nuevos ingresos, con diferencias notables entre campus: el de Mexicali se sitúa en torno a 6,957, Tijuana en 8,666 (considerando Unidad Valle de las Palmas, Unidad Tecate y Unidad Rosarito) y Ensenada en 3,367 (considerando la Unidad San Quintín), en tanto las unidades y extensiones complementan la cobertura en distintos puntos de la entidad.

Tabla 7

Nuevos ingresos a nivel licenciatura de la UABC (2017-2024)

Nuevo Ingreso	**Campus Mexicali**	**Campus Tijuana**	**Unidad Valle de las Palmas**	**Unidad Tecate**	**Unidad Rosarito**	**Campus Ensenada**	**Unidad San Quintín**	**Totales**

2017-2	3,548	3,124	842	282	155	1,694	125	9,770
2018-1	2,995	2,963	568	105	102	1,464	79	8,276
2017-2018	6,543	6,087	1,410	387	257	3,158	204	18,046
2018-2	3,694	3,175	839	280	177	1,786	131	10,082
2019-1	2,993	3,015	568	119	98	1,391	90	8,274
2018-2019	6,687	6,190	1,407	399	275	3,177	221	18,356
2019-2	3,756	3,218	975	333	192	1,919	144	10,537
2020-1	3,365	3,050	610	177	148	1,556	89	8,995
2019-2020	7,121	6,268	1,585	510	340	3,475	233	19,532
2020-2	3,748	3,185	919	335	175	1,844	148	10,354
2021-1	3,153	2,968	568	145	139	1,214	89	8,276
2020-2021	6,901	6,153	1,487	480	314	3,058	237	18,630
2021-2	4,066	3,588	1,175	283	199	1,920	208	11,439
2022-1	3,049	2,915	686	103	91	1,087	165	8,096
2021-2022	7,115	6,503	1,861	386	290	3,007	373	19,535
2022-2	4,275	3,535	1,145	285	139	1,803	192	11,374
2023-1	2,740	2,961	624	71	99	1,016	164	7,675
2022-2023	7,015	6,496	1,769	356	238	2,819	356	19,049
2023-2	4,131	3,500	1,164	279	141	1,928	199	11,342
2024-1	3,184	3,187	747	103	88	952	170	8,431
2023-2024	7,315	6,687	1,911	382	229	2,880	369	19,773
Promedio	6,957	6,341	1,633	414	278	3,082	285	18,989
% total	36.64%	33.39%	8.60%	2.18%	1.46%	16.23%	1.50%	100%

Nota. Datos retomados de *Estadísticas de la Población Estudiantil la UABC*, de 2017-2 a 2024-1, por UABC-CGSEGE, 2025. Los resultados de las filas en gris solo corresponden al ciclo 2023-2024.

La amplia variedad de programas académicos que oferta la UABC responde al interés de los aspirantes por disciplinas que abarcan ciencias exactas y naturales, administración y negocios, ciencias sociales, humanidades, ingeniería, salud, pedagogía, entre otras. Esto se traduce en una competencia significativa para obtener un lugar, pues existe una heterogeneidad de perfiles que buscan formarse en ámbitos específicos.

Aspectos normativos. Como se explicó en el planteamiento del problema, la UABC, según la Ley Orgánica de la UABC (2010), el Estatuto General (UABC, 2019), y el Estatuto Escolar (UABC, 2021), tiene la obligación de realizar por lo menos una convocatoria anual para el proceso de selección de nuevo ingreso a nivel licenciatura.

Con el objetivo de asignar los espacios de manera clara y ordenada, la universidad emite anualmente dos convocatorias para el Concurso de Selección de Nuevo Ingreso en licenciatura. La primera —conocida como convocatoria grande— se anuncia usualmente en marzo y engloba la mayoría de las áreas o troncos comunes, la cual se publica en la página oficial (UABC-CGSEGE, 2025). La segunda, que se presenta entre agosto y septiembre, se restringe a los programas y vacantes remanentes luego de la primera convocatoria (UABC, 2021, artículo 17). De acuerdo con la Tabla 8, durante el ciclo 2023-1 se registraron 28,205 aspirantes y, para 2023-2, 2,435 más, sumando un total de 30,640. De este conjunto, 19,773 obtuvieron un lugar (11,342 en 2023-2 y 8,431 en 2024-1), lo que se tradujo en una tasa global de aceptación del 64.53 %. Estos datos reflejan tanto la magnitud de la demanda de ingreso a la UABC como el grado de competencia entre los postulantes.

Tabla 8

Aspirantes, nuevo ingreso y tasa de aceptación por ciclo (2023-2024)

Ciclo de Aplicación	**Número total de aspirantes**
2023-1	28,205
2023-2	2,435
Total	30,640
Inicio de ciclo de Semestre	**Nuevo ingreso**
2023-2	11,342
2024-1	8,431
Total	19,773
Tasa de aceptación (%)	64.53%

Nota. Elaboración propia basada en los *Reportes Técnicos del ExIES 2023-1 y 2023-2* (Pedroza Zúñiga et al., 2024a, 2024b) y de la página oficial de estadísticas de la UABC (UABC-CGSEGE, 2025).

Dado este contexto, el ExIES juega un papel determinante como instrumento de medición para orientar las decisiones de admisión. Su aplicación oficial en el ciclo 2023-1, con 28,205 aspirantes (Pedroza Zúñiga et al., 2024a), representa un hito significativo en la consolidación de

una estrategia de evaluación institucional. No obstante, para que el ExIES cumpla con las expectativas de equidad y eficacia en la selección, es imprescindible contar con un marco de validez bien fundamentado que avale las interpretaciones de sus puntajes (Kane, 2015; Chapelle, 2021). De ahí que se reafirme la necesidad de realizar un estudio exhaustivo que examine tanto los aspectos psicométricos como las implicaciones educativas y sociales de su uso, a fin de robustecer la imparcialidad y la transparencia en el proceso de admisión universitaria.

Marco teórico

A partir del *Planteamiento del Problema*, donde se destacó la función de los EAI y la importancia de la validez en la toma de decisiones académicas, en esta sección se profundiza en los fundamentos teóricos de la evaluación educativa y su transformación a lo largo del tiempo. En primer lugar, se describe la naturaleza de la evaluación, evaluación educativa y sumativa con el fin de situar el objeto de estudio. Posteriormente, se revisa la evolución del concepto de validez, desde sus orígenes ligados a la medición y correlaciones estadísticas, pasando por la fragmentación en diversos tipos de validez, hasta llegar a la perspectiva unificada; sin dejar de lado el proceso de validación y la relevancia de la confiabilidad e imparcialidad en las pruebas, según la AERA, la APA y el NCME (2014). En este recorrido se introduce el EBA, propuesto por Kane (2006, 2013), el cual propone trascender el mero análisis estadístico para contemplar la interpretación y las consecuencias de las decisiones derivadas de los resultados de la evaluación. Así, se busca ilustrar cómo la reflexión sobre la validez evoluciona desde un abordaje centrado en la medición hacia otro que incluye la interpretación, el uso y las implicaciones éticas de dichas decisiones y lograr situar estas reflexiones en la *Discusión y Conclusiones* de la presente investigación.

De la evaluación educativa a la evaluación sumativa

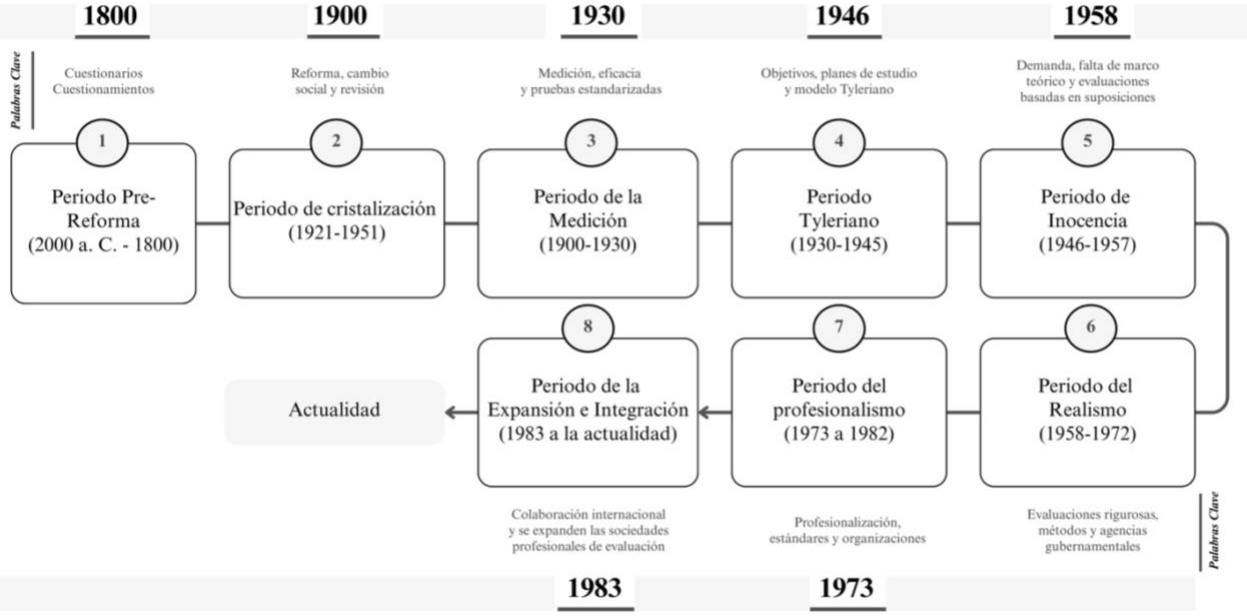
La evaluación se concibe como un proceso sistemático de recolección y valoración de información para determinar el mérito, valor o eficacia de algo (Scriven, 2008). En el ámbito educativo, constituye una herramienta clave para juzgar la calidad de programas, procesos e intervenciones, así como para la toma de decisiones (Stufflebeam & Zhang, 2017). Cuando estos principios se aplican al entorno escolar o formativo, hablamos de evaluación educativa, la cual se enfoca en valorar el grado de logro de objetivos de aprendizaje, la eficacia de los métodos de

enseñanza y la calidad de los recursos didácticos, además de propiciar reflexiones sobre la equidad y la mejora continua en la práctica docente (Stufflebeam & Coryn, 2014).

Esta concepción, en la actualidad, obedece a la expansión de los sistemas educativos, la demanda de evidencia empírica para la toma de decisiones y la reflexión crítica sobre los alcances y limitaciones de la medición. Así, como se observa en la Figura 2, de la interrogación oral y la inspección rudimentaria se pasó a integrar enfoques más amplios que, además de cuantificar resultados, consideran factores contextuales, la equidad y la participación de diversos actores (Chiva et al., 2009; Stufflebeam & Shinkfield, 1985). De esta manera, la evaluación se fue formalizando hasta convertirse en un proceso riguroso y multifacético que sirve no solo para medir, sino también para otorgar valor y propiciar mejoras en la práctica educativa.

Figura 2

Línea del tiempo del concepto de evaluación educativa



Nota. Elaboración propia basada en "Program Evaluation: A Historical Overview" (G. F. Madaus & D. L. Stufflebeam, 2000, en Evaluation in Education and Human Services, vol. 49, pp. 3-22), Evaluation Theory, Models, and Applications (2.ª ed.; D. L. Stufflebeam & A. J. Shinkfield, 2007) y "Historia de la evaluación educativa" (I. Chiva, M. J. Perales & A. Pérez Carbonell, 2009, en Conceptos, metodología y profesionalización en la evaluación educativa, pp. 43-70).

En ese tránsito histórico, la evaluación sumativa cobra especial relevancia al proporcionar un juicio global sobre la valía de un programa o proceso al concluirse, certificando los logros y las competencias alcanzadas y sirviendo a fines de rendición de cuentas institucionales mediante la agregación de múltiples eventos evaluativos a lo largo del ciclo formativo (Scriven, 1967, 1991, 2000). Esta perspectiva subraya su utilidad tanto en el aseguramiento de la calidad como en la certificación de conocimientos y competencias (Bennett, 2015). Por su naturaleza conclusiva, la evaluación sumativa exige criterios rigurosos de validez y confiabilidad (AERA et al., 2014), al tiempo que responde a finalidades diversas, como el acceso a instituciones educativas, la promoción de grado o la obtención de un título (Sambell et al., 2012). Un claro ejemplo son los exámenes de admisión a la educación superior (Hambleton & Zenisky, 2011); como el ExIES, que emiten un puntaje final y determinan el potencial académico de cada aspirante.

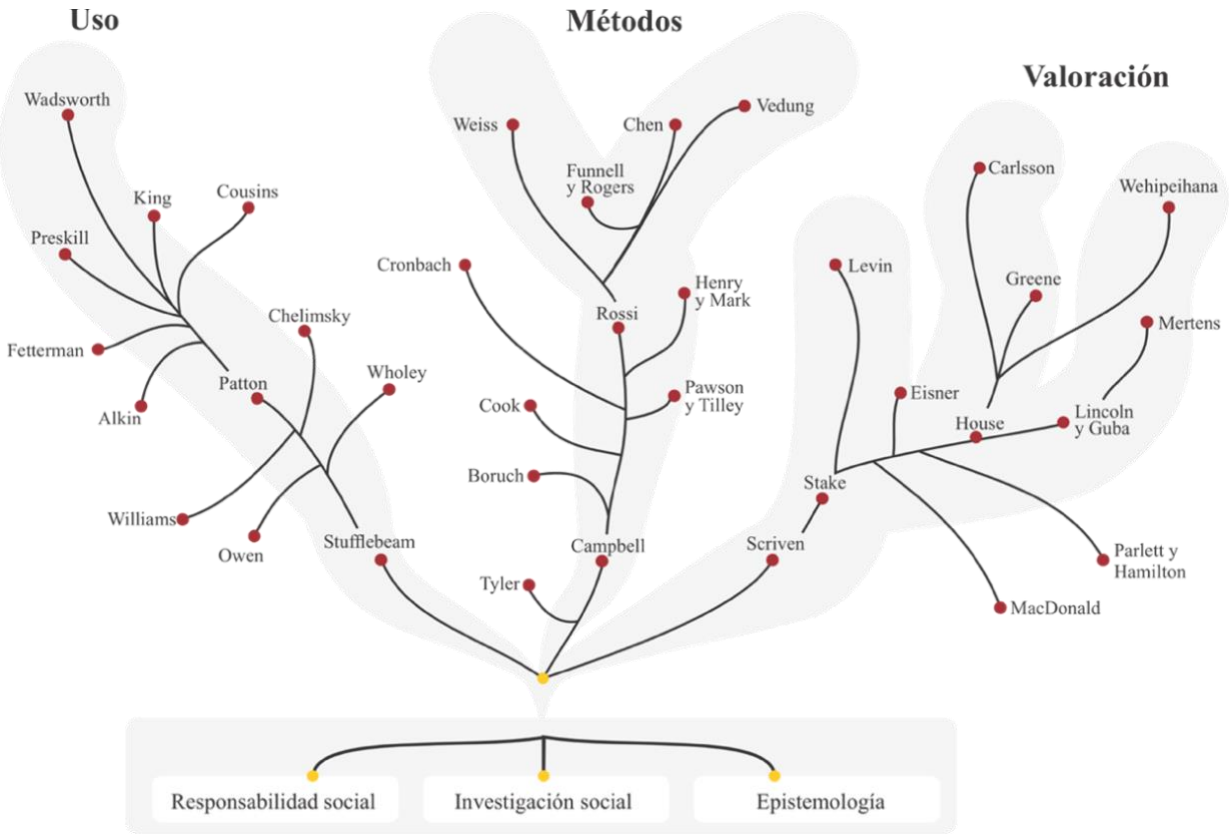
Ahora bien, la complejidad de la evaluación —incluida la sumativa— puede entenderse a partir del Árbol de la Teoría de la Evaluación (Figura 3), propuesto por Alkin (2012). Este modelo agrupa tres grandes enfoques que coexisten en la disciplina: el rigor metodológico, la valoración o juicio de valor y el uso de hallazgos. En particular, la evaluación sumativa se vincula tanto con la rama enfocada en la medición y el control de variables como con aquella que enfatiza la utilidad de los resultados, ya que debe emitir un juicio global y fundamentar decisiones de alto impacto formativo (Scriven, 2008; Stufflebeam & Coryn, 2014).

En ese sentido, la selección y el diseño de técnicas e instrumentos adquieren especial relevancia para garantizar que los resultados reflejen con exactitud las competencias evaluadas, sobre todo, cuando el proceso conduce a determinaciones que afectan la trayectoria académica y profesional del estudiantado. Con base en este árbol teórico, la evaluación sumativa del ExIES se

coloca en la confluencia de los tres enfoques, pues requiere métodos rigurosos (metodología), implica un juicio de valor sobre el desempeño de los aspirantes (valoración) y se traduce en acciones concretas que afectan las trayectorias formativas de los estudiantes (uso).

Figura 3

Árbol de Teoría de la Evaluación



Nota. Adaptado de *Evaluation Roots: A Wider Perspective of Theorists' Views and Influences* (p. 105), de M. C. Alkin, 2012, SAGE Publications.

Desarrollo del concepto de validez

En este contexto, la validez emerge como una condición indispensable para fundamentar adecuadamente las decisiones derivadas de la evaluación educativa, trascendiendo la simple acumulación de datos estadísticos hacia una comprensión profunda y contextual de los resultados obtenidos (Newton & Shaw, 2014; Markus & Borsboom, 2013). De este modo, la validez no se limita únicamente a la precisión en la medición, sino que se expande hacia la interpretación del

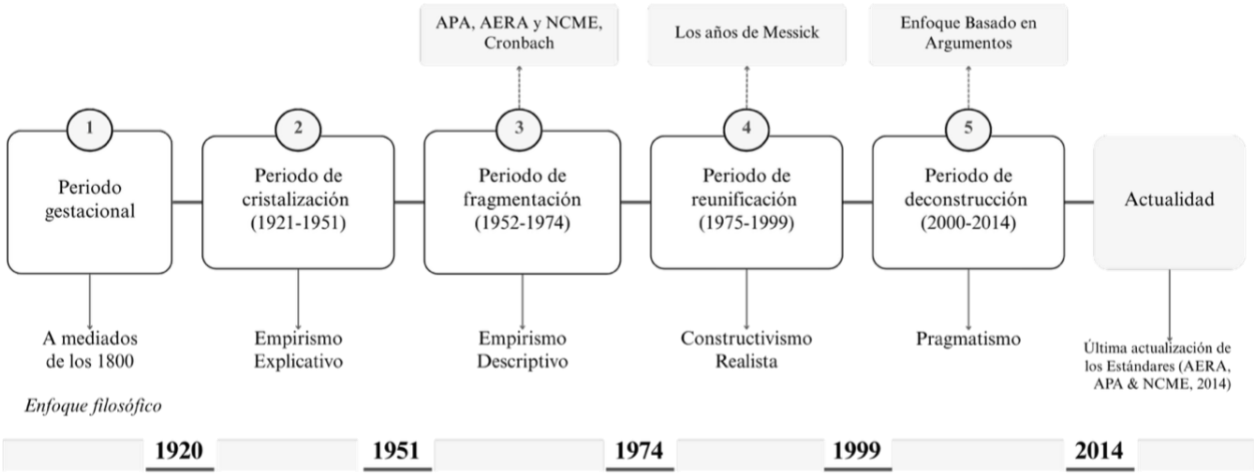
significado y las consecuencias prácticas y éticas de las decisiones basadas en las evaluaciones (Kane, 2015, 2020). En consecuencia, el análisis conceptual de la validez se convierte en un aspecto fundamental para asegurar la legitimidad y pertinencia social de los instrumentos evaluativos, especialmente cuando estos influyen directamente en las trayectorias educativas y profesionales (AERA et al., 2014; García et al., 2017). Por lo anterior, se retoma la visión de la historia conceptual pues, de acuerdo con Koselleck (2000), el lenguaje y los conceptos se transforman junto con la historia, reflejando los cambios sociales que se llevan a la práctica.

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 4 presenta 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 4

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 & D. Borsboom, 2013) con la periodización propuesta en *Validity in Educational & Psychological Assessment* (P. Newton & S. Shaw, 2014), como la reafirmación de *The evolution of concept of validity* (M. Kane & 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 & Borsboom, 2013; Newton & 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 & 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 & 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, mismas que se presentan en la Tabla 9. 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 9

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)

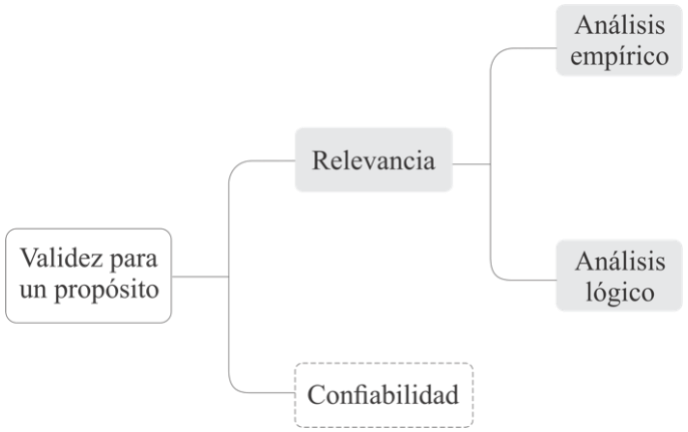
Nota. Elaboración y traducción propia basado en "Validity" (S. Messick, 1989, en R. L. Linn [Ed.], *Educational Measurement*, 3.ª ed., pp. 13-103) y en *The Concept of Validity: Revisions, New Directions, and Applications* (R. Lissitz, 2009).

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 5 ilustra gráficamente el modelo propuesto por Cureton (1951), destacando los componentes clave que configuraron significativamente la evolución conceptual de la validez.

Figura 5

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 & 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 10 muestra la caracterización de estos enfoques según surgían en este periodo y cómo se modificó su pregunta central.

Tabla 10

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,

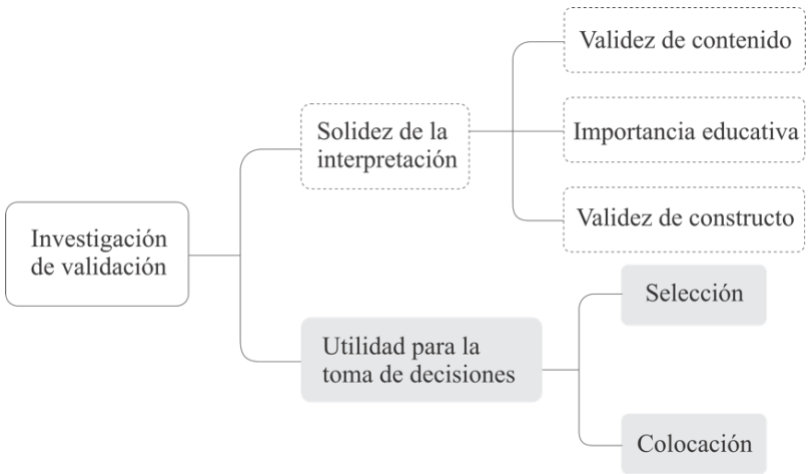
1955) y *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning* (K. A. Markus & D. Borsboom, 2013).

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 6 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 & 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 6

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 & 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

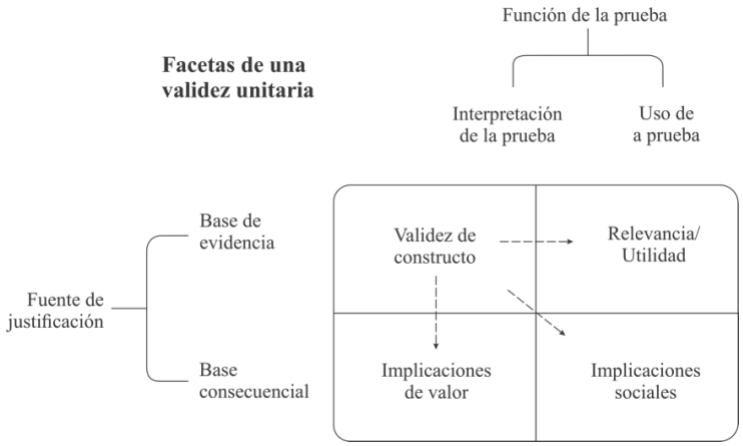
los grupos involucrados, especialmente los más vulnerables o menos favorecidos. Así, la validez no solo implica criterios técnicos y empíricos, sino también consideraciones éticas relacionadas con la equidad y el impacto social.

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 7 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 7

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 & 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.

Con el fin de resumir estas etapas históricas, en la Tabla 11, se expresa cómo el concepto de validez las ha atravesado —desde los periodos de gestación y cristalización, hasta la fragmentación, la re-unificación y la etapa de deconstrucción— para, finalmente, incorporar dimensiones sociales y éticas de la actualidad.

Tabla 11

Evolución del concepto de validez según el periodo, definición, alcance y autores

Periodo	Definición de Validez	Alcance	Autores representativos
Periodo Gestacional (a mediados de 1800-1920)	Reconocimiento de la importancia de evaluar la capacidad de las pruebas para medir lo que se supone que deben medir, sin una definición clara y unificada.	Sólo medir	Diversidad de perspectivas
Periodo de Cristalización (1920-1951)	El grado en que una prueba mide lo que se supone que debe medir", enfatizando el contenido de la prueba y su relación con el constructo evaluado.	Sólo medir	Louis Thurstone, Lee Cronbach, y Paul Meehl
Periodo de Fragmentación (1952-1974)	Surgimiento de perspectivas diversas y falta de consenso sobre la definición de validez.	Medir y decidir	Lee Cronbach, y la AERA, APA y NCME
Periodo de Re-unificación (1975-1999)	Propuesta de una definición ampliada de la validez que incluye el contenido de la prueba y las consecuencias de su uso.	Medir, tomar decisiones y analizar sus consecuencias	Samuel Messick
Periodo de Deconstrucción (2000-2014)	Cuestionamiento del concepto unitario de validez y énfasis en considerar múltiples facetas de la validez.	Medir, tomar decisiones y analizar sus consecuencias desde lo social y cultural	Michel Kane, Linn, Robert L. Brennan, David Carless, Mary James, y Paul Newton, entre otros.
Actualidad (2014-actualidad)	La validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas. (AERA et al., 2014, p.11)	Medir, tomar decisiones y analizar sus consecuencias desde lo social y cultural	AERA, APA, y NCME
	La validez es una propiedad de la interpretación y uso de los resultados. Además, se discute a la validez como parte de las *testing policies*, es decir, forma parte de un conjunto de reglas y directrices, como estándares de calidad, confiabilidad, imparcialidad, transparencia en los resultados, la gestión y retroalimentación, por decir algunos.	Medir, tomar decisiones y analizar sus consecuencias	Samuel Messick
		Medir, tomar decisiones y analizar sus consecuencias desde lo social y cultural, rendición de cuentas, entre otros.	Kane; Chapelle; Sireci; Embretson; y otros.

Nota. Elaboración propia basada en *Standards for Educational and Psychological Testing* (AERA et al., 2014) y en la exposición de las teorías de validez de S. Messick y M. Kane presentada en *Validity in Educational & Psychological Assessment* (P. Newton & S. Shaw, 2014).

Proceso de validación

Como se revisó en la sección anterior, la validez ha enfrentado el problema de operacionalización, es decir, en 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 12). 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 & 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 12

Evolución histórica de la concepción del proceso de validación en los Estándares

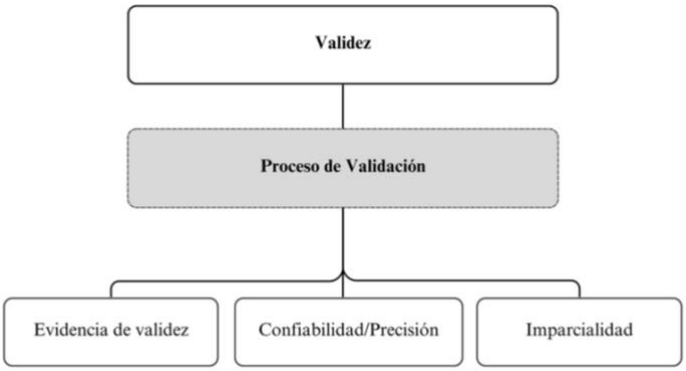
Año	Enfoque dominante	Síntesis del proceso de validación
1954/1955	Recomendaciones técnicas iniciales	Documentación técnica básica sobre las pruebas (AERA & NCME, 1955).
1966/1974	Jerarquía de estándares	Énfasis en desarrollo, uso y reporte jerarquizado de resultados (AERA et al., 1966, 1974).
1985	Linealidad 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 8, 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 8

Integración del Proceso de Validación dentro del concepto de Validez



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

Si bien las evidencias de validez son centrales, éstas 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 sub tipificarlos 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 & Marcoulides, 2011).

Diversos métodos son utilizados para evaluar la confiabilidad, como se presenta en la Tabla 13. 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 13

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	Cronbach et al. (1972); Cronbach et al. (2004)	Permite un análisis más completo y detallado de la varianza de medición.	Puede requerir diseños y análisis estadísticos más complejos.

	través de diseños factoriales.		Identifica fuentes específicas de error.	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 & Novick (1968); Embretson & 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 & 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 & 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 & 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.

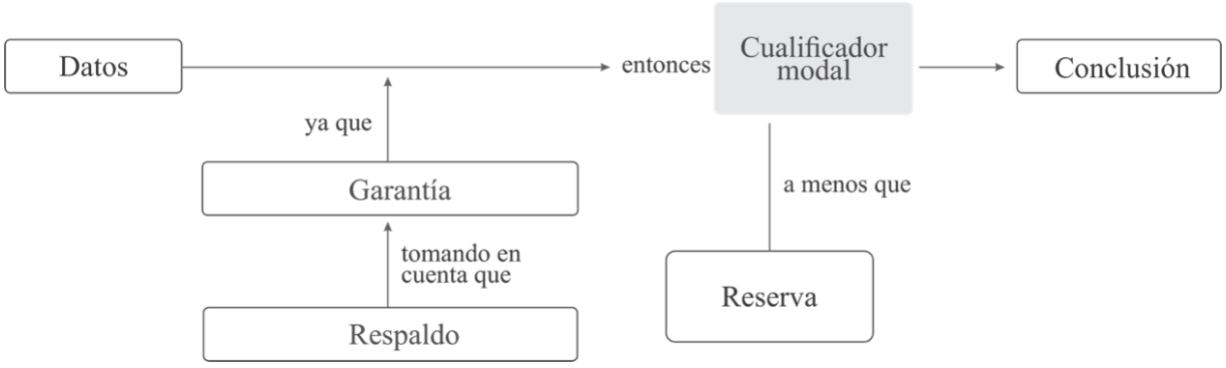
Los eslabones lógicos o inferencias deben ser consistentes y plausibles para sostener que una prueba mide lo que se propone (Kane, 2011, 2013), que a la vez considera las evidencias como refutaciones que pudieran surgir.

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 9

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. En la Tabla 14 se presenta un resumen de su contenido y de los procedimientos que suelen utilizarse.

Tabla 14

Inferencias propuestas por Kane

Inferencia	Refiere a	Ejemplo de garantía
Puntuación (*scoring*)	Se toman información de los datos de los puntajes como afirmación, se establecen criterios y reglas.	Reglas o rúbricas de puntuación. Estandarización de puntajes. Juicio de grupos de expertos.
Generalización (*generalization*)	La puntuación observada del evaluado nos da una estimación de la puntuación del universo evaluado.	Teoría de la generalizabilidad (teoría G). Tamaño de la muestra y cantidad de preguntas.
Extrapolación (*extrapolation*)	Estimación de la función de la prueba en el contexto real.	Análisis entre la relación de la prueba y su función en otros contextos. Ecuación de regresión.

Implicaciones (*Implications*)	Interpretación y toma de decisiones	Aprobación o no aprobación del estándar. Acciones. Consecuencias voluntarias o involuntarias.

Nota. Elaboración propia basada en "Content-Related Validity Evidence in Test Development" (M. Kane, 2006, en S. M. Downing & T. M. Haladyna [Eds.], *Handbook of Test Development*, pp. 131-153) y "A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework" (D. A. Cook, R. Brydges, S. Ginsburg & R. Hatala, 2015).

Estas cuatro inferencias establecen 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 10 y Tabla 15). 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 15

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.
		Garantías[a]	Enunciados que indican que una inferencia puede autorizarse en un contexto determinado.
Particular		Supuestos	Enunciados que aclaran qué evidencia es necesaria.
	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. [a]Las 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 10, 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 10

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 16 se ofrecen ejemplos concretos de investigaciones que ilustran cómo diferentes técnicas pueden aportar evidencia sólida para respaldar cada paso del argumento.

Tabla 16

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

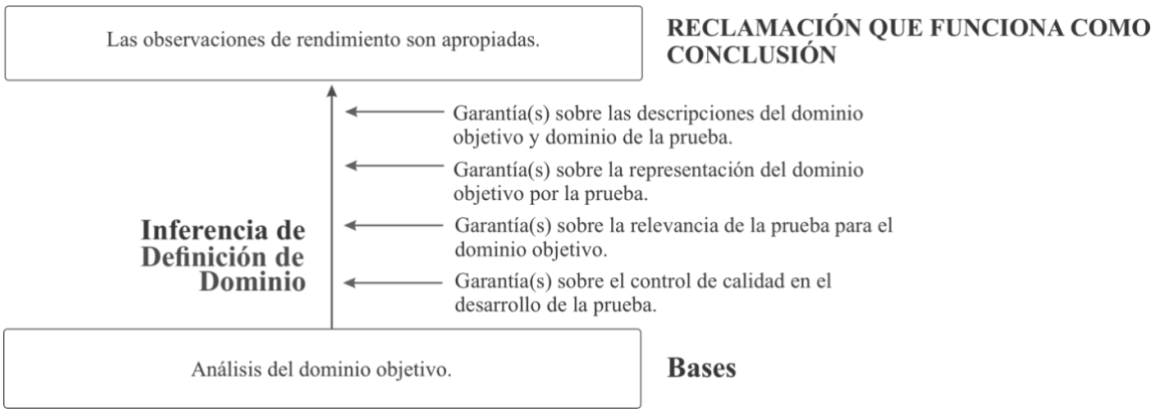
verifica la correspondencia entre la prueba y dichas habilidades (Evaluación) y se comprueba que las puntuaciones se mantienen estables en distintas circunstancias (Generalización). Con la inferencia de Explicación, se dilucida si los puntajes reflejan realmente el constructo teórico propuesto y, mediante 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).

A continuación, se presentan apartados separados de las siete inferencias que Chapelle (2021) propone, dado que permite una comprensión escalonada del EBA y muestra de forma didáctica cómo cada supuesto debe ser defendido con evidencias pertinentes. Así, sin importar si se adopta un modelo con cuatro o siete etapas, la idea central sigue siendo la misma: plantear un conjunto de inferencias claras, sustentarlas con datos y argumentaciones, y fortalecer de manera progresiva la validez de las interpretaciones y usos de los resultados de la prueba. En cada inferencia habrá que retomar la Tabla 15 ya que, a través de figuras, se expresa cómo las garantías precisan el tipo de evidencia que se necesita para validar la información propuesta.

Inferencia de Definición de Dominio. Esta inferencia, ilustrada en la Figura 11, subraya la necesidad de un proceso de ECD que garantice la correspondencia entre el contenido de los ítems y el dominio objetivo (Sireci, 1998, 2008). También, la inferencia de Definición de Dominio, demanda controlar la calidad de los ítems mediante revisiones de expertos, análisis estadísticos y pruebas piloto que respalden la representatividad y la pertinencia de cada tarea o pregunta. En suma, el propósito es comprobar que las observaciones del rendimiento realmente reflejen el dominio definido, sin limitarse a los puntajes obtenidos en la prueba (Chapelle, 2021).

Figura 11

Garantías que soportan la inferencia de Definición de Dominio

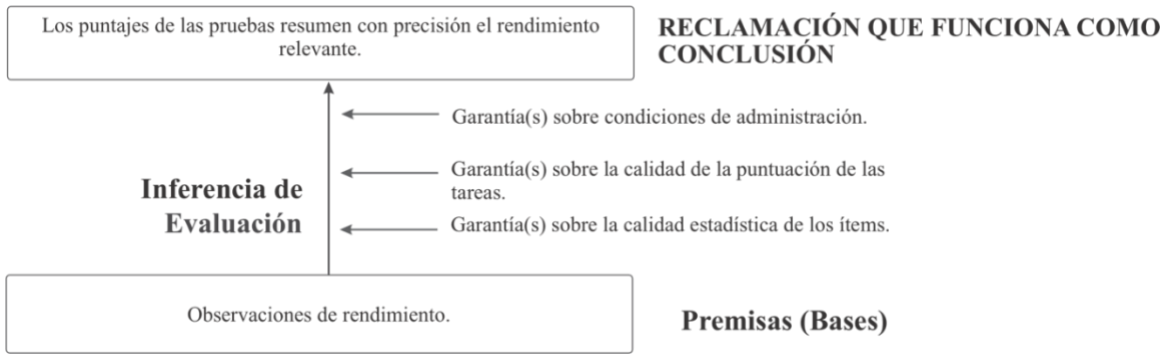


Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 93), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Inferencia de Evaluación. En esta etapa, presentada en la Figura 12, se valora si las puntuaciones registradas son confiables y representan con exactitud el desempeño de los examinados (Chapelle, 2021). Para ello, se revisan aspectos como la estandarización e imparcialidad en la administración, el rigor en la corrección de los ítems y la relevancia de cada tarea con respecto al constructo teórico (Kane, 1992).

Figura 12

Garantías que soportan la inferencia de Evaluación



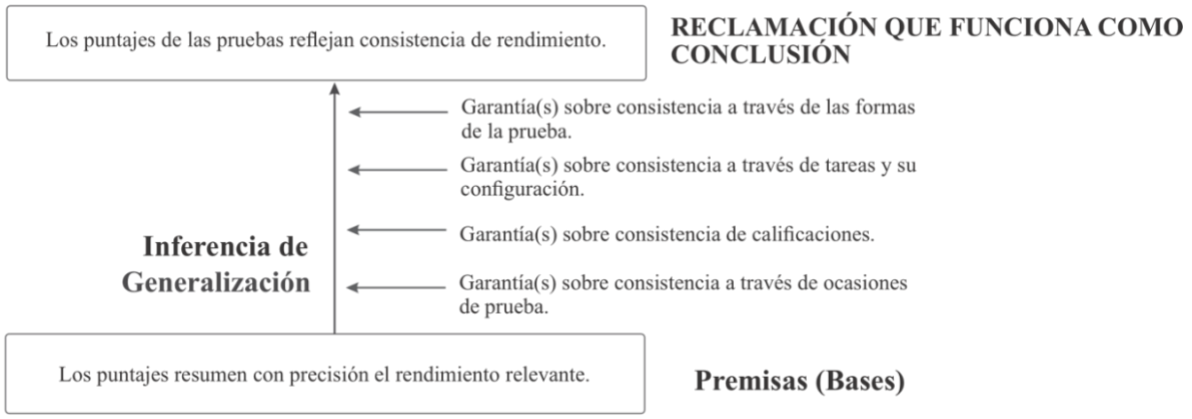
Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 81), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Además, se evalúa la igualdad de acceso a la prueba, lo cual incluye disponer de adaptaciones necesarias para garantizar la equidad. Los manuales de aplicación, las guías de puntuación, las estadísticas de dificultad y discriminación, así como los informes de control de calidad, constituyen las evidencias clave que permiten sostener la precisión de la inferencia de evaluación (Chapelle, 2021).

Inferencia de Generalización. A diferencia de la anterior, esta inferencia (ilustrada en la Figura 13) se enfoca en la posibilidad de extrapolar las puntuaciones a un conjunto más amplio de tareas o situaciones similares (Chapelle, 2021). Su objetivo es determinar si, en caso de cambiar algunos ítems o contextos de evaluación, los resultados se mantendrían estables.

Figura 13

Garantías que soportan la inferencia de Generalización



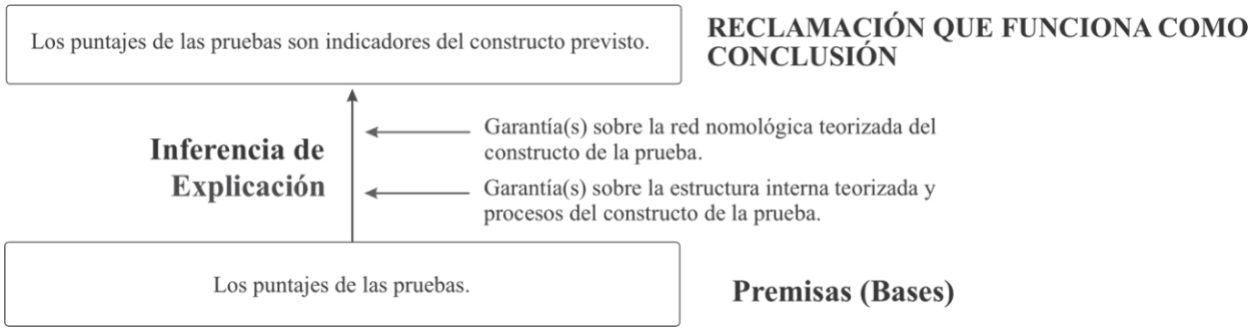
Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 74), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Los datos de confiabilidad (coeficientes de consistencia interna, retest, etc.), los estudios de equivalencia (comparación entre distintas versiones de la prueba) y los análisis de estructura interna (para verificar la dimensionalidad) brindan la evidencia de que los puntajes pueden generalizarse a otras condiciones más allá de las observadas (Kane, 2006).

Inferencia de Explicación. Esta inferencia, referida en la Figura 14, aborda la interpretación sustantiva de los puntajes, planteando que estos reflejan un rasgo o constructo subyacente (Messick, 1989; Chapelle, 2021). Para sostener tal afirmación, se requiere articular una red nomológica que defina las relaciones entre ese constructo y otras variables conceptualmente afines.

Figura 14

Garantías que soportan la inferencia de Explicación



Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 56), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

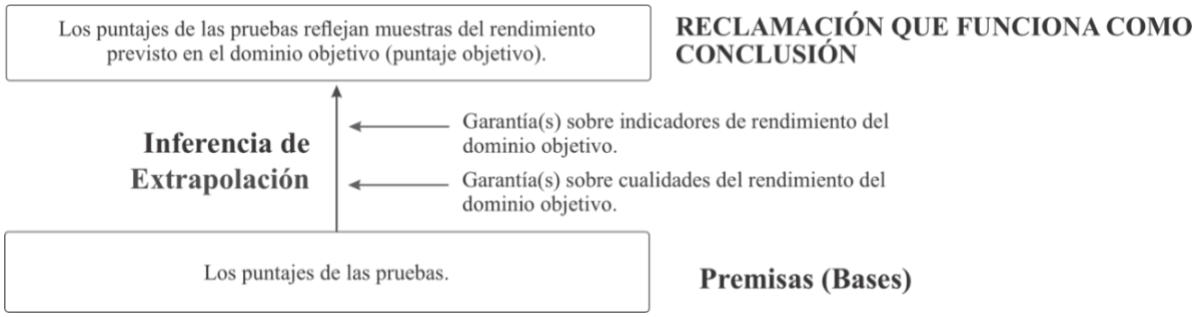
También resulta esencial confirmar la estructura interna mediante análisis factorial u otras metodologías, de modo que se demuestre la coherencia entre lo teorizado y la manera en que las puntuaciones se organizan (Cronbach, 1971). Las evidencias empíricas incluyen correlaciones con medidas relevantes, estabilidad temporal de los rasgos y validaciones cruzadas con otros instrumentos.

Inferencia de Extrapolación. Basada en la tradición iniciada por Cronbach y Meehl (1955), esta inferencia, vista en la Figura 15, se orienta a demostrar que el desempeño registrado en la prueba se extiende a contextos reales o más amplios que el escenario evaluativo (Chapelle, 2021). En otras palabras, si los puntajes altos de un examen de idiomas predicen el éxito académico en cursos avanzados, o si las destrezas evaluadas en un examen de ingreso se

manifiestan de manera consistente en las asignaturas iniciales de la carrera. Para ello, se requiere correlacionar los puntajes con medidas de rendimiento futuro o concurrente, comparar cualitativamente las tareas de la prueba con las exigencias del entorno real y revisar la alineación entre los ítems y los desafíos propios del dominio de aplicación.

Figura 15

Garantías que soportan la inferencia de Extrapolación



Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 63), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

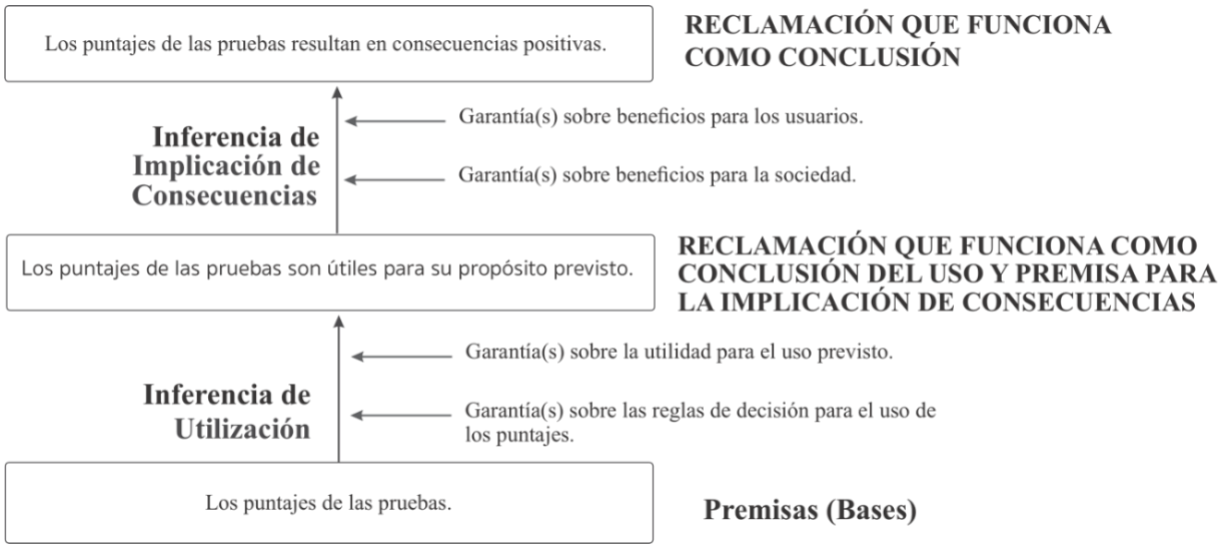
Inferencia de Utilización e Implicación de Consecuencias. Finalmente, la Figura 16 muestra cómo las puntuaciones se aplican con un fin determinado y, en consecuencia, generan repercusiones para las personas y a la sociedad (Chapelle, 2021). Aquí se analizan las decisiones adoptadas con base en los resultados —por ejemplo, la admisión o exclusión de un candidato— y se valoran los efectos deseados e indeseados de dichas decisiones.

Estas inferencias se pueden analizar de forma dividida, pero también se pueden unir en su análisis ya que son consecuentes (Chapelle, 2021). En el caso de los exámenes de ingreso a la educación superior, como indica Chapelle (2021), las consecuencias pueden incluir, entre otras, la eficiencia en la asignación de plazas, la motivación de los estudiantes para prepararse mejor académicamente o la reducción de sesgos en la selección; sin embargo, también pueden emerger efectos no deseados, como la presión económica para costear cursos de preparación, el estrés y la

ansiedad excesivos, la marginación de aspirantes con perfiles no contemplados por el examen o la sobrevaloración de resultados cuantitativos en detrimento de otras habilidades.

Figura 16

Garantías que soportan la inferencia de Utilización e Implicación de Consecuencias



Nota. Adaptado de *Argument-Based Validation in Testing and Assessment* [traducción propia] (p. 40), por C. A. Chapelle, 2021, SAGE Publications. Copyright 2021 de SAGE Publications.

Tal enfoque es coherente con las aportaciones de Cureton (1951), Cronbach (1971) y Messick (1989), quienes subrayaron la importancia de sopesar la utilidad, la relevancia práctica y los posibles impactos adversos. Si bien los desarrolladores esperan beneficios como la selección más precisa de aspirantes y la transparencia de los procesos de admisión, se deben reconocer y documentar los costos sociales, las posibles desigualdades que puede generar el examen y la llamada enseñanza orientada a la prueba (o washback), todo ello para preservar la solidez del argumento y mantener un equilibrio entre los beneficios y los costos asociados al uso de la prueba (AERA et al., 2014). Por tanto, las inferencias de Utilización e Implicación de Consecuencias incluyen no solo la funcionalidad práctica del examen, sino también su efecto formativo, ético y social, requiriendo que las instituciones realicen una supervisión continua y evalúen el impacto real de las decisiones tomadas.

Identificar qué inferencias proponer

Finalmente, el evaluador debe reconocer las afirmaciones, inferencias, que orienten la interpretación y el uso de las puntuaciones (AERA et al., 2014). Cada inferencia se respalda con garantías y suposiciones, además de evidencias que refuercen la validez (Kane, 2013). Para Chapelle (2021), construir un Argumento de Validez implica tres etapas:

1. Identificación de las inferencias y afirmaciones requeridas.
2. Recopilación de la evidencia necesaria, diseñando estudios empíricos.
3. Interpretación de los hallazgos y valoración de su pertinencia para sostener el argumento.

La Tabla 17 muestra un conjunto de preguntas para la planificación de un argumento de interpretación/uso de una prueba existente. Cada pregunta aclara qué inferencias pueden plantearse y qué investigaciones son esenciales para respaldarlas (Chapelle, 2021).

Tabla 17

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.

Para pruebas que se construyen por primera vez, la Tabla 18 presenta un esquema similar en el que se explora la existencia o la necesidad de un dominio claramente definido, la forma en que la calificación y administración pueden influir en los puntajes, la utilidad de esos puntajes y las consecuencias que se esperan del uso de la prueba. Cada pregunta busca clarificar las inferencias que podrían requerirse y las investigaciones que harían falta para confirmarlas o rechazarlas.

Tabla 18

Guía para planificar un argumento de interpretación/uso sobre una prueba nuevo

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.

El objetivo final es analizar cada inferencia de manera lógica y argumentada, en este sentido, Chapelle (2021) subraya que la comunicación entre los investigadores y los diferentes actores involucrados (comunidades discursivas, según Kane, 2006) resultan esenciales para unificar criterios y fortalecer los argumentos de validez. Por ende, el rol de los validadores también entra en juego: es preciso definir si el evaluador es parte del equipo desarrollador o un consultor externo, así como si la prueba ya existe o está creándose. Aquellos con un rol interno se concentran en diseñar argumentos que respalden las interpretaciones de sus propias pruebas, mientras que los externos pueden adoptar miradas más críticas o enfocarse en aspectos distintos.

Antecedentes: aplicación del EBA en la literatura

La validez, como se mencionó en el *Marco Teórico*, constituye uno de los conceptos más relevantes en el ámbito de la evaluación educativa y psicológica, pues de ella depende que los resultados de las pruebas reflejen con precisión aquello que deben medir. En las últimas décadas han surgido debates y propuestas sobre cómo conceptualizar y evaluar la validez de manera integral (Newton & Shaw, 2014; Chapelle, 2021). Bajo este panorama, el EBA—principalmente formulado por Kane (2006, 2013) y retomado por Chapelle (2015, 2021)—ha cobrado fuerza como una estrategia metodológica que no se limita a acumular datos psicométricos, sino que destaca la construcción de argumentos lógicos y coherentes para respaldar la interpretación y el uso de los puntajes.

En la búsqueda de investigaciones que esclarecieran la operatividad de este enfoque, se identificaron algunas RSL. Por ejemplo, Hatala et al. (2015) realizaron una revisión sistemática sobre la validez en herramientas de evaluación de habilidades técnicas en el área de la salud y, aunque siguieron rigurosamente el marco de Kane, los análisis y orientaciones se centraron en contextos de formación clínica, sin extrapolar directrices o procedimientos aplicables directamente a la educación universitaria general o a exámenes de admisión académica. De modo similar, Cheng y Sun (2015) examinan en detalle los efectos de alto impacto de los exámenes de idiomas bajo el EBA, pero reconocen que la mayor parte de la evidencia y discusiones giran en torno a contextos de enseñanza de lenguas, omitiendo el desarrollo de ejemplos, pasos operativos o marcos argumentativos explícitos para validaciones fuera de ese ámbito disciplinar.

Asimismo, Lavery et al. (2020) han destacado que, si bien existe consenso sobre la importancia de la validación argumentada, la literatura académica tiende a discutir el EBA en términos generales, priorizando la justificación teórica sobre procedimientos operativos o

estudios de caso detallados para ámbitos universitarios. Además, en su revisión sistemática de artículos de revisión por pares, advierten que la mayoría de los trabajos sobre validación argumentada describen los principios del enfoque y sus fundamentos, pero rara vez proporcionan guías prácticas o esquemas estructurados aplicables directamente a la validación de exámenes. Los autores subrayan que, fuera del ámbito de la salud o de la enseñanza de lenguas, existe una escasez de estudios que documenten paso a paso cómo proceder con el EBA en contextos educativos.

Por lo anterior, se realizó una RSL, cuyo método se explicita en el Apéndice A, con el objetivo de identificar cómo se operacionaliza el EBA en estudios empíricos del área educativa mediante el modelo de Elementos Preferidos para Informes de Revisiones Sistemáticas y Metaanálisis (PRISMA, por sus siglas en inglés; Page et al., 2021), en particular aquellos que utilizan instrumentos de evaluación a nivel universitario con el fin de establecer un marco metodológico claro y coherente en los procesos de validación desde este enfoque.

Al establecer un objetivo concreto para esclarecer los antecedentes, se proponen seis apartados: las características de los artículos revisados; cómo especifican el AIU estos estudios; identificar las inferencias que se utilizan en la evaluación del argumento; las técnicas utilizadas por inferencia; así como comprender cómo evalúan la validez global del argumento; y, finalmente, el análisis de la operacionalización del EBA.

Características de los estudios

Como se observa en la Tabla 19 se identificaron 28 artículos que mencionaron de forma explícita el uso del EBA en sus estudios. Más de la mitad de los artículos revisados (18 de 28, aproximadamente el 64%) provienen de Norteamérica, con una representación destacada también en Asia (4 estudios, 14%), Medio Oriente (3 estudios, 11%) y Europa (3 estudios, 11%).

Este predominio es consistente con estudios anteriores que documentan la aplicación de enfoques de validación en evaluaciones educativas en estas regiones (Chapelle et al., 2008; Kane, 2013). La distribución geográfica sugiere que el EBA tiene una fuerte presencia en contextos educativos específicos, particularmente en el ámbito anglosajón. Las revistas más representativas fueron *Language Testing in Asia* y *Language Testing*, que en conjunto agrupan un tercio de los estudios (10 de 28, 36%), destacando una concentración temática en pruebas de lengua (Chapelle, 2021; Koizumi et al., 2016).

Esta inclinación temática se relaciona con la variabilidad en la expresión del AIU, donde presentan definiciones literales hasta preguntas, objetivos, propósitos e hipótesis. Aproximadamente un tercio de los estudios (9 de 28, 32%) formularon el AIU como objetivos claros, mientras que otros optaron por hipótesis (5 de 28, 18%) o propósitos más generales (7 de 28, 25%). Este rango de formulaciones refleja la flexibilidad del EBA y su capacidad de adaptación a diferentes contextos y necesidades de investigación. No obstante, la variabilidad en la claridad y especificidad de estas declaraciones puede influir en la comprensión y aplicación de los marcos de validación, como sugieren Chapelle (2008) y Kane (2013). Los estudios con declaraciones más específicas, como los objetivos, ofrecen guías más claras para la validación, mientras que aquellos con propósitos más amplios presentan un contexto más general, pero menos detallado.

Tabla 19

Distribución anual de artículos por revista (2014–2023)

Revista	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Totales	
AERA Open	0	0	0	0	0	0	0	0	1	0	1	3.6%
Advances in Health Sciences Education	0	0	0	0	1	1	0	0	0	0	2	7.1%
Assessing Writing	0	0	0	0	1	0	0	1	0	0	2	7.1%
BMC Medical Education	1	0	0	0	0	0	1	0	0	1	3	10.7%

Educational Assessment	0	0	0	0	0	0	1	0	0	0	1	3.6%
Frontiers in Education	0	0	0	0	0	0	0	1	0	0	1	3.6%
Journal of Mathematics Teacher Education	0	0	0	0	0	0	0	0	0	1	1	3.6%
Journal of Teacher Education	1	0	0	0	0	0	0	0	0	0	1	3.6%
Language Assessment Quarterly	0	1	0	0	0	1	0	0	1	0	3	10.7%
Language Testing	1	1	0	1	0	1	0	0	0	0	4	14.3%
Language Testing in Asia	0	0	2	0	2	0	0	1	1	0	6	21.4%
Practical Assessment, Research & Evaluation	0	0	0	0	0	0	0	0	1	0	1	3.6%
Studies in Applied Linguistics & TESOL at Teachers College, Columbia University	0	0	0	0	0	0	1	0	0	0	1	3.6%
Zeitschrift für Didaktik der Naturwissenschaften	0	0	0	1	0	0	0	0	0	0	1	3.6%
Totales	3	2	2	2	4	3	3	3	4	2	28	100%
	10.7%	7.1%	7.1%	7.14%	14.3%	10.7%	10.7%	10.7%	14.3%	7.1%	100%	

Nota. Elaboración propia basada en el análisis de los 28 artículos distribuidos en 14 revistas. Los totales incluyen el número absoluto de artículos por año y el porcentaje relativo por revista.

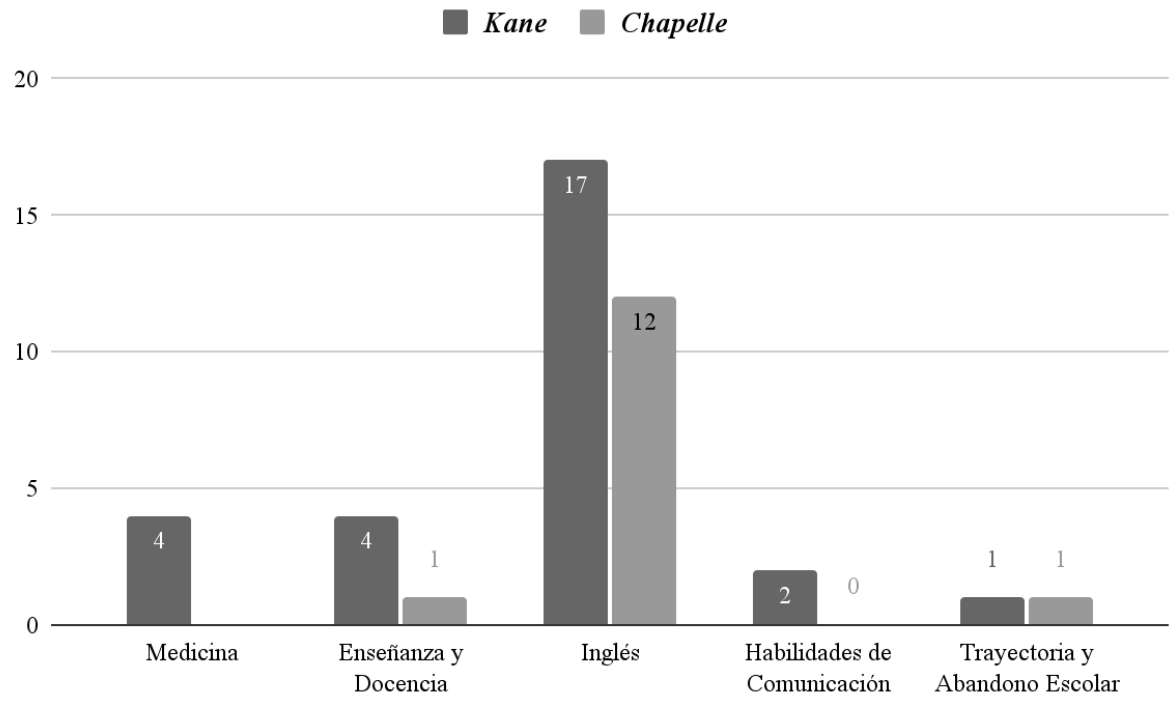
Y, como se observa en la Tabla 19, las fechas de publicación abarcan de 2014 hasta 2023, con una frecuencia de publicación anual de entre 2 y 4 artículos. Asimismo, las revistas con mayor número de publicaciones incluyen *Language Testing in Asia* (6 artículos) y *Language Testing* (4 artículos), seguidas *de Language Assessment Quarterly* y *Practical Assessment*, *Research & Evaluation* (3 artículos cada una).

Por otra parte, la Figura 17 muestra la aplicación de los enfoques de validación de Kane y Chapelle en diversas áreas de conocimiento. Donde, si se cita a Chapelle (2012, 2021) o Chapelle, Enright y Jamieson (2008, 2012) se cita a Kane. Así, los temas que se alinean con el idioma inglés citan mayoritariamente a Kane (17). En el área de medicina se tienen 4 estudios (véase Gráfica 20), subrayando la necesidad de validaciones precisas en evaluaciones médicas (Gotch & French, 2020). En enseñanza y docencia, hay 4 estudios de Kane y solo 1 de Chapelle, indicando un potencial para mayor investigación en esta área crítica para la calidad educativa (Chapelle et al., 2010). Habilidades de comunicación presenta 2 estudios de Kane y ninguno de

Chapelle, mientras que trayectorias escolares tienen 1 estudio cada uno, sugiriendo oportunidades de expansión en estas áreas (Kane, 2013).

Figura 17

Distribución de estudios por tema según referencia a Kane o Chapelle



Definición del AIU

El EBA, por su estructura, comienza con la identificación de los elementos clave que sustentan en el AIU. Esto incluye las afirmaciones que se desean realizar a partir de los puntajes, las inferencias que llevan a esas afirmaciones, las garantías que justifican dichas inferencias, y los supuestos subyacentes que deben cumplirse para que las inferencias sean válidas (Chapelle et al., 2008; Kane, 2013). La importancia del AIU radica en que proporciona una base explícita y estructurada para validar las decisiones derivadas de los resultados de las pruebas. Sin embargo, los estudios revisados en la Tabla 20 muestran que el AIU se declara en diferentes formatos, como definiciones literales, preguntas de investigación, objetivos, propósitos e hipótesis. Por otra

parte, **es** importante señalar que, en todos los estudios, los autores tratan de explicar de forma breve qué es el EBA con el fin de contextualizar la expresión del argumento.

Tabla 20

Formas de Expresar el AIU según los resultados de los estudios

Tipo de definición	Cita original	Traducción
Definición literal	"The AIU in this study is that the test scores between the expert raters will be in agreement when they use a previously developed POCUS assessment tool to score participants who perform ultrasound of the heart, aorta, and abdomen." (Sheppard et al., 2023, p. 2)	La AIU en este estudio es que las puntuaciones de la prueba entre los evaluadores expertos coincidirán cuando utilicen una herramienta de evaluación POCUS previamente desarrollada para puntuar a los participantes que realizan ecografías del corazón, la aorta y el abdomen. (Sheppard et al., 2023, p. 2)
Pregunta	"Validity studies are thus an important segment of the LMT instrument adaptations research aimed at addressing crucial questions: Are the adapted LMT items measuring the same construct as in the U.S.? Can they be used in other settings to measure the MKT—its level or growth—the same way they are used in the U.S.?" (Marcinek et al., 2023, p. 307)	Los estudios de validez constituyen, por tanto, un segmento relevante de la investigación sobre adaptaciones del instrumento LMT, orientada a responder preguntas cruciales: ¿Miden los ítems LMT adaptados el mismo constructo que en EE. UU.? ¿Se pueden emplear en otros contextos para medir el MKT—su nivel o crecimiento—de la misma manera que se utiliza en EE. UU.? (Marcinek et al., 2023, p. 307)
Objetivo	"For this aim, we examine the validity of the interpretation and use for assessing second language (L2) writing proficiency at a university in Japan, when the interpretation and use are made based on scores derived from Criterion®." (Koizumi et al., 2016, p. 2)	Para este propósito, examinamos la validez de la interpretación y el uso en la evaluación de la competencia en escritura en L2 en una universidad de Japón, cuando la interpretación y el uso se basan en las puntuaciones derivadas de Criterion®. (Koizumi et al., 2016, p. 2)
Propósito	"The purpose of this study was to develop and validate a computerized adaptive English proficiency testing (E-CAT) system for Taiwanese EFL university students." (Heng-Tsung et al., 2022, p. 163)	El propósito de este estudio fue desarrollar y validar un sistema de evaluación adaptativa por computadora de competencia en inglés (E-CAT) para estudiantes universitarios taiwaneses de EFL. (Heng-Tsung et al., 2022, p. 163)
Hipótesis	"We hypothesize that the InCoPrA is a feasible, acceptable and reliable method to provide a meaningful and supportive validated interpretation of ICS and Professionalism skills of the learners." (Abu Dabrh et al., 2020, p. 2)	Hipotetizamos que el InCoPrA es un método factible, aceptable y fiable para proporcionar una interpretación validada, significativa y de apoyo sobre las habilidades de ICS y profesionalismo de los aprendices. (Abu Dabrh et al., 2020, p. 2)

Nota. Traducción propia.

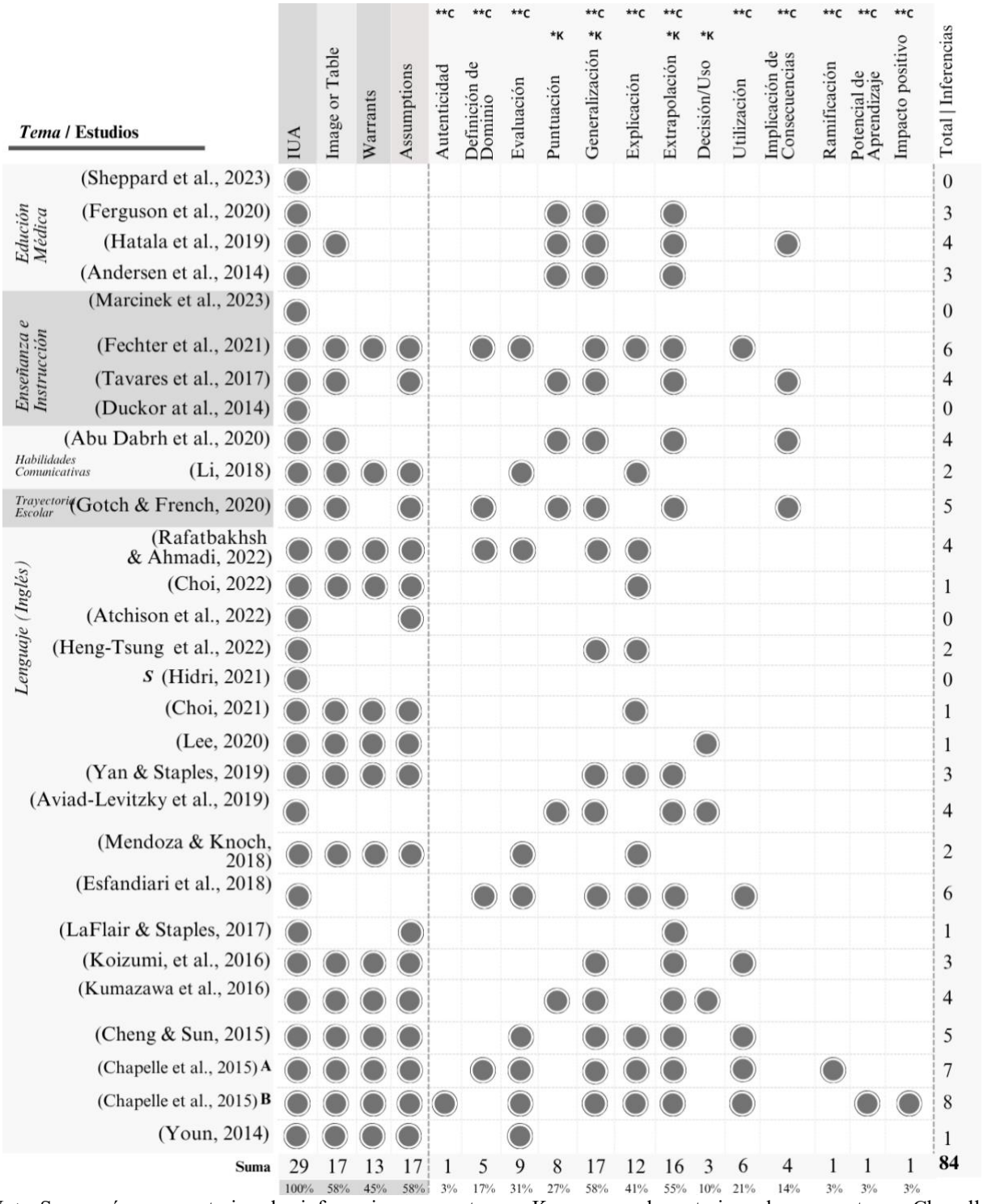
Aunque un objetivo y un propósito pueden parecer similares, no siempre mantienen la misma claridad. Un objetivo tiende a ser más específico y concreto, estableciendo metas claras y medibles para la investigación, mientras que un propósito puede ser más amplio y descriptivo, enfocándose en la razón subyacente del estudio. Esta variabilidad en la declaración del AIU puede afectar la claridad y la comprensión del marco de validación, ya que formatos más precisos como los objetivos pueden proporcionar una guía más directa para la validación, mientras que los propósitos y las hipótesis pueden ofrecer un contexto más amplio, pero potencialmente menos específico. A pesar de estas diferencias, la diversidad en la forma de declarar el AIU refleja la flexibilidad y adaptabilidad de los marcos teóricos de validación.

Inferencias utilizadas

Según el EBA, un argumento está integrado por inferencias, garantías y supuestos acordes con el Modelo de Toulmin (1958/2003). La Figura 18 resume el uso de inferencias por artículo, en total 23 de los 28 presentan alguna inferencia de forma explícita dando un total de 80 inferencias. Se identificó un uso constante de la inferencia Generalización (17) y de Extrapolación (16). Además, sólo un subconjunto de autores (Fechter et al., 2021; Li, 2018; Rafatbakhsh y Ahmadi, 2022; Choi, 2021, 2022; Lee, 2020; Yan & Staples, 2019; Mendoza & Knoch, 2018; Koizumi et al., 2016; Kumazawa et al., 2016; Cheng & Sun, 2015; Chapelle et al., 2015; Youn, 2014) describen las garantías y supuestos, reflejando un desajuste entre la teoría y la práctica sobre cómo representar un AIU (Lavery et al., 2020).

Figura 18

Resumen de las inferencias abarcadas en cada estudio



Nota. Se marcó con un asterisco las inferencias propuestas por Kane, y con dos asteriscos lo propuesto por Chapelle. Para los estudios de Chapelle et al. (2015) se marcó A y B para diferenciar los estudios.

Por otro lado, es importante resaltar que quienes sí reflejan garantías y supuestos, así como proponer una imagen o tabla como guía, suelen citar a Chapelle (2012, 2021) o a Chapelle

con sus múltiples colaboradores (Chapelle et al., 2008, 2010; Chapelle & Voss, 2013; Chapelle & Douglas, 2006). Es interesante que Chapelle, Cotos y Lee (2015) proponen otras inferencias, en su estudio A hacen uso de la inferencia de ramificación con el fin de evaluar el impacto y los efectos secundarios del uso de un sistema de evaluación automatizada de la escritura.

En cuanto al estudio B de Chapelle et al. (2015), desarrollan como base la inferencia de Autenticidad en lugar de Definición de Dominio, en sentido práctico son similares, ya que la inferencia de autenticidad busca asegurar que las tareas de escritura se alineen con las convenciones y expectativas de las investigaciones relacionadas a su disciplina. Asimismo, agregan dos inferencias en lugar de Implicación de Consecuencias (Chapelle, 2021): aprendizaje potencial e impacto positivo. La primera evalúa el potencial del sistema para promover el aprendizaje de los estudiantes y la segunda evalúa si los estudiantes tienen una experiencia de aprendizaje positiva.

Los estudios alineados a Chapelle fueron: Fechter et al. (2021) de Enseñanza e Instrucción*;* Li (2018) de Trayectoria Escolar; Rafatbakhsh y Ahmadi (2022), Heng-Tsung et al. (2022), Choi (2021), Yan & Staples (2019), Mendoza y Knoch (2018), Esfandiari et al. (2018), Koizumi, et al. (2016), Cheng & Sun (2015), Chapelle et al. (2015), y Youn (2014) de Lenguaje (véase Figura 18). También la inferencia de Explicación (12) fue desarrollada en gran medida, lo cual puede deberse al uso de métodos cuantitativos que se han utilizado de forma más común como el Análisis Factorial Exploratorio (AFE). Las siguientes inferencias más relevantes son Evaluación (9) y Puntuación (8), así como Definición de Dominio (5) que desde la teoría es la base para la definición y alineación de los usos posibles de los puntajes (Chapelle, 2021).

Asimismo, quienes desarrollaron más inferencias por estudio fueron Chapelle, Cotos y Lee (2015) con 7 para el estudio A y 8 para el B. Le siguen Fechter et al. (2021) y Esfandiari et

al. (2018) con 6 inferencias por estudio. Y, como se observa en la Figura 20, los temas apegados al área médica suelen hacer uso de las inferencias de Kane, retomando Puntuación, Generalización y Extrapolación (Ferguson et al., 2020; Andersen et al., 2014), aunque hay otros estudios que abarcan, además de esas inferencias, la inferencia de Implicación de Consecuencias (Hatala et al., 2019; Tavares et al., 2017; Abu Dabrh et al., 2020), donde el caso de Gotch y French (2020) es importante, ya que ubican la Definición de Dominio como un punto relevante para iniciar con el proceso de validación aunque no lo establecen como una inferencia. Asimismo, hay quienes abarcaron la inferencia de decisión o uso (Aviad-Levitzky et al., 2019; Kumazawa, Shizuka, Mochizuki, & Mizumoto, 2016; Lee, 2020).

En este sentido, se puede afirmar que hay un uso predominante de inferencias de Generalización y Extrapolación en la mayoría de los estudios. Estas dos inferencias son las más frecuentemente desarrolladas, con 17 de los 28 estudios (61%) empleando explícitamente Generalización y 16 estudios (57%) utilizando Extrapolación, lo que subraya su papel central en la validación de evaluaciones educativas (Fechter et al., 2021; Yan & Staples, 2019). Sin embargo, el enfoque de Chapelle (2021) destaca que, aunque el marco es flexible, no siempre se representan todas las inferencias posibles, lo que puede limitar la profundidad de la validación en algunos estudios. Este enfoque indica una prioridad en asegurar que las puntuaciones de las pruebas sean confiablemente generalizables más allá de las condiciones específicas de prueba y puedan extrapolarse a dominios más amplios o desempeños futuros, aspectos críticos de la validez de las pruebas (Chapelle, 2021; Kane, 2013).

Evaluación del argumento de Validez Global

En un tercer momento, el EBA propone recopilar y evaluar la evidencia que respalda las inferencias (Kane, 2006), aunque también se podrían crear las evidencias (Chapelle, 2021), esto

dependería del papel del evaluador, ya sea como interno o externo. Para evaluar el Argumento de Validez global Kane (2013) propone tres aspectos clave: claridad, coherencia y plausibilidad. La claridad refiere a la especificación precisa y comprensible de las inferencias y suposiciones necesarias para justificar el uso de los puntajes de una prueba, por ende, el proceso de validación comienza con la especificación del AIU. Este paso garantiza que las afirmaciones y los objetivos de la prueba estén claramente articulados desde el principio (Kane, 2013; Chapelle, 2021).

Por otro lado, la coherencia es la relación lógica y estructurada entre las evidencias y las inferencias. En este proceso, se recopilan evidencias que respalden o refuten las inferencias formuladas, asegurando que las diferentes partes del argumento estén conectadas de manera consistente y fundamentada (Kane, 2013). Y, la plausibilidad se relaciona con la credibilidad del argumento en su conjunto, es decir, si las conexiones entre las inferencias, evidencias y suposiciones son lo suficientemente robustas para sostener las conclusiones sobre los puntajes (Chapelle, 2021). Esta evaluación, en su conjunto, se entrelaza, si no hay claridad, faltará coherencia y sin éstas tampoco plausibilidad. El fin es integrar todas las evidencias, determinando el grado en que las inferencias son válidas y confiables, asegurando que las conclusiones derivadas del uso de los puntajes estén justificadas de manera empírica y lógica.

Según la Figura 18, aunque en algunos estudios los autores proporcionan una descripción detallada del proceso de validación (Choi, 2021, 2022; Chapelle et al., 2015; Fechter et al., 2021), incluyendo la Explicación de las inferencias, evidencias y suposiciones clave, en otros, los procedimientos para evaluar el argumento se dan por sentados, sin ofrecer una justificación explícita o un desglose detallado de cómo se lleva a cabo cada evaluación (Marcinek et al., 2023; Lee, 2020).

Aunque algunos autores ofrecen reflexiones importantes sobre el proceso de validación, no siempre se explicita cómo se evalúan las inferencias clave del AIU. Por ejemplo, Sheppard et al. (2023) señalan: "The results of this study do not provide enough evidence to support the AIU in the scoring domain. Therefore, the POCUS assessment tool requires further modification and testing prior before it can be used for reliable undergraduate POCUS assessment" (p. 6) [Los resultados de este estudio no proporcionan evidencia suficiente para respaldar el AIU en el ámbito de la puntuación. Por lo tanto, la herramienta de evaluación POCUS requiere más ajustes y pruebas antes de poder emplearse de manera fiable en el nivel de pregrado] (traducción propia). Esta afirmación evidencia una falta de respaldo empírico claro, aunque no se detalla cómo se valoran específicamente las inferencias formuladas. En un enfoque más estructurado, Chapelle et al. (2015) afirman: "The system of warrants and rebuttals in the interpretive argument is built dynamically in response to research needs and research outcomes" (p. 403) [El sistema de garantías y refutaciones dentro del argumento interpretativo se construye de forma dinámica, adaptándose a las necesidades y los resultados de la investigación] (traducción propia), lo que sugiere flexibilidad en la construcción del argumento, pero sin una evaluación formal de claridad, coherencia o plausibilidad.

Del mismo modo, Yan y Staples (2019) reconocen que "The framework is comprehensive, but its application in writing assessment presents challenges in evaluating all inferences, particularly extrapolation" [El marco es amplio, pero su aplicación en la evaluación de la escritura presenta retos al momento de valorar todas las inferencias, en especial la extrapolación] (traducción propia), señalando limitaciones específicas en la aplicación del EBA. En la misma línea, Mendoza y Knoch (2018) concluyen que "The validation efforts focused primarily on scoring and generalization, with limited attention to extrapolation and

consequences" [Los esfuerzos de validación se centraron en la puntuación y la generalización, prestando menos atención a la extrapolación y las consecuencias] (traducción propia), lo que refleja una visión parcial del Argumento de Validez global. Los autores señalan las inferencias donde se tiene mayor desarrollo y en cuáles se tiene limitaciones, pero no realizan una valoración global del argumento. Es decir, no se aborda de manera explícita si las inferencias fueron coherentes ni si las conexiones entre las evidencias y las conclusiones eran plausibles.

Por otra parte, la expresión del argumento puede representarse mediante un diagrama esquemático (Chapelle, 2021), lo que sugiere una claridad en este aspecto. Sin embargo, ninguno de los autores menciona explícitamente la evaluación global del AIU en términos de claridad, coherencia y plausibilidad. Así, en la práctica, muchos estudios emplean diversos formatos gráficos y tablas para visualizar y organizar el AIU de manera clara, facilitando la interpretación de los datos y la justificación de las inferencias.

Como se muestra en la Figura 18, 17 de los 28 artículos analizados incluyen algún tipo de tabla o diagrama que articula estas relaciones (Hatala et al., 2019; Fechter et al., 2021; Tavares et al., 2017; Abu Dabrh et al., 2020; Li, 2018; Gotch & French, 2020; Rafatbakhsh & Ahmadi, 2022; Choi, 2022, 2021; Lee, 2020; Yan & Staples, 2019; Mendoza & Knoch, 2018; Koizumi, et al., 2016; Kumazawa et al., 2016; Cheng & Sun, 2015; Chapelle et al., 2015), siguiendo modelos como el propuesto por Chapelle (2021), que organiza las inferencias, garantías y evidencias de manera explícita. Tavares et al. (2018) utilizaron un diagrama en su metodología para ilustrar el proceso de validación basado en la estructura del AIU y el Argumento de Validez, véase Figura 19. En esta figura, el proceso comienza con la formulación del AIU, seguido de la identificación y priorización de las afirmaciones e inferencias más críticas, como las relacionadas con la Puntuación, la Generalización, la Extrapolación y las implicaciones.

Figura 19

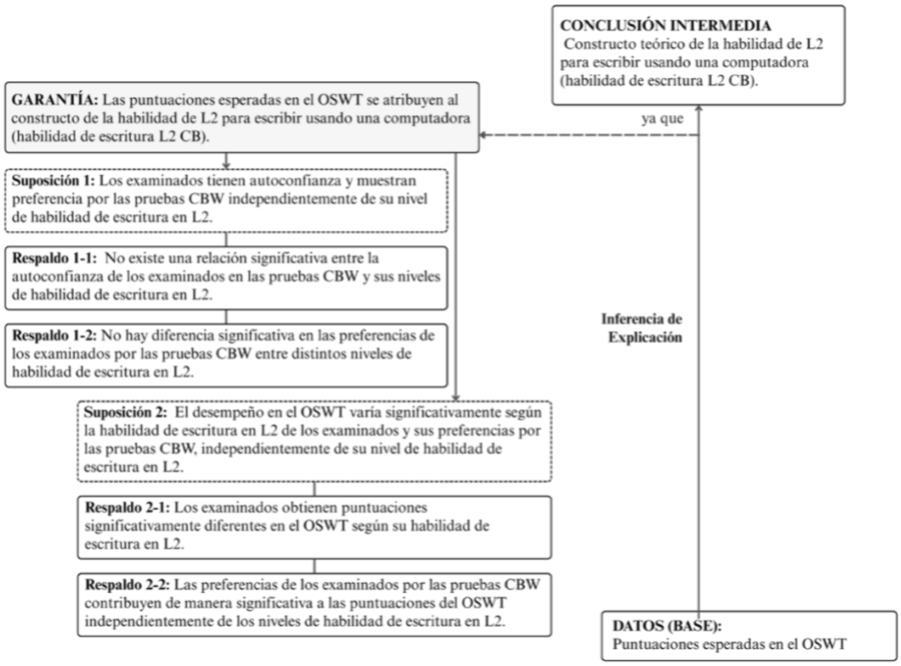
Esquema del proceso de validación propuesto por Tavares et al. (2018)

Nota. Adaptado de "Applying Kane's Validity Framework to a Simulation-Based Assessment of Clinical Competence" (p. 327) por W. Tavares, R. Brydges, P. Myre, J. Prpic, L. Turner, R. Yelle y M. Huiskamp, 2018, *Advances in Health Sciences Education, 23*(2), 323-338, (https://doi.org/10.1007/s10459-017-9800-3). Copyright 2018 de Tavares et al.

Luego, se crea un plan para poner a prueba estas afirmaciones, y finalmente evalúan las evidencias empíricas y los argumentos lógicos para formular un juicio sobre la validez. Además, Nomura et al. (2020), Dabrh et al. (2020), Rafatbakhsh y Ahmadi (2022), Tavares et al. (2017), Yan y Staples (2019), Mendoza y Knoch (2018), Li (2018), proponen un formato de tabla para expresar las inferencias, donde incluyen o no las garantías, supuestos y las evidencias que se recopilaron por cada uno de los supuestos; esto recordando que los supuestos emergen de las garantías (Chapelle, 2021). El argumento, en ocasiones, se agregó en una fila dentro de la tabla, como en el caso de Mendoza y Knoch (2018) para definir el argumento. Por otro lado, Choi (2021; 2022) y Fechter et al. (2021) retoman la propuesta realizada por Chapelle et al. (2008) basado en el modelo de Toulmin (1958/2003) para expresar la conclusión (afirmación), el uso de las puntuaciones, la garantía y sus supuestos, así como la evidencia por cada supuesto; pero la fraccionaron según a la inferencia a analizar y revisar (Figura 20).

Figura 20

Ejemplo de representación de las inferencias y fuentes de evidencia



Nota. Adaptado de "Validity of Score Interpretations on an Online English Placement Writing Test" [traducción propia] (p. 7), por Y. Choi, 2022, *Language Testing in Asia, 12*, Article 42, (https://doi.org/10.1186/s40468-022-00187-0). Copyright 2022 de Y. Choi.

En la Tabla 21 se expresa un modelo de cómo, en general, se puede expresar un argumento incluyendo sus inferencias, garantías, supuestos y evidencias. Se identificó que cuando hay más de una inferencia los autores suelen inclinarse por este tipo de tablas. En este sentido, una inferencia puede tener diversas garantías y cada garantía diferentes suposiciones y éstas una variedad de evidencias. Asimismo, no es necesario hacer uso de todas las inferencias, como hemos visto en el apartado anterior, por lo que se pueden proponer AIU intermedios.

Tabla 21

Modelo para organizar un Argumento de Validez a partir del EBA

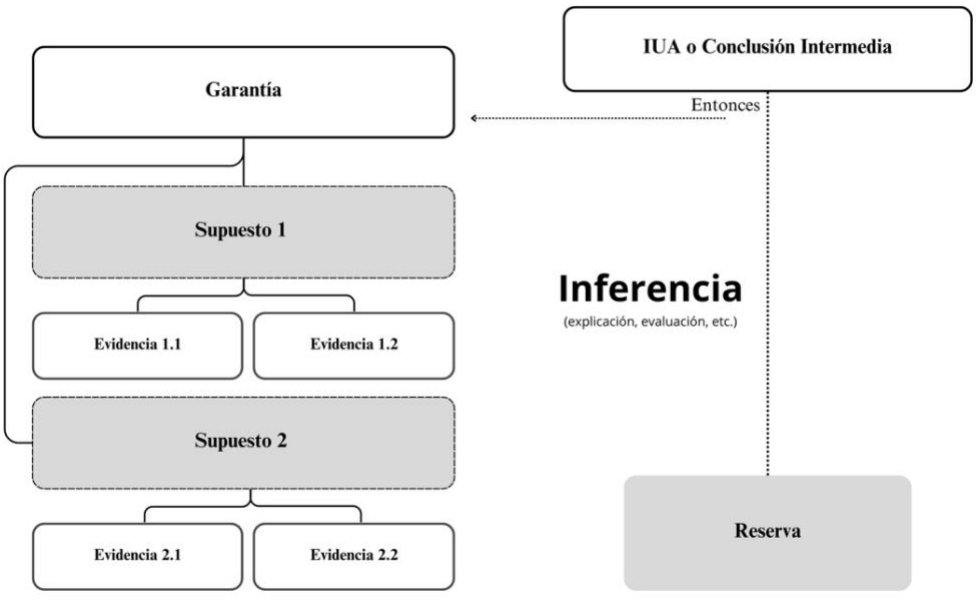
AIU o Conclusión Intermedia			
Calificador			
Inferencia	**Garantía**	**Supuesto**	**Evidencia**
Evaluación	Garantía 1	Supuesto 1.1	Evidencia 1.1.1
		Supuesto 1.2	Evidencia 1.1.2

		Supuesto 1.3	Evidencia 1.1.3
	Garantía 2	Supuesto 2.1	Evidencia 2.1.1
		Supuesto 2.2	Evidencia 2.2.1
			Evidencia 2.2.2
Explicación	Garantía 1	Supuesto 1.1	Evidencia 1.1.1
		Supuesto 1.2	Evidencia 1.1.2
		Supuesto 1.3	Evidencia 2.1.1
			Evidencia 2.1.2
			Evidencia 2.1.3
Extrapolación	Garantía 1	Supuesto 1.1	Evidencia 1.1.1
		Supuesto 1.2	Evidencia 1.1.2

Finalmente, la Figura 21 es el modelo que se logró extraer de los estudios de Choi (2021; 2022) y Fechter et al. (2021). Esta, Figura 21, se apega al modelo de Toulmin (1958/2003) para construir su argumento, respetando el fundamento, así como el cualificador, pero en este caso es por cada una de las inferencias, es decir, este diagrama se adapta a estudios que se enfocan a una sola evidencia. No obstante, cuando se presentan más garantías y suposiciones el diagrama se puede volver muy amplio.

Figura 21

Modelo de expresión de una sola inferencia



Implementación metodológica del EBA

Como se analizó en el apartado anterior, para abordar la limitación de la operacionalización y mejorar la claridad y coherencia del argumento, la representación visual de las inferencias es clave. En más de la mitad de los estudios analizados (17 de 28, 61%), se incluyen diagramas y tablas que articulan las relaciones entre inferencias, garantías y evidencias, lo cual mejora significativamente la comprensión del proceso de validación basado en argumentos (Choi, 2022; Fechter et al., 2021). En este sentido, la Tabla 27 retoma todos los aspectos del modelo de Toulmin (1958/2003), facilitando una presentación sistemática y completa del argumento (Chapelle, 2021). Sin embargo, el hecho de que cerca del 40% de los estudios (11 de 28) no utilicen representaciones visuales indica una oportunidad para fortalecer el proceso de validación mediante una representación más exhaustiva.

Un aspecto clave en esta revisión es la evaluación de la validez global, a través de la claridad, coherencia y plausibilidad, dimensiones propuestas por Kane (2013). A pesar de estas herramientas y marcos teóricos, se observa que los estudios no realizan una evaluación global de la validez ni una expresión explícita del nivel o grado de validez alcanzado. Algunos estudios, como los de Choi (2021) y Chapelle et al. (2015), ofrecen una descripción detallada de los procesos de validación, mientras que otros lo presentan de manera implícita, sin detallar cada inferencia o justificación (por ejemplo, Marcinek et al., 2023; Lee, 2020). Chapelle (2021), a diferencia de Kane (2006), no se enfoca en proponer una evaluación explícita de la claridad, coherencia y plausibilidad. Su enfoque radica en la organización de las inferencias, garantías y supuestos, bajo la premisa de que esta estructura, por sí sola, proporciona un nivel o grado de validez. Así, se centra en la presentación y articulación de estas relaciones, como un indicador implícito de la validez, lo cual marca una diferencia crítica en cómo se conceptualiza la

evaluación en el EBA. Sin embargo, esto plantea interrogantes sobre la suficiencia de esta organización para reflejar un grado adecuado de validez y sobre la necesidad de cualificadores modales para evaluar dicho grado. La falta de una evaluación explícita de estos aspectos se puede atribuir a la flexibilidad del EBA y a la ausencia de mecanismos estandarizados, lo cual permite interpretaciones diversas en la aplicación de estos principios de validación (Lavery et al., 2020; Chapelle, 2021).

Esta situación evidencia la falta de claridad sobre cómo realizar la evaluación global de la validez, un aspecto crítico que requiere mayor atención. La ausencia de directrices explícitas y procedimientos estandarizados para llevar a cabo una evaluación global dentro del marco del EBA puede deberse a la complejidad de integrar múltiples inferencias y evidencias en un juicio cohesivo (Gotch & French, 2020; Kane, 2013). El enfoque de Chapelle (2021) sobre la cadena de inferencias, aunque flexible, necesita ser aplicado de manera más completa para abordar esta falta de claridad. Partiendo de lo anterior, las conclusiones puntuales son las siguientes:

1. Hay una clara inclinación hacia la valoración de exámenes del idioma inglés y del área médica, representando juntos más del 70% de los estudios analizados.
2. La expresión del AIU es variable, optando por varios formatos: objetivos, hipótesis o definiciones, donde aproximadamente un tercio de los estudios (32%) utilizan objetivos claros.
3. Uso predominante de Inferencias de Generalización y Extrapolación, presentes en más del 60% de los estudios.
4. La representación visual de las inferencias aporta tanto a la claridad como a la coherencia del argumento; sin embargo, cerca del 40% de los estudios no la emplean.

5. No se realiza una evaluación global de la validez, ni una expresión explícita del nivel o grado de validez en ninguno de los estudios revisados.
6. Hace falta claridad sobre cómo realizar la evaluación global de la validez, lo cual es un aspecto crítico que requiere mayor atención en futuras investigaciones.
7. El EBA expresado por Chapelle (2021) es la mejor propuesta sobre su operacionalización.

El EBA ha demostrado ser una herramienta valiosa y relevante para estructurar y evaluar la validez de la interpretación de los puntajes en pruebas, al ofrecer un desglose detallado de inferencias, garantías y suposiciones. Sin embargo, como destacan Durson y Li (2021), muchos estudios no implementan el EBA de forma completa, lo que genera inconsistencias en la evaluación de inferencias clave, como la Extrapolación y la Generalización, debilitando el Argumento de Validez.

Por último, esta situación plantea preguntas importantes: ¿por qué la claridad, coherencia y plausibilidad no se evalúan de manera explícita y sistemática en todos los estudios? ¿Es suficiente, en términos de validez, la presentación visual de tablas o diagramas, como el modelo de Toulmin (1958/2003), o sería necesario también especificar el cualificador modal que indique el grado de validez alcanzado? ¿Es posible desarrollar criterios estandarizados para evaluar la claridad, coherencia y plausibilidad en el EBA que puedan aplicarse de manera consistente en distintos contextos? Revisiones sistemáticas como las de Lavery et al. (2020) y Durson y Li (2021) mencionan la necesidad de desarrollar prácticas más transparentes y estructuradas que aseguren una evaluación explícita de estas dimensiones en futuros estudios de validación, por lo que tratar de desarrollar criterios podría ser beneficioso para la claridad del EBA.

Diseño metodológico

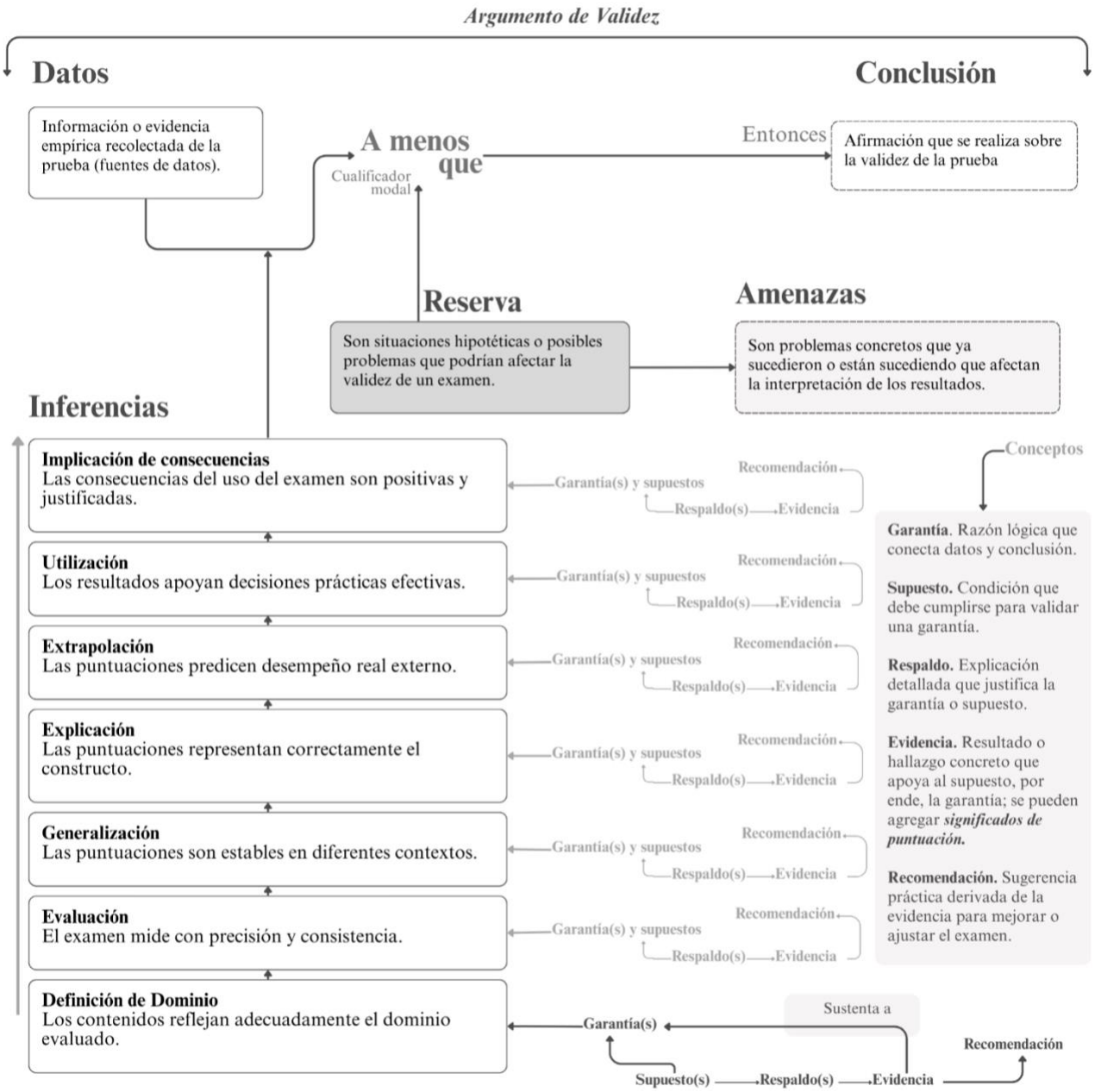
El presente diseño metodológico busca responder al objetivo general, el cual refiere a la evaluación de la interpretación y del uso de los puntajes del ExIES (2023-1) como criterio de admisión a los programas de licenciatura de la UABC. Para cumplir este propósito se adopta un enfoque evaluativo ya que, como apuntan Jornet et al., (2010; 2020), se orientan a la construcción y validación de instrumentos a través de procesos minuciosos, por lo que forma parte de la investigación evaluativa.

Por lo anterior, se retoma el EBA, el cual, gracias a su flexibilidad, articula de manera argumentativa y sistemática múltiples fuentes de evidencia documental, cualitativa y cuantitativa alrededor de un AIU (Chapelle, 2021). Esta selección responde a que si bien los Estándares (AERA et al., 2014) indican qué evidencias deben recabarse, el EBA especifica cómo concatenarlas y organizarlas, de modo que los argumentos se concatenen a través de siete inferencias. Al mismo tiempo, estas inferencias permiten integrar los tres elementos esenciales de dichos Estándares—evidencias de validez, confiabilidad e imparcialidad—dentro de un argumento lógico, coherente y explícito.

Una dificultad en la aplicación práctica del EBA ha sido el cómo articular la teoría con la práctica. Por ende, la Figura 22 funciona como una fórmula donde se concentran todos los elementos del EBA. Para su elaboración se retomó la expresión clásica del modelo de Toulmin (1958/2003), véase Figura 9, de Kane (2006, 2013); y la Figura 10 propuesta por Chapelle (2021), donde muestra de forma clara la concatenación entre inferencias y su relación con las garantías y supuestos. Asimismo, se tomó en cuenta la Tabla 15 para la definición conceptual.

Figura 22

Elementos clave para el diseño del proceso de validación según el EBA



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; C. A. Chapelle, 2021, SAGE Publications).

Considerando que esta metodología permite una evaluación externa al equipo del ExIES se propone basarse en fuentes documentales. A continuación, se presenta la descripción de las fuentes de datos, así como las características detalladas del instrumento como objeto de estudio (ExIES) y el procedimiento, profundizando particularmente en sus etapas y fases operativas;

donde se contemplan la valoración de las siete inferencias alineados a los objetivos de la presente investigación.

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 proporcionados 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 el ciclo 2023-2, 2024-1 y 2025-1. En la Tabla 22 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 22

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 le Evaluación de ítems del Nuevo Examen de Selección de aspirantes a ingresar a la Universidad Autónoma de Baja California (Caso & 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 & 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:

- 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.

En la Tabla 23 del ExIES se pueden observar las especificaciones. El área de Lectura se centra en la evaluación del contenido, el análisis discursivo y la síntesis de diversas fuentes. La sección de escritura revisa el desarrollo temático y la adecuación a las normas gramaticales del español. Por su parte, en Matemáticas, se abordan aspectos como las ecuaciones lineales, análisis de datos, funciones exponenciales y conceptos avanzados relacionados con el área y el volumen.

Tabla 23

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 24 detalla los Niveles de Demanda Cognitiva (NDC) del ExIES. La Comprensión implica entender el significado del material, con un 8% total. Aplicación, con el 46% total, se refiere a usar el material en situaciones nuevas. El nivel de Análisis representa el 5% y trata sobre descomponer el material. La Síntesis tiene el 11% y se enfoca en integrar partes para crear un conjunto. Por último, la Evaluación, que es el 30% total, implica juzgar el valor del material.

Tabla 24

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 25, 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. De estos ítems, el 30% son ancla, es decir, son ítems que se incluyen en todas las versiones del examen, lo que permite equiparar resultados entre distintas poblaciones o períodos. El restante 70% corresponde a ítems que aparecen en una sola versión.

Tabla 25

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).

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. Estos ítems piloto fueron de recién elaboración y se evaluaron para decidir su posible incorporación futura, sin considerarse para el desempeño de los aspirantes. Además, se agregó también una subversión adaptada especialmente para estudiantes con discapacidad visual grave.

Participantes en la elaboración del instrumento

El proceso de elaboración del ExIES reúne a cinco tipos de participantes con funciones diferenciadas (véase Tabla 26). 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 26

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

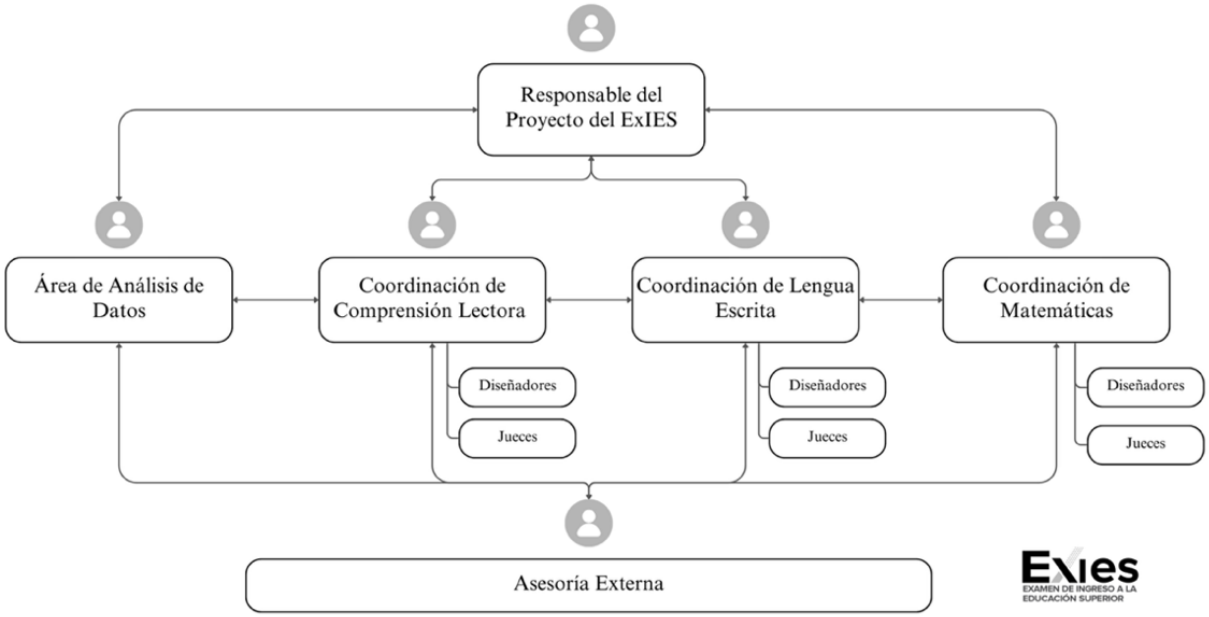
Nota. Elaboración propia y retomando usuarios del *Examen de ingreso a la educación superior (ExIES) 2023-1: Reporte técnico* (p. 10), por Pedroza Zúñiga et al., 2024a.

Equipo ExIES. El equipo, véase Figura 23, 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.

Figura 23

Organigrama del Equipo del ExIES



Nota. Las flechas indican la comunicación retroalimentada y constante entre los miembros.

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 27.

Tabla 27

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 28 detalla esta distribución.

Tabla 28

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,

Procedimiento

El procedimiento se basa en la Figura 22. El proceso de validación inicia con los datos: toda la información empírica que se obtiene de la prueba (resultados de ítems, respuestas de encuestados, entrevistas, etc.). Estos datos alimentan una cadena ascendente de las siete inferencias; cabe recordar que pueden existir otro tipo de inferencias y esto depende del razonamiento que se realice a partir de los datos obtenidos. Cada salto –entre inferencia– se legitima mediante garantías (razones lógicas que conectan la premisa con la nueva afirmación) apoyadas en supuestos (condiciones que deben cumplirse) y en su respaldo (explicación de por qué el supuesto sustenta la garantía; puede ser sustentada de forma teórica, empírica o normativa). Sobre estas piezas se colocan las evidencias (observaciones o análisis concretos que confirman el respaldo). Si la evidencia es sólida, la inferencia se acepta, por ende, la cadena progresa. Así, la estructura asegura que ninguna afirmación sobre la prueba se dé por válida sin un puente lógico explícito y un soporte empírico observable.

Cuando la serie de inferencias culmina, se emite la conclusión: un juicio global sobre la validez de la prueba para el uso pretendido. Sin embargo, el modelo incorpora la Reserva (circunstancias hipotéticas que, de llegar a darse, invalidarían parte del argumento), con el fin de identificar objeciones o refutaciones. Así, el o la analista transforma cada evidencia en recomendaciones de mejora—p. ej. refinar ítems o ampliar la muestra, etc.— lo que cierra el ciclo evaluativo. En conjunto, el resultado del proceso no es sólo un veredicto, sino un

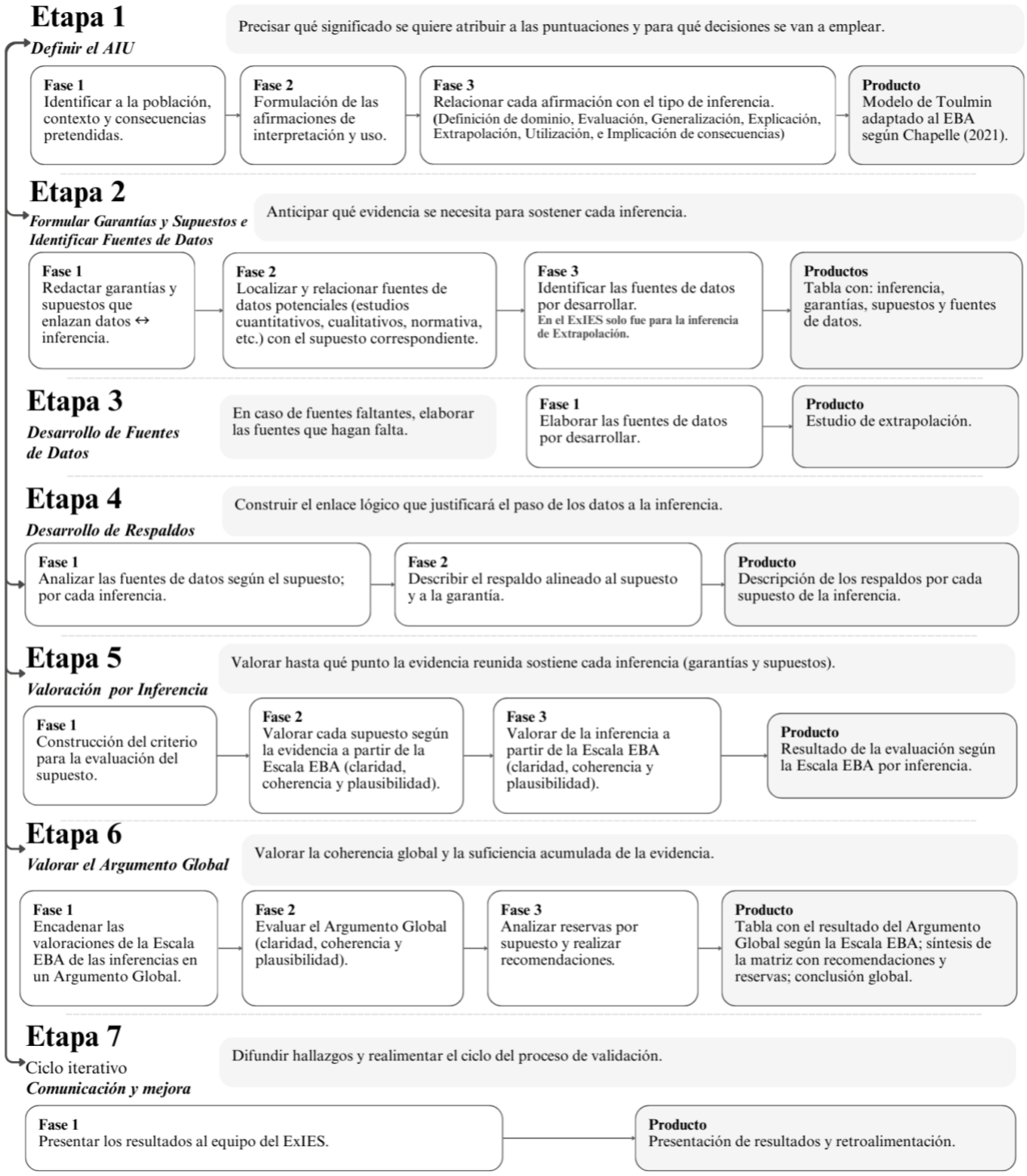
Argumento de Validez documentado: un mapa lógico-empírico que muestra cómo se pasó de los datos brutos a la decisión práctica, con sus garantías, condiciones y limitaciones claramente trazadas. Sin embargo, según lo encontrado en los *Antecedentes*, esta expresión visual puede variar y se ha encontrado que la mejor forma de expresar el Argumento de Validez es en formato de tabla.

Siguiendo esta lógica, en la Figura 24 se presenta el procedimiento seguido para el objetivo de la presente investigación, retomando los tres pasos de Kane (2006, 2013, 2020): 1) formulación del AIU; 2) Evaluación crítica del argumento; 3) conclusión global sobre la validez; y retomando 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. Las flechas (véase Figura 24) entre etapas indican que el proceso no es estrictamente lineal, sino iterativo, permitiendo, en cualquier momento, revisar y ajustar aspectos fundamentales del AIU, añadir nuevas fuentes de datos, ampliar su descripción o redefinir respaldos y garantías según las necesidades.

Asimismo, los resultados de la presente investigación se desarrollaron según las etapas; la primera etapa es el planteamiento del AIU, el cual es un borrador del Argumento de Validez; de la etapa dos a la cinco es la presentación de cada inferencia siguiendo las mismas fases para cada una; una vez presentadas las inferencias continúa la etapa seis sobre la valoración del Argumento Global, que refiere a la valoración del Argumento como Unidad según Kane (2006, 2013), para finalmente dar recomendaciones generales. Al concluir los resultados se procede a la discusión y conclusiones, donde se pretende retomar y contrastar lo encontrado tanto en el *Marco Teórico* como en los *Antecedentes*.

Figura 24

Procedimiento del proceso de validación del ExIES a partir del EBA



Etapa 1: Definir el AIU

La intención de esta primera etapa fue precisar el uso e interpretación que se le darían 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 a 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

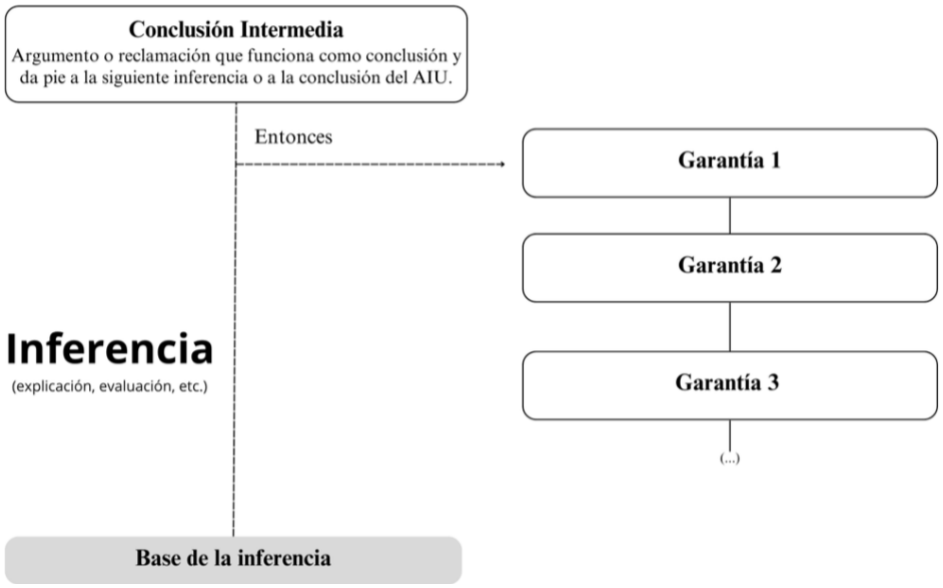
En la segunda etapa se especificaron las evidencias necesarias para respaldar cada inferencia planteada. Así, se formulan las garantías para después alinear los supuestos que, a su vez, relacionan lógicamente con las inferencias; es decir, se parte de seleccionar las fuentes posibles que fundamenten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

Fase 1. La Figura 25 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 25

Esquelete 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 16, donde se dan los ejemplos de investigaciones cualitativas o cuantitativas por inferencia como guía para identificar las evidencias; así mismo, la Tabla 17 sobre el cómo desarrollar el AIU para una prueba existente; y la Tabla 18 para una prueba nueva; propuestas por Chapelle (2021).

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 (Apéndice B). Por otra parte, un ejemplo de la estructura se muestra en la Tabla 29.

Tabla 29

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 Descripción del argumento	G1.1. Descripción de la garantía	S.1.1.1 Descripción del supuesto	F1.1.1 Descripción del dato
(…)	(…)	(…)	(…)

Para la realizar el Apéndice B, se elaboró la *Guía con indicaciones generales para proponer Inferencias, Garantías, Supuestos y Tipos de Evidencia para la Validación de EAI según el EBA* (Apéndice C), con el fin de orientar esta investigación como futuras investigaciones de otros EAI.

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, etc.

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

Tabla 30

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; véase Tabla 31. 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. Este panorama reafirma la pertinencia de integrar en la evaluación educativa un procedimiento integral para mejorar la calidad de los argumentos de validez y orientar acciones concretas que fortalezcan las inferencias menos robustas.

Tabla 31

Comparación sobre las escalas de GRADE, CERQual y Escala EBA (propuesta)

Aspecto	GRADE	CERQual	EBA (propuesta)
Contexto de aplicación	Evaluaciones clínicas y formulación de recomendaciones (p. ej., guías de práctica clínica).	Revisión y síntesis de evidencia cualitativa (p. ej., política pública, ciencias sociales).	Evaluación de la validez de inferencias en pruebas educativas, con base en los Estándares (AERA et al., 2014).
Niveles	Alta, Moderada, Baja, Muy baja	Alta, Moderada, Baja, Muy baja	Alta, Moderada, Baja, Muy baja

Criterios de clasificación	- Diseño del estudio (aleatorizado u observacional). - Riesgo de sesgo. - Consistencia/inconsistencia. - Directriz/indirectriz. - Precisión. - Sesgo de publicación. (Pueden incrementar o reducir el nivel de evidencia.)	- Relevancia del contexto. - Coherencia metodológica. - Suficiencia de datos. - Solidez conceptual. (Cada criterio puede afectar la confianza final en la evidencia.)	- Claridad de la inferencia. - Coherencia (relación entre supuestos, evidencias y conclusiones). - Plausibilidad (credibilidad de la relación entre evidencias y uso del puntaje). (Se suman valores de 1 a 4 en cada dimensión para un puntaje global.)
Interpretación general	- Alta: Muy alta confianza en el efecto estimado. - Moderada: Confianza razonable; podrían existir ajustes. - Baja: Incertidumbre relevante. - Muy baja: Escasa confianza, evidencias inconsistentes.	- Alta: Elevada confianza en que el hallazgo cualitativo refleja la realidad. - Moderada: Se requiere evidencia adicional, pero el hallazgo es plausible. - Baja: Varios vacíos o inconsistencias. - Muy baja: Falta de solidez y coherencia.	- Alta: Evidencia robusta, sin contradicciones relevantes. - Moderada: Evidencia sólida con algunos puntos mejorables. - Baja: Evidencia limitada o con vacíos notables. - Muy baja: Evidencia muy escasa o que contradice la inferencia.
Uso principal	Apoyar la toma de decisiones en salud, balanceando riesgos/beneficios y coste-efectividad.	Brindar confianza sobre hallazgos cualitativos en revisiones sistemáticas.	Fundamentar la validez de cada inferencia (p. ej., Definición de Dominio, evaluación, Generalización, etc.) en pruebas educativas.

Nota. Elaboración propia basada en "GRADE Guidelines: 1. Introduction—GRADE Evidence Profiles and Summary of Findings Tables" (G. Guyatt et al., 2011), "Applying GRADE-CERQual to Qualitative Evidence Synthesis Findings: Introduction to the Series" (S. Lewin et al., 2018), "Content-Related Validity Evidence in Test Development" (M. Kane, 2006, en S. M. Downing & T. M. Haladyna [Eds.], *Handbook of Test Development*, pp. 131-153) y *Argument-Based Validation in Testing and Assessment* (C. A. Chapelle, 2021).

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. Si bien son guías, como se mencionó en el *Marco Teórico*, no son correspondencias estrechas, pero nos da un margen para expresar este ejercicio sobe el proceso de validación del ExIES, y que también funcione para futuras investigaciones. 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 32 detalla dicha vinculación, acompañada de una breve justificación fundamentada a través de Chapelle (2021) y los Estándares (AERA et al., 2014).

Tabla 32

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 33 presenta una plantilla diseñada (producto) para expresar la evaluación de cada inferencia según los tres criterios especificados (claridad, coherencia y plausibilidad).

Tabla 33

Estructura de las Escalas para le 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 33, 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 34.

Tabla 34

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 34. Por ende, con el fin de interpretar el LV global (LVg), en la Tabla 35 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 (Kane, 2013; Chapelle, 2021).

Tabla 35

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 36 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 36

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 si no para los tomadores de decisiones. Además, es importante aclarar que el producto final de esta etapa son las recomendaciones sintetizadas posterior a la evaluación mediante la escala propuesta.

Consideraciones éticas

- Confidencialidad institucional: 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.
- Transparencia en la autoría y uso de información: Se citaron las fuentes con apego a las Normas APA (2020), respetando derechos de autor y propiedad intelectual.
- Evitar conflicto de intereses: La investigadora es externa al ExIES, asegurando independencia en la evaluación. Jornet et al. (2010), señala que un posible conflicto de interés es ser juez y parte en los procesos de evaluación, con lo cual se separaron explícitamente los roles de evaluación y de gestión.

- Lineamientos: Se siguieron las consideraciones y criterios de los Estándares de la AERA et al. (2014) para el proceso de validación de pruebas. Asimismo, se atendieron criterios de confiabilidad, equidad y uso adecuado de los puntajes, dejando constancia documental de procedimientos y resultados.

Resultados

En este apartado se presentan los resultados del ExIES y responde de forma directa a los siete objetivos específicos de la investigación (véase Tabla 1): (1) valorar la congruencia entre el contenido del ExIES y el dominio curricular; (2) verificar la alineación de los procedimientos de aplicación con los estándares técnicos; (3) estimar la confiabilidad y estabilidad de las puntuaciones; (4) examinar la estructura interna y las correlaciones entre áreas conforme al modelo teórico; (5) reunir evidencia predictiva respecto del desempeño universitario; (6) revisar el uso institucional de los puntajes en procesos de admisión; y (7) analizar las consecuencias de las decisiones desde enfoques de equidad y efectividad; y, por ende, alinearse al objetivo general.

La descripción se organiza en nueve apartados. El primero introduce el AIU del ExIES como mapa que articula las inferencias del estudio, y los ocho apartados restantes desarrollan, en el orden lógico del AIU, las inferencias que sostienen la validez de las interpretaciones y usos del examen: (2) Definición de dominio, (3) Evaluación, (4) Generalización, (5) Explicación, (6) Extrapolación, (7) Uso de los puntajes en admisión, (8) Consecuencias y (9) la Valoración del Argumento Global. En las siguientes secciones se detalla cada inferencia siguiendo el procedimiento propuesto en el *Diseño Metodológico*. Cada inferencia contiene: a) una introducción a la inferencia, su alineación a los Estándares (AERA et al., 2014), y algunos antecedentes; b) la definición de las garantías, supuestos y fuentes; c) los respaldos específicos vinculados a las garantías y supuestos, es decir, la descripción de los resultados por supuesto; d) la evaluación (claridad, coherencia y plausibilidad), y recomendaciones, más un resumen por cada supuesto contemplando la evidencia, recomendaciones y reservas; y, e) comentarios finales con relación al *Marco Teórico* y los *Antecedentes* de la presente investigación.

Argumento de Interpretación y Uso

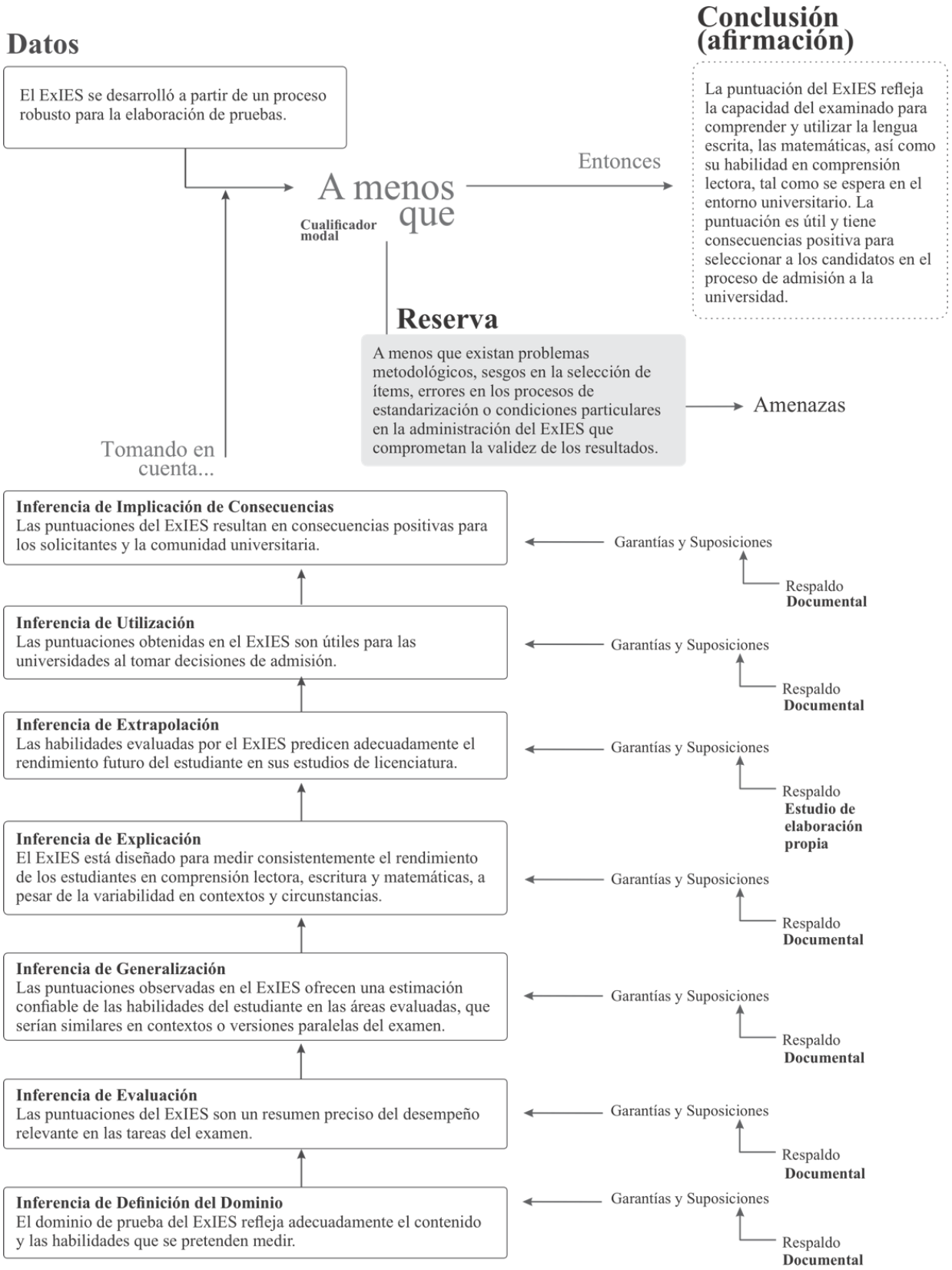
El primer paso, a partir del EBA, fue la elaboración de la afirmación o conclusión a manera de borrador. Es decir, describir con claridad la conclusión del argumento pues es lo que define cómo se interpretarán los puntajes. En el caso del ExIES lo que se pretende evidenciar es la siguiente afirmación (véase Figura 26):

> 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.

Con el fin de organizar de forma lógica los argumentos que dan pie a la conclusión, la Figura 26 exhibe el AIU; el cual es un primer borrador que expresa de forma escalonada las siete inferencias que conformarán el Argumento de Validez del ExIES. La secuencia lógica del AIU parte de verificar que el examen cubre el dominio pertinente; continúa demostrando que sus puntajes resumen con precisión ese desempeño, se mantienen consistentes en contextos paralelos y reflejan las habilidades realmente evaluadas. Sobre esta base, se evidencia que las puntuaciones predicen el rendimiento universitario, sirven como criterio confiable de admisión y, finalmente, generan consecuencias positivas tanto para los aspirantes como para la comunidad académica. Por otra parte, la tabla del Apéndice B detalla para cada inferencia sus garantías, supuestos y fuentes de datos para lograr visualizar el argumento completo; esta tabla se dividió según las inferencias y se expresan en las siguientes secciones de resultados.

Figura 26

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; C. A. Chapelle, 2021, SAGE Publications).

Inferencia de Definición de Dominio

Según los Estándares (AERA et al., 2014)—particularmente los estándares 1.0, 1.2, 4.1, 11.13 y 11.14—, para asegurar dicha validez es imprescindible delimitar con claridad el constructo evaluado y justificar adecuadamente la selección del contenido en relación con el desempeño académico esperado. Por tanto, el objetivo que se responde en esta sección es sobre la inferencia de Definición de Dominio. Ésta evalúa específicamente hasta qué punto los ítems del ExIES cubren de forma suficiente y representativa los contenidos y habilidades esenciales para ingresar a la educación superior. Por ende, se busca confirmar la adecuada alineación del instrumento con los marcos formativos nacionales e internacionales relevantes. Considerando lo anterior, a continuación, se explicitan las garantías propuestas para esta inferencia.

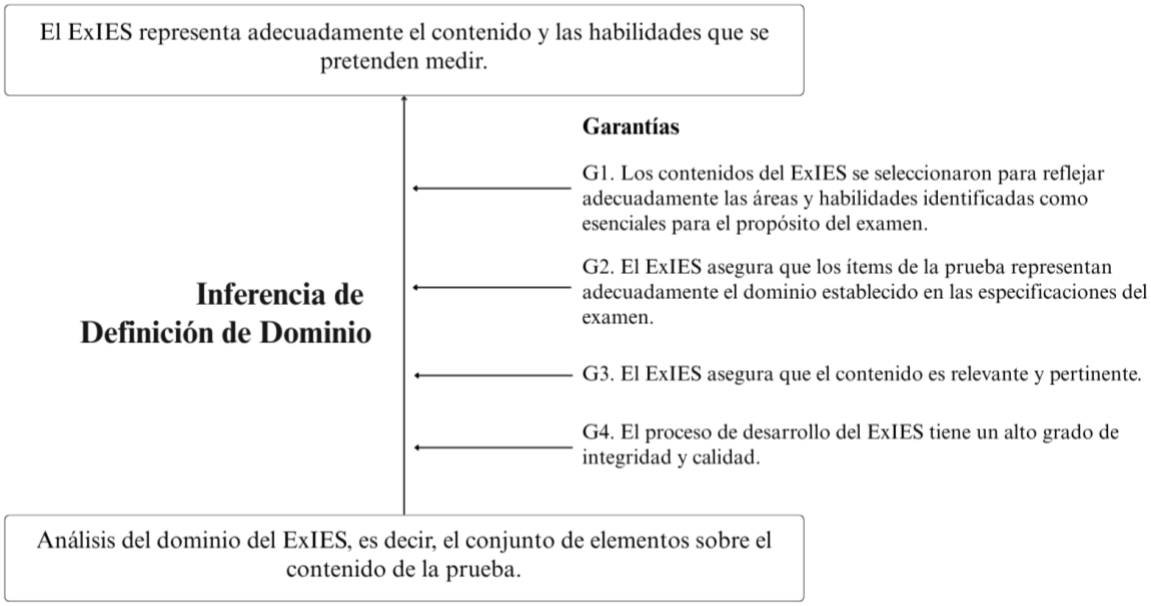
Definición de las garantías, supuestos y fuentes de la inferencia de Definición de Dominio

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 esta afirmación, se contemplan cuatro garantías; véase la Figura 27. La Garantía 1.1 (adecuación de los contenidos) afirma que el examen cubre las áreas, y habilidades esenciales de la EMS, mediante tres supuestos: concordancia del contenido, cumplimiento de las especificaciones y un proceso de revisión y actualización permanente; la Garantía 1.2 (representación del dominio en los ítems) profundiza en la correspondencia ítem–especificación señalando que el diseño respeta proporciones de habilidades y niveles cognitivos, que existe una revisión para afinar los ítems y que opera un sistema continuo de ajuste; la Garantía 1.3 (relevancia y pertinencia) exige que los ítems sean sometidos a revisiones internas y externas y que se elimine material irrelevante o sesgado, evidencia que queda registrada en juicios expertos y manuales (R.3.1, R.3.2); y, por último, la Garantía 1.4 (integridad y calidad en el desarrollo)

recoge el tema organizativo: equipos cualificados y capacitados, procesos rigurosos con revisiones múltiples y control sistemático mediante análisis de calidad de los ítems.

Figura 27

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 37, 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.

Tabla 37

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 le Evaluación de ítems del Nuevo Examen de Selección de aspirantes a ingresar a la Universidad Autónoma de Baja California (Caso & 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 & Díaz, 2016). En la Tabla 38 se compara el enfoque disciplinar de la EMS con el Proyecto DeSeCo y la evaluación PISA.

Tabla 38

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

Nota. Elaboración propia basada en Caso et al. (2017) y, Caso y Díaz (2016).

Como se mencionó, el ExIES parte de tres áreas (Pedroza Zúñiga et al., 2024a; Caso et al., 2017; Caso & Díaz, 2016). Éstas se definen a continuación:

- 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.

  En este sentido,

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 (Pedroza Zúñiga et al., 2024a; Caso et al., 2017; Caso & Díaz, 2016):

- 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 le Evaluación de ítems (Caso & 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). En la Tabla 23, como ya se ha mencionado en el apartado de *Metodología,* se pueden observar las especificaciones de las tres áreas. El área de Lectura se centra en la evaluación del contenido, el análisis discursivo y la síntesis de diversas fuentes. La sección de escritura revisa el desarrollo temático y la adecuación a las normas gramaticales del español. Por su parte, en Matemáticas, se abordan aspectos como las ecuaciones lineales, análisis de datos, funciones exponenciales y conceptos avanzados relacionados con el área y el volumen.

Asimismo, en la Tabla 24 se detallan los NDC del ExIES, los cuales se basan en la Taxonomía de Bloom (Bloom et al., 1956); aunque no hay una citación o referencia a la taxonomía, es una deducción a partir de la tabla. El NDC de comprensión implica entender el significado del material, con un 8% total. Aplicación, con el 46% total, se refiere a usar el material en situaciones nuevas. El análisis representa el 5% y trata sobre descomponer el material. La síntesis tiene el 11% y se enfoca en integrar partes para crear un conjunto. Por último, la evaluación, que es el 30% total, implica juzgar el valor del material.

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 39). 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 39

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. Como se observa en el Apéndice D*,* estos cambios son evidentes en los subcontenidos de las áreas de Lectura, Lengua Escrita y, especialmente, Matemáticas, que fue el área con mayores modificaciones durante el proceso. En la Tabla 40 se resumieron estos cambios, es decir el número de ítems que tuvieron cambios, que evidencia el cumplimiento del seguimiento a los cambios en las especificaciones.

Tabla 40

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 aumenta 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).

Cada conjunto de contenidos de los manuales está adaptado a su área específica, con particularidades que reflejan las habilidades y conocimientos necesarios para Lengua Escrita, Matemáticas y Lectura. Las diferencias entre los conjuntos de contenido parecen estar principalmente en los detalles específicos y en la aplicación práctica de los principios generales (Pedroza Zúñiga et al., 2023a, 2023b, 2023c). Las homologaciones se pueden ver en:

- Proceso de diseño, construcción y validación: Todos siguen el mismo marco metodológico para asegurar la validez del examen.
- Desarrollo de ítems: Aunque con contenido diferente, la estructura para el desarrollo de ítems se mantiene.
- Demanda cognitiva: Todos evalúan la demanda cognitiva, aunque los detalles específicos varían según el área de conocimiento.
- Tipos de ítems y pasos para su elaboración: Siguen una metodología similar con adaptaciones específicas para el área de contenido respectiva.
- Evaluación y entrega de ítems: Tienen procesos similares para la evaluación y entrega de ítems desarrollados.

Por otro lado, también se cuentan con especificaciones por componente del ExIES. En la Tabla 41 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 41

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, etc.). 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).

De esta manera, las especificaciones recogen todo el marco metodológico para la elaboración de ítems en cada área —desde la definición del contenido y nivel cognitivo, hasta la forma de redactar instrucciones y distractores— garantizando la alineación con las proporciones y competencias definidas en la tabla de especificaciones y en los manuales de elaboración de ítems (Pedroza Zúñiga et al., 2023a, 2023b, 2023c). Esto permite que cada ítem cumpla con los criterios de validez de contenido y la correcta representación de las habilidades y conocimientos que se busca medir.

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, y

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 42); 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 42

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. Copyright 2023 por L. H. Pedroza Zúñiga et al.

Por lo anterior, se cuenta con una base de datos para el seguimiento de las subversiones, es una matriz de trazabilidad NDC–subcontenido–ítem (2024c), con el fin de guardar un control sobre los ítems de los que se compone cada forma y subversión: (a) proporciones NDC/subcontenido (tabla de especificaciones), (b) listado de ítems incluidos (ID ítem, subcontenido, NDC, posición en la forma), (c) conteo observado y proporción observada por NDC/subcontenido, (d) desviación respecto al objetivo y acción tomada (aceptar/editar/retirar/pilotear). Además, se cuenta con una base de datos por área sobre los resultados del análisis Rasch y estadísticas según ítem-forma (2023s), lo que da la posibilidad de mejorar el seguimiento y descarte en posteriores procesos de jueceo.

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 43 que ejemplifica lo realizado en el componente Lectura.

Tabla 43

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 43, 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 tienen 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.

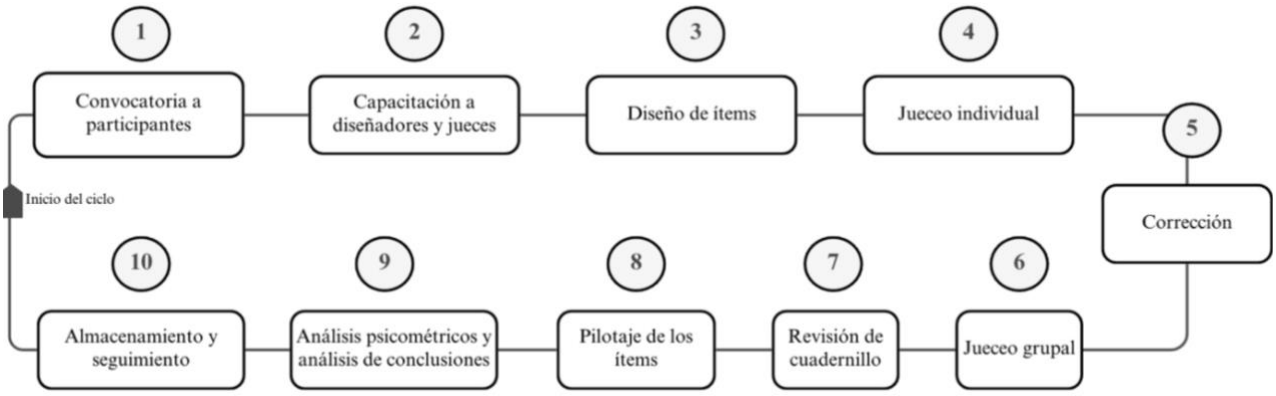
A partir de dichos resultados, se implementa un proceso de retroalimentación continua que identifica a los diseñadores que requieren apoyo adicional en su capacitación o, en casos

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 28, 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 28

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 & 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 44 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.

En el primer supuesto (S1.1.1, 83.3%), el constructo y las áreas del ExIES se encuentran documentados en el Manual técnico y la Guía de evaluación de ítems, con apoyo en competencias de EMS y marcos internacionales como PISA/DeSeCo. El puntaje 10/12 refleja coherencia y plausibilidad altas, pero cierta debilidad en claridad, pues no se explicita el mapeo dominio–subcontenido–NDC ni su relación con los usos previstos del puntaje, lo que limitaría la alineación plena con los estándares 1.0–1.2 (AERA et al., 2014; Sireci & Faulkner-Bond, 2014). El segundo supuesto (S1.1.2, 100%) alcanzó la calificación máxima gracias a especificaciones completas y trazadas a múltiples fuentes, en cumplimiento del estándar 4.1 (AERA et al., 2014; Lane et al., 2016). En contraste, el tercero (S1.1.3, 75%) obtuvo 9/12, dado que, aunque existen ciclos anuales de revisión documentados, no se incluye evidencia explícita sobre la transición normativa hacia el MCCEMS/NEM (Acuerdo 09/08/23), lo que reduce la plausibilidad. Esta situación podría resolverse con tablas de concordancias y actas de consulta con expertos (AERA et al., 2014; Sireci & Faulkner-Bond, 2014).

Tabla 44

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	11 (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	11 (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	10 (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	11 (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	11 (91.7%, Alta)
	Global	37 (84.1%, Moderada)	37 (84.1%, Moderada)	38 (86.4%, Alta)	112/ 132 (84.8%, Moderada)

En cuanto a la segunda garantía, el supuesto S1.2.1 (75%) recibió 9/12 porque no se reporta de manera sistemática la comparación entre las metas de especificación y lo efectivamente observado, lo que afecta la coherencia y plausibilidad. Una mejora consistiría en publicar balances de cumplimiento y bitácoras de ajustes (Haladyna & Rodriguez, 2013; Sireci & Faulkner-Bond, 2014). El supuesto S1.2.2 (75%) presenta una situación semejante: aunque se describe el proceso de revisión de ítems, faltan criterios psicométricos explícitos, como umbrales de misfit o reglas de decisión ante hallazgos de DIF, lo que limita su plausibilidad (AERA et al., 2014; Bond & Fox, 2015; Holland & Thayer, 1988). Algo similar ocurre con el supuesto S1.2.3 (75%), pues, si bien se evidencia un proceso iterativo de diseño, jueceo, pilotaje y análisis con trazabilidad al NDC, no se dispone de reportes periódicos que documenten desviaciones y ajustes, lo que justificaría una valoración intermedia. Incorporar reportes sistemáticos ayudaría a reforzar la transparencia, como recomiendan Mislevy et al. (2003, 2004).

La tercera garantía muestra puntajes más altos. El supuesto S1.3.1 (91.7%) refleja un proceso robusto de revisión interna y externa, con eliminación de ítems irrelevantes; sin

embargo, la claridad se limita al no ofrecer ejemplos documentados de ítems rechazados o ajustados, lo que explica la pérdida de un punto. El supuesto S1.3.2 (91.7%) confirma la existencia de procesos preventivos contra sesgos y se refuerza con análisis de DIF, aunque también aquí se detecta falta de claridad, que podría mejorarse mediante un manual breve de sesgos frecuentes con ejemplos del ExIES (AERA et al., 2014).

Finalmente, la cuarta garantía presenta resultados consistentes. El supuesto S1.4.1 (83.3 %) demuestra que los desarrolladores de ítems reciben capacitación, pero la plausibilidad se ve limitada al no existir indicadores de desempeño individuales; por ello, un tablero con métricas de aceptación y calidad reforzaría la evidencia (Lane et al., 2016; Haladyna & Rodriguez, 2013). El supuesto S1.4.2 (91.7%) acredita un proceso riguroso de elaboración, jueceo, pilotaje y análisis, documentado conforme a los estándares 4.1 y 11.13; sin embargo, podría ampliarse con la incorporación explícita del marco ECD como narrativa de preguntas y evidencias (Mislevy et al., 2003, 2004). El supuesto S1.4.3 (91.7%) respalda el uso de análisis avanzados como TRI, confiabilidad y equiparación, aunque carece de umbrales y reglas de decisión publicadas, lo que restringe la plausibilidad (AERA et al., 2014; Bond & Fox, 2015; Kolen & Brennan, 2014).

Inferencia de Evaluación

La valoración de la inferencia de Evaluación en el ExIES se centra en demostrar que las puntuaciones obtenidas reflejan, de manera válida y confiable, el nivel real de desempeño de los sustentantes (Kane, 2006, 2013; Chapelle, 2021). Para sustentar este proceso, se toman en cuenta los Estándares, específicamente el 6.1, 6.2, 6.4, 6.5, 7.2, 8.1, 8.2 y 9.3 establecidos por la AERA, la APA y el NCME (2014), los cuales garantizan la estandarización en la aplicación, la equidad en el tratamiento de los sustentantes, la confidencialidad en el manejo de resultados y la claridad en la interpretación de las puntuaciones.

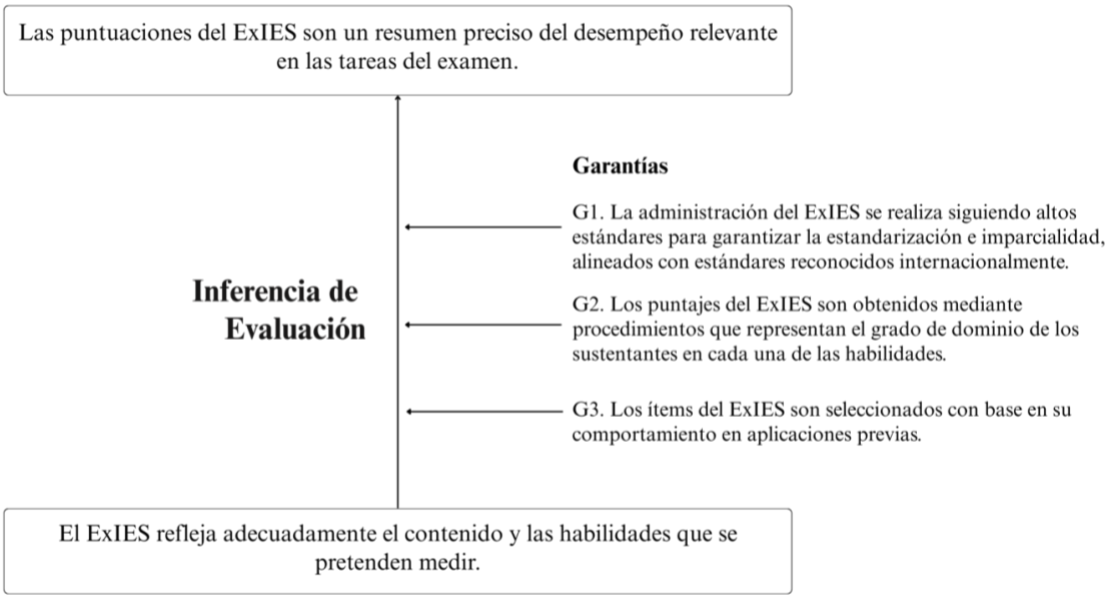
En concreto y en consonancia con los Estándares (AERA, et al., 2014), esta inferencia implica capacitar al personal de aplicación, ofrecer materiales de práctica a los aspirantes y establecer protocolos de seguridad e instrucciones claras contra la deshonestidad académica. Tomando lo anterior en consideración, a continuación, se describen las garantías pertinentes.

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

En la Figura 29 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 examen; sustentado por la premisa base de que este refleja adecuadamente las habilidades y contenidos que pretende medir.

Figura 29

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 45 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 45

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	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

	condiciones de aplicación.	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 basado en sus respuestas. S2.2.2 Se realiza un procedimiento operativo después de la estimación Rasch por cada área.	F2.2.1.1 Técnica Rasch, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) 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 46 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 46

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 47 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 47

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 & Rodriguez, 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 48, 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 48

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 post-examen	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, sección 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). Éste 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. Se enfatiza la importancia de seguir protocolos de seguridad y estandarización para garantizar la integridad, validez y equidad en la evaluación, documentando además los principales retos y eventos atípicos detectados en el proceso

Según la Tabla 49, 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. También se reportaron incidentes aislados como intentos de soborno, daños a los materiales de aplicación, conductas inapropiadas y dificultades para controlar la presencia de familiares en las sedes.

Tabla 49

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 50, 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. Además, se especifican los procedimientos para detectar y reportar conductas sospechosas, y se establece que, antes de iniciar el examen, los aplicadores y supervisores deben recordar a los sustentantes las reglas y sanciones, reforzando así la transparencia y la integridad del proceso de evaluación.

Tabla 50

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. Sin embargo, esta sección también tiene áreas de mejora.

Una de las principales debilidades —de la guía— es la falta de énfasis en las consecuencias adicionales, ya que, aunque se menciona la cancelación del examen, no se detallan otras posibles sanciones, como la prohibición de volver a presentar el examen en futuras aplicaciones. Además, el formato del texto podría mejorar en cuanto a visibilidad y accesibilidad, ya que la información sobre las consecuencias está dispersa en el texto. Reorganizar esta información en un formato más claro, con un título dedicado a las consecuencias de actos deshonestos, haría que el mensaje sea más concreto y fácil de recordar.

Por otra parte, sí se explicitan estas prohibiciones en el manual del aplicador y en la presentación utilizada en la capacitación para la aplicación y supervisión del ExIES (Pedroza Zúñiga et al.,2023c, 2023k), éstas son necesarias para volver a repasar de forma oral las sanciones y repercusiones sobre comportamientos deshonestos. Sin embargo, aunque el ExIES asegura instrucciones claras, la falta de detalles sobre la implementación efectiva de estas estrategias de comunicación sugiere que el sistema podría beneficiarse de un enfoque más

estructurado, es decir, sería ideal establecer algún diagrama de flujo o tabla que indique los medios de información antes del examen, ya que durante la aplicación sí quedan claras las sanciones propuestas. En este sentido, es relevante que se implemente una retroalimentación formal, como encuestas post-examen, para confirmar si los candidatos comprendieron las instrucciones y las consecuencias de deshonestidad.

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 51, 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. Además, los protocolos de seguridad garantizan que se mantenga un ambiente controlado y que todos los sustentantes enfrenten las mismas condiciones sin interrupciones ni distracciones en cada una de las sedes.

Tabla 51

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 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 52) 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. Por lo que podría ser importante seguir asegurando la presencia de personal calificado en todas las sedes, incorporar recursos audiovisuales para la instrucción de sustentantes y establecer estrategias para controlar el acceso de familiares en las instalaciones.

Tabla 52

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).

La implementación de un proceso estandarizado en la administración del ExIES presenta varias fortalezas. En primer lugar, los manuales proporcionan instrucciones claras y detalladas para preparar las aulas, distribuir los materiales y controlar el tiempo, asegurando que todos los

sustentantes enfrenten las mismas condiciones. Esto garantiza que los puntajes sean comparables y que las diferencias reflejen las habilidades de los candidatos y no factores externos. Además, los protocolos de seguridad descritos en los manuales aseguran que el ambiente del examen esté controlado, minimizando la posibilidad de distracciones o interrupciones. Las áreas de mejora, ya mencionadas, son importantes además de que transparenta su seguimiento.

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 basado 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 & 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 53 se muestra la descripción de las evidencias psicométricas para esta suposición.

Tabla 53

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 57 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 ($\bar{p}$) —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.

También, la Tabla 57, expresa las 11 subversiones de la prueba, los cuales son resultado de diferentes combinaciones de ítems según las mejores métricas. Se emplearon 200 ítems de Lectura, 200 ítems de Lengua Escrita y 188 del área de Matemáticas. Así, se observa que la tasa de acierto promedio es más alta en el área de Lectura, con un 54.5%, lo que indica que los participantes obtuvieron más respuestas correctas en esta área en comparación con Lengua Escrita (49.86%) y Matemáticas (37.36%). Esta última área muestra la tasa de acierto más baja, lo cual sugiere que los ítems de Matemáticas fueron percibidos como más difíciles por los examinados. En términos de dificultad media, los ítems de Matemáticas tienen el valor más alto (.58), seguido por Lengua Escrita (.54) y Lectura (.51), lo que refuerza la observación de que los ítems de Matemáticas fueron los más complejos. Se recomienda aplicar un modelo logístico de dos parámetros (2 PL) para posteriores pruebas.

Tabla 54

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 55 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, depende 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 55 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 55

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).

Por otro lado, el Apéndice E es un ejemplo de cómo se presentan los resultados por ítem en cuanto a con Infit/Outfit MNSQ que se encuentran dentro o fuera del rango de ajuste

recomendado. La Tabla 55 muestra un listado que resume los ítems que, en cada subversión y cada área (Lectura, Lengua Escrita y Matemáticas), se encuentran fuera del rango para el modelo Rasch (Infit/Outfit MNSQ entre 0.70 y 1.30).

En función de lo anterior, el ExIES mantiene buenos resultados. Aun cuando existen algunos ítems cuyo ajuste podría refinarse (al exceder los rangos recomendados de Infit/Outfit MNSQ), el número de casos es reducido en relación con el total de ítems. Además, dichos ítems mantienen, por lo general, índices de discriminación por encima del mínimo deseable. Esto confirma que el modelo Rasch se está utilizando con eficacia para estimar la habilidad de los sustentantes, resultando en puntajes que reflejan de manera válida su nivel de desempeño en cada área y versión evaluada.

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:

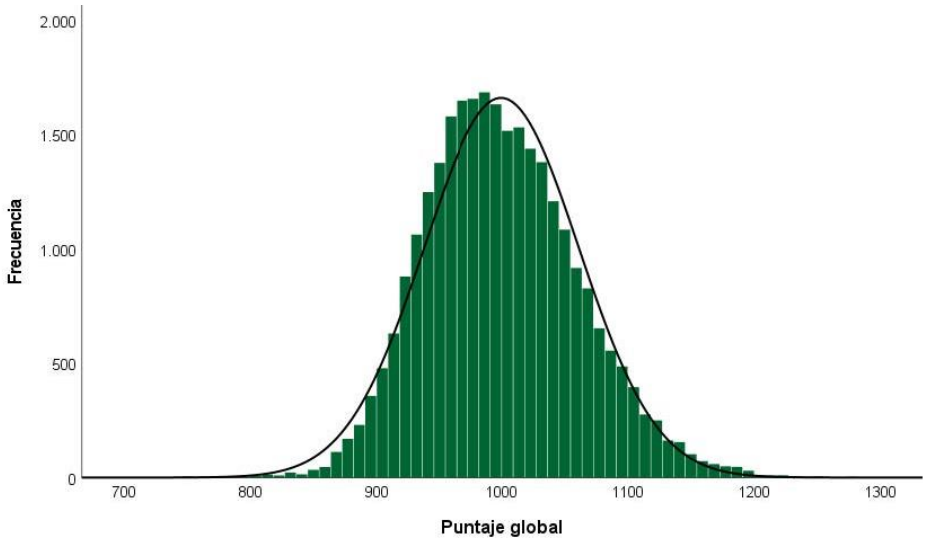
a. Depuración previa de reactivos: se excluyen del cómputo los ítems que muestran mal funcionamiento (p. ej., correlación punto-biserial negativa o desajustes fuera de criterios), debido a que degradan la consistencia interna y agregan ruido a la medición; antes de retirarlos se realiza revisión experta 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 inter-cohortes.
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 30.
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 30

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 & Fox, 2015; Kolen & Brennan, 2014; Tavakol & Dennick, 2011; Haladyna & Rodriguez, 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 detallado 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).

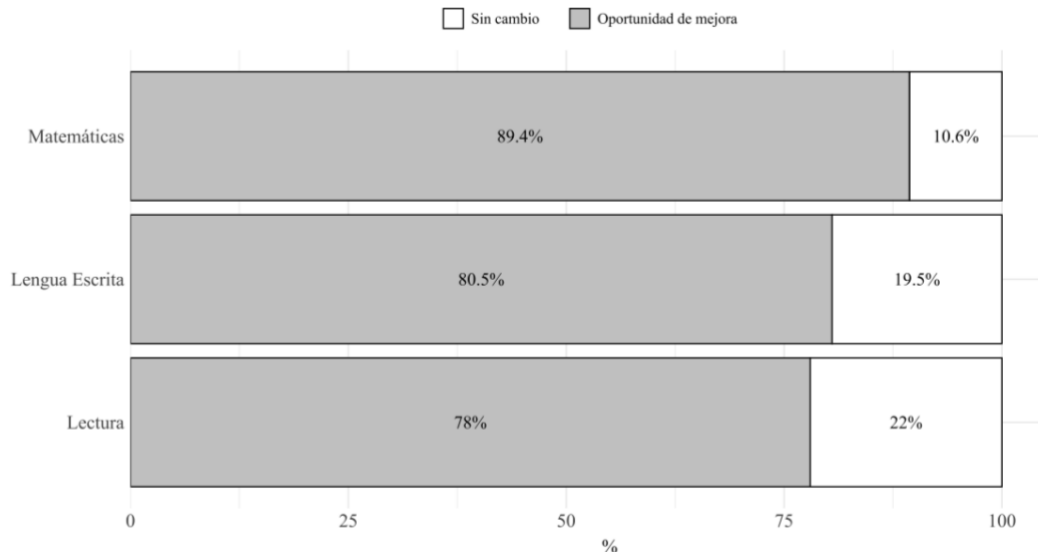
Los ítems que fueron sometidos a pilotaje en versiones previas, en este caso el pilotaje 2022-2, se reevaluaron para confirmar su comportamiento psicométrico en nuevas aplicaciones. Este procedimiento garantiza que los ítems seleccionados mantengan una coherencia en su dificultad y discriminación entre diferentes cohortes de sustentantes; véase Apéndice F. Además, se emplean ítems ancla y criterios psicométricos adicionales para asegurar la comparabilidad

entre diferentes versiones del examen. De este modo, se garantiza que las diferencias en los puntajes obtenidos por los sustentantes reflejan de manera precisa sus habilidades, eliminando posibles sesgos o inconsistencias derivadas de la variabilidad en la dificultad de los ítems.

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 31): un 22% en Lectura, 19.5% en Lengua Escrita y 10.6% en Matemáticas (Pedroza Zúñiga et al., 2024a).

Figura 31

Distribución porcentual de ítems con oportunidad de mejora, por área, en el ExIES 2023-1



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

Evaluación de la inferencia de Evaluación

La valoración global de la inferencia de Evaluación, presentada en la Tabla 56, 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. La primera, vinculada con la administración del examen, integra cinco supuestos. En el S2.1.1 (91.7%), la formación del personal se encuentra normada y respaldada por manuales de aplicadores y supervisores, así como por materiales de capacitación que detallan funciones, listas de verificación y cadena de custodia. Esta evidencia asegura la estandarización y la imparcialidad del proceso, en concordancia con los Estándares 6.1 y 6.5 (AERA, APA & NCME, 2014; Lane et al., 2016). El puntaje de 11/12 refleja un nivel alto de claridad y plausibilidad, con una coherencia evaluada como moderada.

En el S2.1.2 (83.3%), la Guía del Sustentante ofrece ejemplos de reactivos y orientaciones para la gestión del tiempo, lo que promueve la equidad informativa y la familiarización con las condiciones de la prueba, conforme a los Estándares 8.1 y 8.2 (AERA et al., 2014; Pedroza Zúñiga et al., 2023l). El puntaje de 10/12 proviene de la coherencia alta (4), mientras que la claridad y la plausibilidad se valoraron como moderadas (3 y 3).

Tabla 56

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)

El S2.1.3 (83.3%) documenta la aplicación de procedimientos de seguridad, tales como el resguardo de materiales y la cancelación de intentos de fraude, lo que se alinea con los Estándares 6.4, 7.2 y 9.3 (AERA et al., 2014; Haladyna & Rodriguez, 2013). Su calificación de 10/12 refleja una claridad alta (4) y coherencia y plausibilidad moderadas (3 y 3), debido a la variabilidad reportada en las sedes de aplicación.

En el S2.1.4 (83.3%), las instrucciones sobre deshonestidad académica se comunican en la Guía del Sustentante y los manuales de aplicación, además de reforzarse en la fase operativa, en consonancia con el Estándar 8.2 (AERA et al., 2014). La puntuación de 10/12 se explica por la coherencia alta (4) y por niveles moderados de claridad y plausibilidad (3 y 3). Asimismo, el S2.1.5 (83.3%) refleja la estandarización de condiciones de administración, como la preparación

de salones, el control de tiempos y la distribución y recuperación de materiales, lo cual garantiza igualdad de oportunidades, de acuerdo con los Estándares 6.1 y 6.2 (AERA et al., 2014; Lane et al., 2016). El puntaje de 10/12 corresponde a una plausibilidad alta (4) y a claridad y coherencia moderadas (3 y 3).

En relación con la segunda garantía, vinculada con los procedimientos de obtención de puntajes, el S2.2.1 (100 %) confirma que los resultados se estiman mediante el modelo Rasch, lo que permite controlar las características de los ítems y obtener una medida válida de la habilidad de los sustentantes, en consonancia con los Estándares 5.1 y 5.2 (AERA et al., 2014; Chapelle, 2021). Aunque recibió la máxima puntuación global (12/12), la claridad y la plausibilidad fueron valoradas como moderadas (3 y 3), debido a que la documentación técnica no detalla exhaustivamente los fundamentos teóricos ni los procedimientos operativos.

Finalmente, la tercera garantía se centra en la selección de ítems. En el S2.3.1 (83.3%), se constata la existencia de un proceso de pilotaje previo para analizar la dificultad y el ajuste psicométrico de cada ítem antes de integrarlo en el examen operativo. Este procedimiento, en línea con el Estándar 4.1 (AERA et al., 2014), garantiza que los ítems incluidos sean válidos y pertinentes. El puntaje de 10/12 refleja una coherencia alta (4), mientras que la claridad y la plausibilidad se mantuvieron en un nivel moderado (3 y 3).

Inferencia de Generalización

El objetivo que responde a esta sección es evaluar la confiabilidad y estabilidad de las puntuaciones del ExIES, mediante el análisis de su comportamiento psicométrico a lo largo del tiempo y entre diferentes formas del examen. La revisión se sustenta en los Estándares 4.10, 2.1, 6.1, 5.2, 7.1 y 4.1 (AERA et al., 2014), que exigen evidencia de estabilidad y equiparación formal entre formas alternas para sustentar decisiones educativas válidas y justas (Kane, 2006;

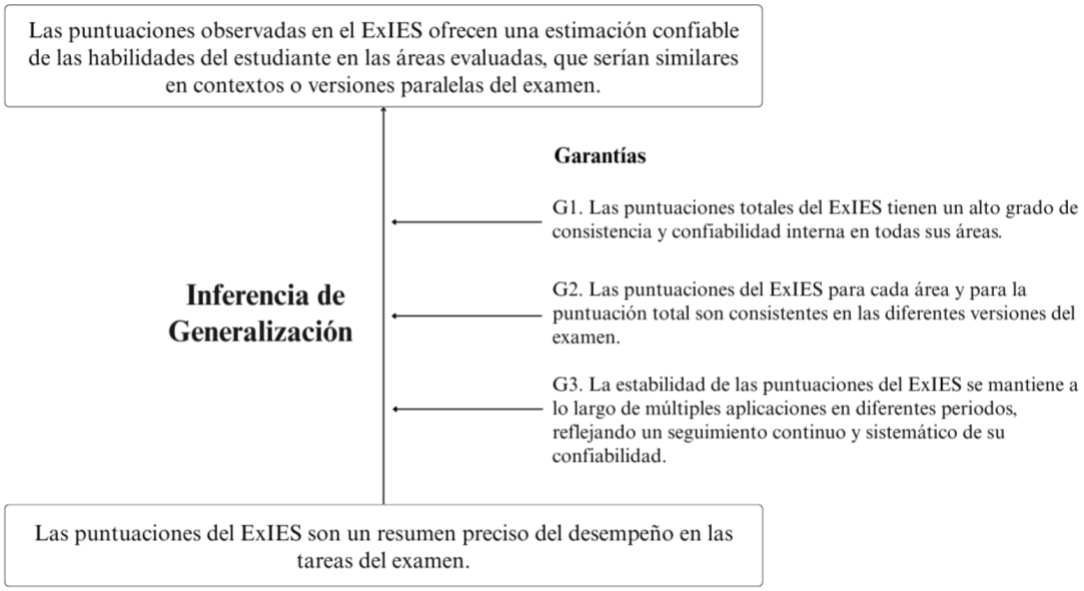
Chapelle, 2021).

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

La conclusión de Generalización, véase la Figura 32, en conjunto con sus garantías, busca asegurar que los puntajes obtenidos sean consistentes y confiables en diversas versiones y aplicaciones del examen. Estas garantías respaldan la estabilidad y la confiabilidad de las puntuaciones en las áreas evaluadas.

Figura 32

Argumento de la inferencia de Generalización como parte de la Validez del Argumento



La G3.1 (consistencia y confiabilidad interna) sostiene que las diferencias entre sustentantes reflejan rendimiento real y no errores de construcción o aplicación; su evidencia principal son coeficientes psicométricos aceptables (p. ej., alfa de Cronbach y otros indicadores). Asimismo, la G3.2 (equivalencia entre formas) exige que las distintas versiones del examen (Formas A, B, etc.) mantengan niveles de dificultad y propiedades psicométricas comparables, de modo que los puntajes sean equiparables entre aplicaciones. Y, la G3.3 (estabilidad temporal) requiere el monitoreo sistemático en múltiples aplicaciones y cohortes para verificar la

reproducibilidad de los coeficientes (supuesto 3.3.1) y detectar variaciones que demanden corrección, asegurando así la validez sostenida de la inferencia de generalización. Por otro lado, la Tabla 57 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 57

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 58 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 58

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 59 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 59

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	*AB*1	*AB*2	*AB*3	*AB*4	*AB*5	*A*6	*A*7	*A*8	*A*9	*A*10	*B*6	*B*7	*B*8	*B*9	*B*10
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 60 y en la Figura 33, 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 60

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

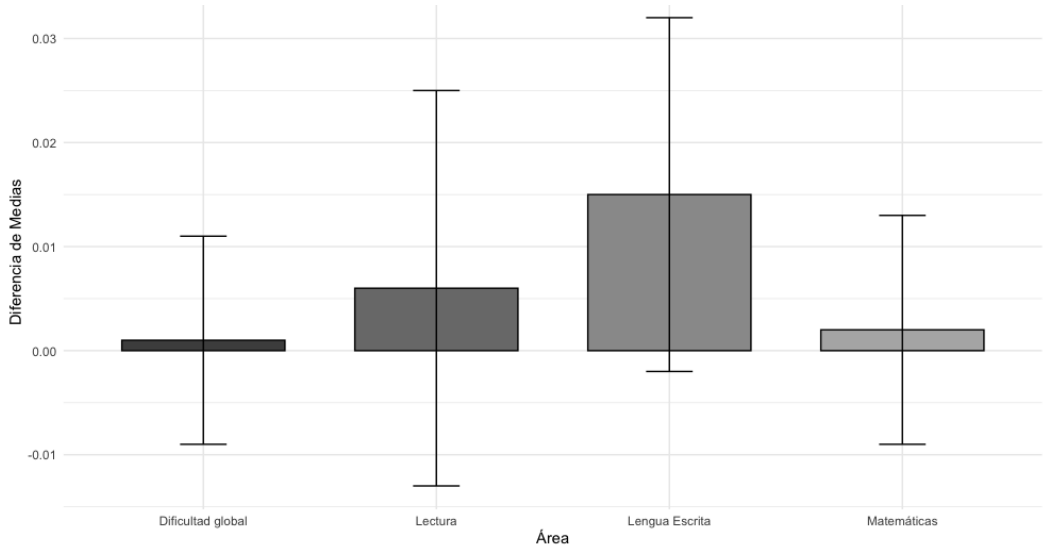
Área	*p* valor	Diferencia de medias	Diferencia de error estándar
Lectura	0.734	0.006	0.019
Lengua Escrita	0.380	0.015	0.017
Matemáticas	0.862	0.002	0.011
Dificultad global	0.864	0.001	0.010

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).

Este proceso asegura que los sustentantes enfrentan exámenes con niveles de dificultad similares, lo que contribuye a la imparcialidad y la validez de la inferencia de Generalización. La equiparación adecuada es esencial para garantizar que las versiones del examen mantengan la coherencia necesaria para hacer comparaciones justas entre los puntajes de los sustentantes, lo que fortalece la validez de las inferencias basadas en los resultados del ExIES.

Figura 33

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 61 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. Estos resultados refuerzan la validez de la equiparación en la mayoría de las áreas, aunque destaca una discrepancia en el área de Lengua Escrita, la cual podría requerir una revisión adicional (Kolen & Brennan, 2014). No obstante, en el Reporte Técnico 2023-1 se menciona que no se encontraron diferencias significativas.

Tabla 61

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 62 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 62

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 resultados 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 63 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. La primera garantía se refiere a la consistencia y confiabilidad interna de todas las áreas evaluadas. En el supuesto S3.1.1 (83.3%), la consistencia interna por área se mantiene dentro de los rangos aceptables: Lectura alcanza un coeficiente alfa de Cronbach de .73, Lengua Escrita de .77 y Matemáticas de .82, mientras que el puntaje global se ubica en .88 (Tabla 57). Estos valores se encuentran por encima del umbral mínimo de .70 recomendado para decisiones educativas grupales y cercanos al criterio de .80 que se considera deseable en contextos de mayor precisión (Nunnally & Bernstein, 1994; Tavakol & Dennick,

2011). El puntaje de 10/12 refleja claridad suficiente, aunque la coherencia y la plausibilidad fueron evaluadas como moderadas.

Tabla 63

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)

La segunda garantía, correspondiente a la equivalencia entre formas del examen, obtuvo resultados sólidos. El supuesto S3.2.1 (91.7%) documenta la equiparación de las versiones A y B del ExIES mediante calibración concurrente con ítems ancla y pruebas t, sin diferencias significativas en dificultad durante 2023-1. Este hallazgo confirma la comparabilidad entre formas y la validez de los puntajes derivados de ellas, en concordancia con los Estándares 4.10,

5.2 y 6.2 (AERA et al., 2014; Kolen & Brennan, 2014). La valoración de 11/12 (4-4-3) se explica por la consistencia metodológica y el respaldo empírico obtenido.

La tercera garantía aborda la estabilidad de la confiabilidad en distintas aplicaciones. En el supuesto S3.3.1 (91.7%), los coeficientes alfa de Cronbach se mantuvieron estables a lo largo de los periodos revisados, con valores globales entre .87 y .89 y alfas por área en rangos de .72 a .84 (Pedroza Zúñiga et al., 2024a, 2024b). Estos resultados confirman la reproducibilidad de la consistencia interna en diferentes cohortes y condiciones de aplicación, cumpliendo con los estándares que demandan estabilidad temporal de la medición (Nunnally & Bernstein, 1994; Tavakol & Dennick, 2011). La puntuación de 11/12 refleja que la claridad y la plausibilidad de la evidencia son altas, con una coherencia también favorable.

Inferencia de Explicación

La inferencia de Explicación en el ExIES se orienta a demostrar que las puntuaciones obtenidas reflejan adecuadamente las habilidades subyacentes que se pretende medir —Lectura, Lengua Escrita y Matemáticas— y que dichas habilidades se manifiestan como rasgos relativamente estables en los sustentantes. Los Estándares identificados para esta inferencia fueron el 1.0, 5.1, 1.1, 1.13, 1.16 y 3.2 (AERA et al., 2014), que exigen evidencia empírica sobre la correspondencia entre la estructura teórica y el comportamiento observable de los ítems. Dentro del EBA (Kane, 2013; Chapelle, 2021), la validez explicativa constituye un vínculo central entre lo que se mide y cómo se interpreta. Así, a continuación, se expresa la definición de las garantías para después evidenciarlas de esta inferencia.

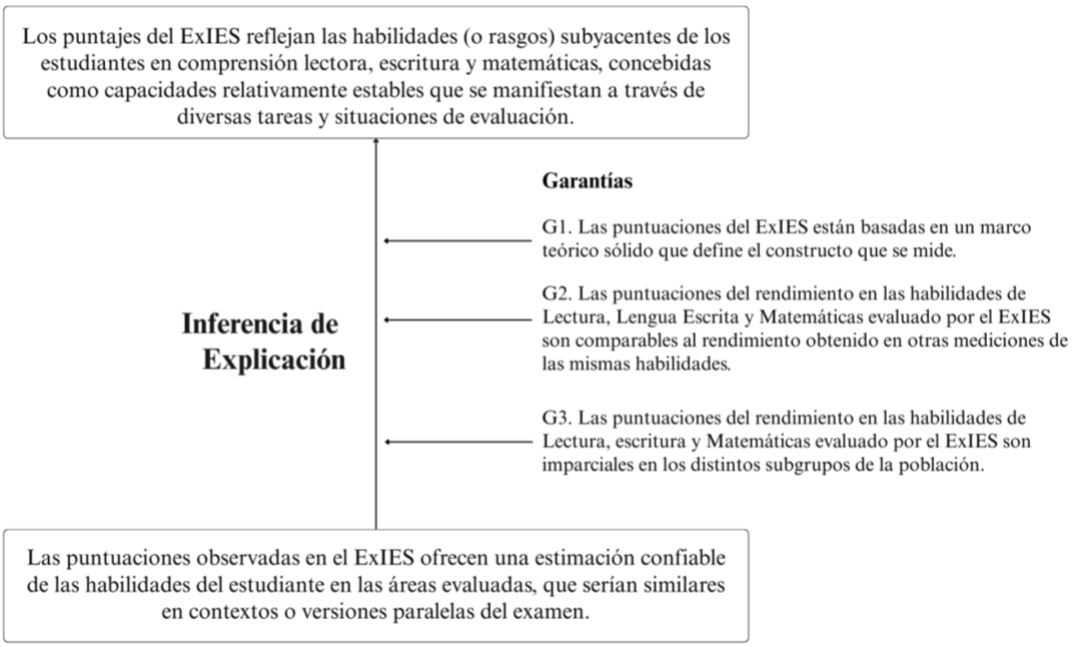
Definición de las garantías, supuestos y fuentes de la inferencia de Explicación

Con base en lo anterior, esta inferencia se estructura en tres frentes (véase la Figura 34). G4.1 (fundamento teórico) asegura que los ítems y áreas se apoyan en un marco conceptual claro

que define el constructo medido, de modo que cada ítem traduzca con precisión las capacidades pretendidas (sostenida por tres supuestos que vinculan tarea, contenido y proceso de respuesta).

Figura 34

Argumento de la inferencia de Explicación como parte de la Validez del Argumento



G4.2 (validez convergente) verifica que las puntuaciones del ExIES sean comparables con otras mediciones del mismo dominio —por ejemplo, EXANI II y el promedio de bachillerato— mediante el análisis de correlaciones y asociaciones que confirmen la convergencia entre indicadores académicos establecidos (S4.2.1–S4.2.3); y G4.3 (imparcialidad entre subgrupos) exige comprobar la ausencia de sesgo diferencial en los ítems a través de análisis DIF (S4.3.1), garantizando que las diferencias en puntajes respondan a las competencias evaluadas y no a condiciones socioeconómicas u otras variables contextuales.

Así, la Tabla 64 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.

Tabla 64

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 & 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 64, 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 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 lavann, 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 65) 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 65

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 66, 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, 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 66

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 67. En la Forma A, Lectura y Matemáticas concentran el mayor número de cargas fuertes según el criterio ≥ .35 (≈8 ítems cada una), mientras que Lengua Escrita presenta menos cargas fuertes (4) y una alta proporción de cargas negativas, lo que sugiere revisar reactivos

invertidos o su redacción. En la Forma B, Lengua Escrita mantiene buen desempeño global, pero con el nuevo umbral su patrón se desplaza hacia la moderación (4 fuertes y 17 moderadas) y continúa sin ítems no significativos ($p \geq .05 = 0$); en contraste, Lectura y Matemáticas registran 7 y 4–5 ítems no significativos, respectivamente. En conjunto, esto respalda la solución de tres factores, con focos claros de mantenimiento en los ítems con cargas débiles o negativas.

Tabla 67

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 evaluado 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 académicos.

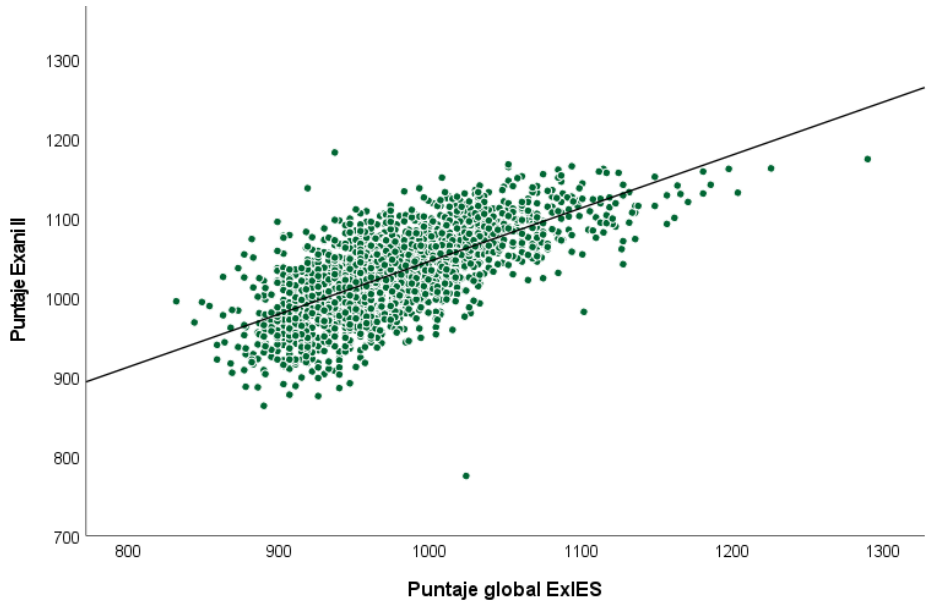
- Cálculo de correlaciones: Se empleó la función cor.test de R para obtener coeficientes de Pearson (r), con interpretación según Schober et al. (2018): r ≥ .70 indica correlación fuerte, r = .50–.69 moderada, y r = .30–.49 baja pero relevante para contextos educativos.
- Análisis adicional: Se exploró 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 35 y la Tabla 68, 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 35

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 68

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 & 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 (1,937). Se realizó una conversión a la regresión lineal elaborada. Y, como se observa en la Tabla 69**,** 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 69

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 basado en datos adaptados de *Estudio de validez concurrente con los puntajes de EXANI II* (Pedroza Zúñiga & 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 & 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 & Gómez Monárrez, 2025c). El procedimiento fue exhaustivo y combinó métodos clásicos y modernos, siguiendo criterios de robustez estadística y buenas prácticas internacionales (Zieky, 1993; García et al., 2016):

- 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): |logOR| < .43 (DIF leve), .43 ≤ |logOR| < .64 (moderado), |logOR| ≥ .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 70 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)| ≥ 0.43) y severo (|log (OR)| ≥ 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 70

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

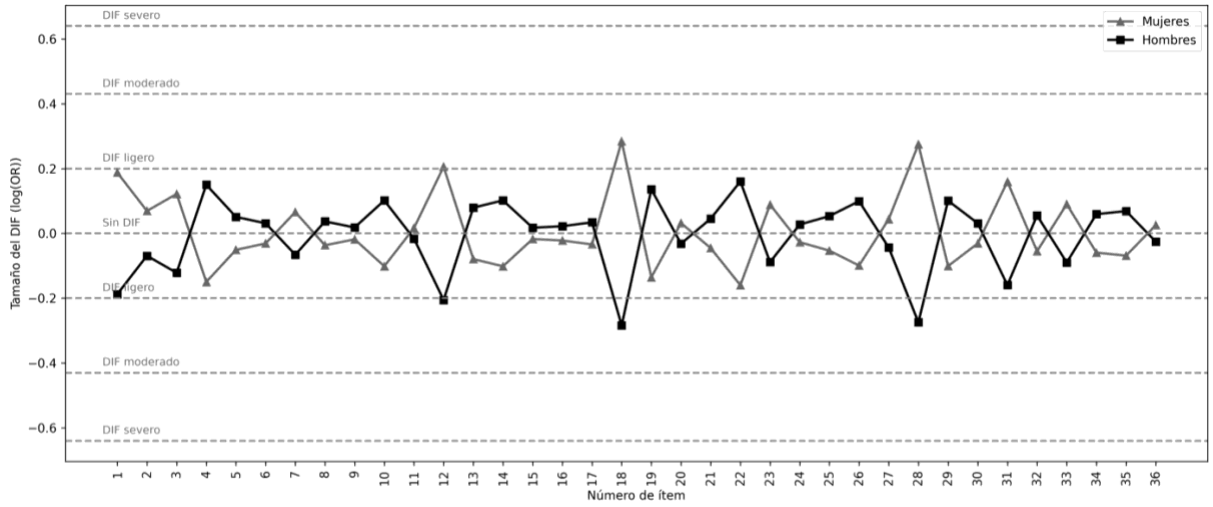
Comprensión	A	36	28	3	5
Comprensión	C	36	30	3	3
Lengua Escrita	A	36	34	1	1
Lengua Escrita	C	36	27	4	5
Matemáticas	A	50	42	5	3
Matemáticas	C	50	44	4	2
Total	–	244	205 (84%)	20 (8.19%)	19 (7.78%)

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

Lectura. En la Figura 36 se muestra el análisis del DIF por sexo, correspondiente al área de Lectura (Forma A). De manera visual, puede observarse que los ítems se distribuyen de forma simétrica en torno al eje central (log (OR) = 0), sin evidenciar separación relevante entre las curvas de hombres y mujeres. De acuerdo con los umbrales que propone Zieky (1993) para la interpretación del log (OR), todos los ítems se ubican en el rango de DIF leve, lo que indica la ausencia de magnitudes moderadas o severas de DIF. Este hallazgo sugiere que el comportamiento de los ítems resulta comparable entre sexos y que, en caso de existir diferencias mínimas, no serían sistemáticas.

Figura 36

DIF por sexo en Lectura (Forma A) del ExIES 2023-2

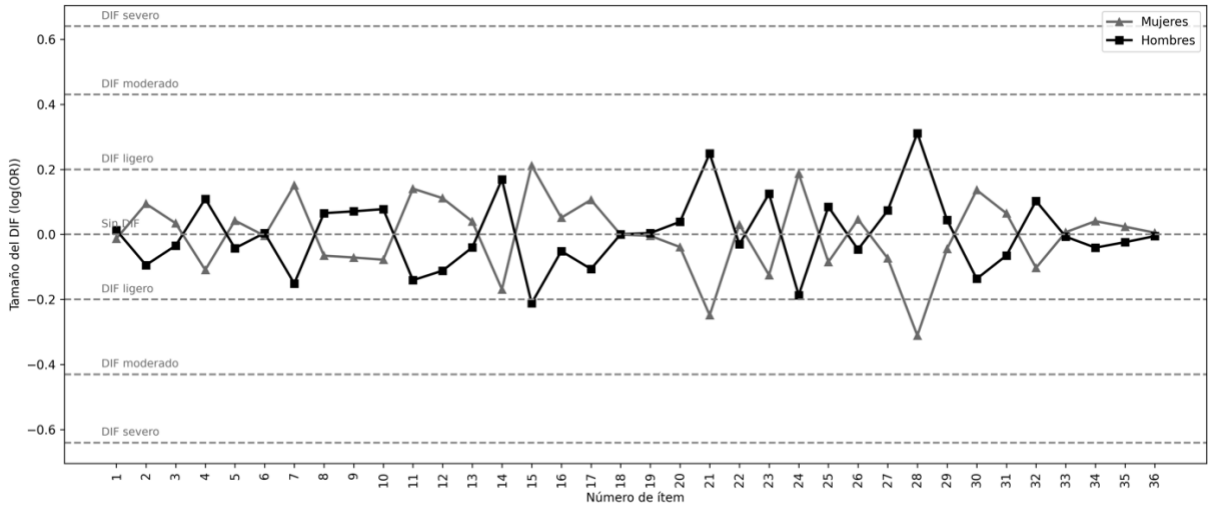


Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 5, Figura 1).

De manera similar, la Figura 37 muestra el log (OR) para la Forma C de Lectura. Nuevamente, se observa que los 36 ítems se ubican cerca del eje log (*OR*)=0. De acuerdo con los criterios establecidos (±0.43 para DIF moderado y ±0.64 para DIF severo), ningún ítem rebasa dichos límites. Así, si bien la Tabla 89 indica un leve incremento de ítems que pudieran favorecer a uno u otro grupo (30 sin DIF, 3 a favor de hombres y 3 a favor de mujeres), la magnitud de esos valores se mantiene en la categoría de DIF leve. Esto significa que la prueba de Lectura (Forma C) tampoco muestra signos de favorecer sistemáticamente a un sexo cuando se contempla a la muestra completa.

Figura 37

DIF por sexo en Lectura (Forma C) del ExIES 2023-2



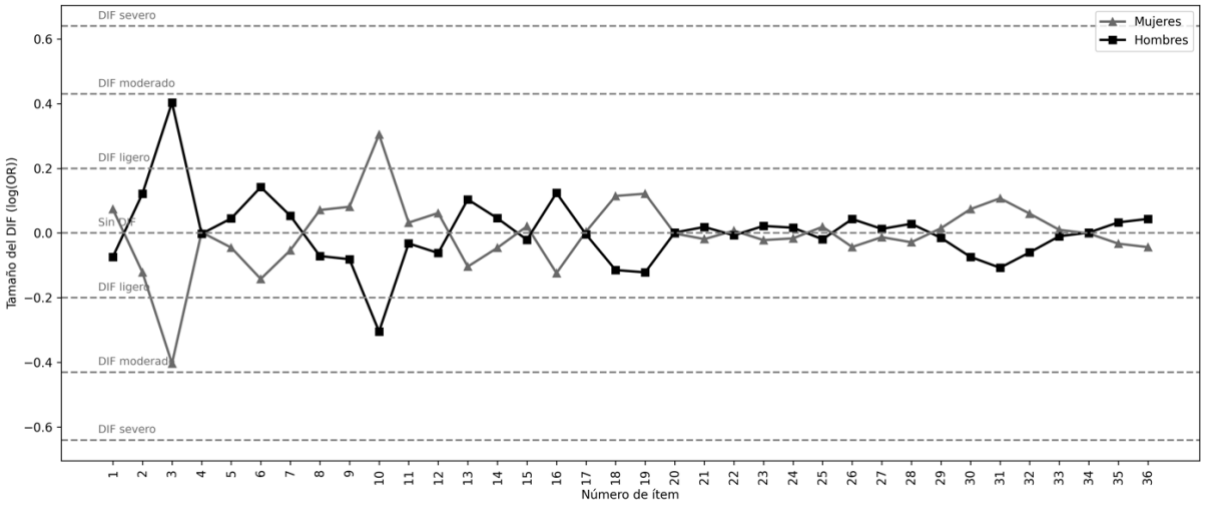
Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 5, Figura 6).

Lengua Escrita. En el análisis global por sexo para la Forma A de Lengua Escrita, se aprecia que la mayoría de los 36 ítems se concentran junto al eje central (log (OR)=0), sin rebasar los límites de ±0.43; 34 ítems carecen de DIF significativo, 1 favorece a hombres y 1 favorece a mujeres. Destaca únicamente el ítem 3, como se observa en la Figura 38, cuya

proximidad al umbral sugiere una posible tendencia, motivo suficiente para que sea revisado con mayor detenimiento.

Figura 38

DIF por sexo en Lengua Escrita (Forma A) del ExIES 2023-2



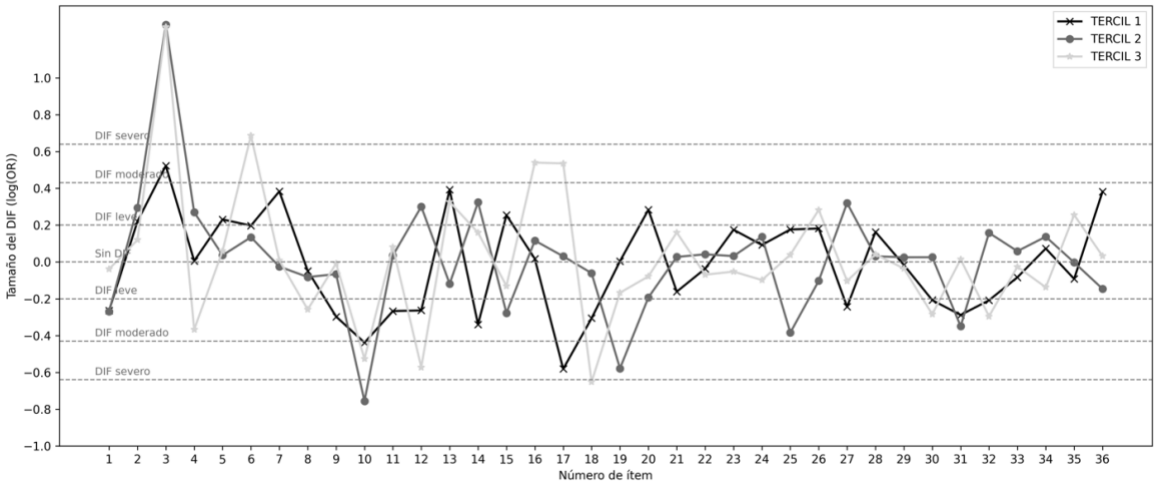
Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 9, Figura 5).

La Figura 38 demuestra otros hallazgos, por ejemplo, el tercil inferior muestra 3 ítems con DIF moderado (3, 10 y 17), el tercil medio presenta un ítem moderado (19) y dos severos (3 y 10), mientras el tercil superior asciende a 6 (4 moderados y 2 severos). Resulta evidente la consistencia con que el ítem 3 aparece como problemático en varias bandas, lo que confirma su elevado potencial de sesgo y subraya la necesidad de examinar sus características (contenido, estructura, tipo de lenguaje, etc.).

Como se observa en la Figura 39, la Forma C de Lengua Escrita exhibe un patrón parecido: la distribución de puntos alrededor de log (*OR*) = 0 no supera ±0.43, lo que sugiere la ausencia de DIF relevante para la comparación global hombres–mujeres; aquí 27 ítems se hallan sin DIF, mientras 4 resultan a favor de hombres y 5 a favor de mujeres, pero con magnitudes que permanecen en la categoría leve.

Figura 39

DIF por tercil de habilidad y sexo en Lengua Escrita (Forma A)



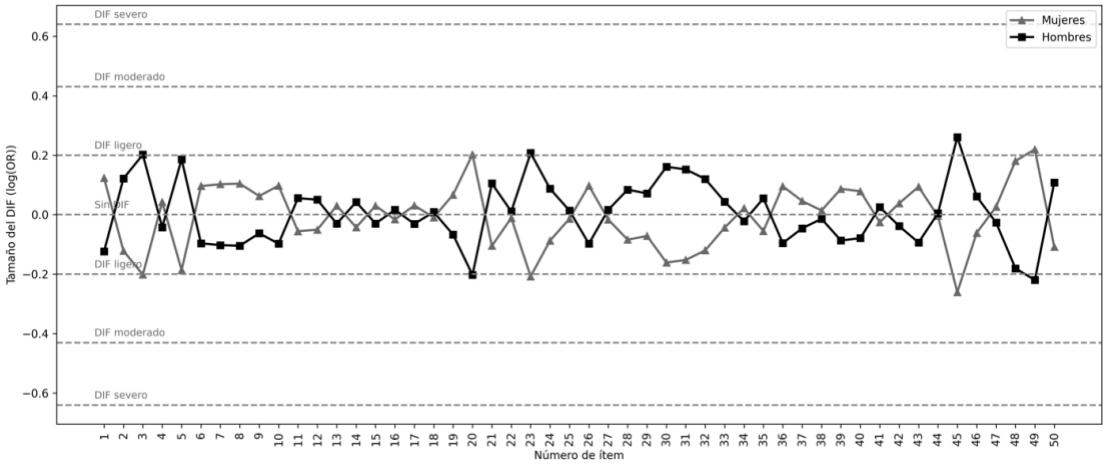
Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 10, Figura 6).

Matemáticas. En la comparación general por sexo, la mayoría de los 50 ítems se ubican cerca del eje central y no alcanzan ±0.43; 42 ítems no mostraron DIF, 5 favorecieron a hombres y 3 a mujeres, pero sin rebasar umbrales de moderado o severo. Este dato da cuenta de que, en un primer análisis global, se percibe un balance razonable en la dificultad de los ítems para ambos sexos. En la Figura 40 se puede observar cómo la mayoría de los ítems de la Forma A se distribuyen en torno al eje central (log (*OR*)).

Asimismo, la Figura 41 muestra el análisis de funcionamiento diferencial del ítem (DIF) por sexo para el área de Matemáticas en la Forma C del ExIES 2023-2. En este gráfico, cada punto representa el tamaño del efecto del DIF (log (OR)) para un ítem específico, comparando el comportamiento entre hombres y mujeres. Se observa que la mayoría de los ítems se agrupan en torno al eje central (log (OR) = 0), lo que indica ausencia de sesgo sistemático a favor de alguno de los sexos.

Figura 40

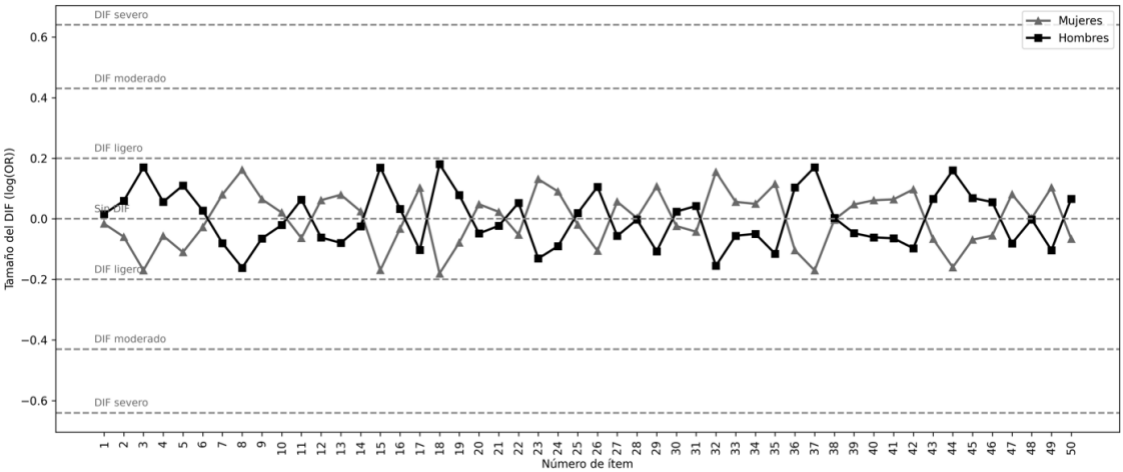
DIF por sexo en Matemáticas (Forma A) del ExIES 2023-2



Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez,2025c, p. 13, Figura 9).

Figura 41

DIF por sexo en Matemáticas (Forma C) del ExIES 2023-2



Nota. Figura tomada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 14, Figura 11).

Además, ninguno de los ítems rebasa los umbrales establecidos para DIF moderado (±0.43) o severo (±0.64), de acuerdo con los criterios de Zieky (1993). Esto sugiere que, en el análisis global, las diferencias encontradas corresponden a DIF leve, sin impacto significativo en la imparcialidad de la medición. En consecuencia, el comportamiento de los ítems en esta forma

de la prueba puede considerarse equivalente entre hombres y mujeres, apoyando la validez y equidad del instrumento para la población evaluada.

Análisis DIF por terciles. El desglose por terciles revela focos de sesgo concretos que no aparecen en el análisis global, este análisis se encuentra en el Apéndice F. 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). La síntesis del Apéndice F (Tabla 88) identifica 27 reactivos únicos con DIF moderado/severo, el 37.5 % de los 72 ítems de esta área; no obstante, sólo 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 (Apéndice F) 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 71 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. La primera garantía es sobre el AFC, esta garantía se apoya en tres supuestos, la evidencia confirma la estructura tripartita prevista, pero con solidez moderada: (S4.1.1, 9/12; 75%) el AFC reproduce tres factores con RMSEA aceptable y CFI/TLI < .90, lo que sugiere ajuste global suficiente pero no óptimo (Hu & Bentler, 1999; AERA et al., 2014). Además, (S4.1.2, 10/12; 83.33%) las correlaciones positivas entre el puntaje global y cada área sostienen la coherencia interna, si bien el patrón entre dominios muestra tensiones conocidas (verbal-cuantitativo), por ello la coherencia y plausibilidad se puntuó con 3. Finalmente, (S4.1.3, 9/12; 75%) la mayor parte de los ítems carga en su factor teórico, pero persisten cargas débiles/negativas en segmentos específicos que conviene depurar. Como mejora, se recomienda:

(a) reescritura/piloteo de reactivos con saturaciones < .20 o negativas; (b) contrastar modelos jerárquicos o bifactor y estimación para datos categóricos (p. ej., WLSMV); y (c) articular explícitamente el marco teórico-de tareas (ECD) en el informe técnico para reforzar el nexo constructo, tarea, proceso de respuesta (Mislevy et al., 2003, 2004; Kane, 2013; Chapelle, 2021; AERA et al., 2014).

Tabla 71

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 cuanto a la segunda garantía sobre validez convergente, la convergencia externa queda sustentada con nivel moderado-alto, pero no exento de márgenes de mejora. Según el S4.2.1, (10/12; 83.33%) el ExIES se asocia de forma positiva y significativa con el EXANI-II en el cruce disponible, aunque la calificación de claridad (3) refleja que la documentación del emparejamiento muestral y el tratamiento de atenuación/restricción de rango pueden transparentarse mejor (AERA et al., 2014; Schober et al., 2018). Asimismo, el S4.2.2 (10/12; 83.33%) muestra correlaciones de moderadas a fuertes (p. ej., Lectura–Comprensión Lectora > .50), mientras Matemáticas tiende a valores más contenidos, consistentes con la literatura de admisión; esto explica que la plausibilidad obtuviera un puntaje de 3 (Kolen & Brennan, 2014; Morales et al., 2015).

La última garantía, sobre imparcialidad entre grupos, está respaldada de forma moderada (S4.3.1, 10/12; 83%**)**, primero porque el análisis DIF no pertenece al examen 2023-1, si no al 2023-2. En un segundo lugar, el 84% de los reactivos no presenta DIF y las diferencias residuales se ubican mayormente en el rango leve; no obstante, el análisis por terciles de habilidad identifica focos moderados/severos en un subconjunto acotado de ítems, lo que justifica coherencia/plausibilidad sea igual a 3 y orienta acciones de mantenimiento (Holland & Thayer, 1988; Zieky, 1993; Dursun & Li, 2021). Por lo que se podría institucionalizar umbrales y rutas de acción (retirar, reescribir, volver a pilotear), monitorear DIF interseccional (sexo por plantel y área), y documentar una bitácora de decisiones para convertir estas debilidades en una ventaja de control continuo de sesgos, en línea con los Estándares 3.2, 3.6 y 3.7 (AERA et al., 2014).

Inferencia de Extrapolación

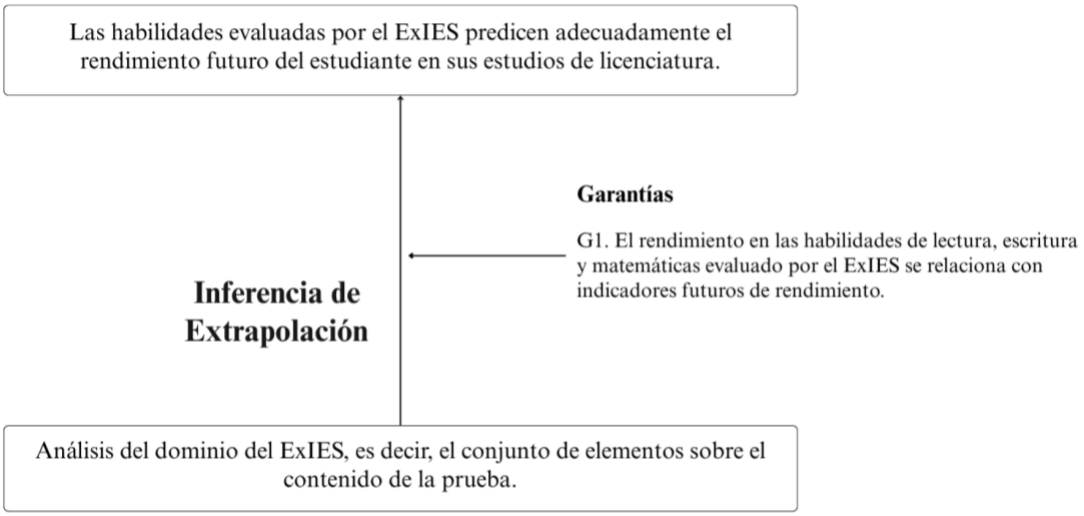
El resultado de esta inferencia se alinea con su objetivo que es evaluar la evidencia predictiva del ExIES respecto al desempeño universitario, considerando la medida en que los puntajes anticipan, con validez empírica, los logros académicos iniciales en la educación superior. Este análisis se enmarca en los Estándares 1.19, 1.20, 5.1, 12.13 (AERA, APA y NCME, 2014), los cuales establecen que las inferencias predictivas deben ser respaldadas con evidencia estadística clara, análisis de error de predicción, y confirmación empírica de la relación entre los resultados del examen y los criterios externos de desempeño. Adicionalmente, estos lineamientos recomiendan considerar la estabilidad de las predicciones a través de métodos como validación cruzada y control de variables contextuales.

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

En la Figura 42 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.

Figura 42

Argumento de la inferencia de Extrapolación como parte de la Validez del Argumento



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 72 muestra la garantía, supuesto y fuentes de datos que respaldan la inferencia de Extrapolación del ExIES. El supuesto establece que el rendimiento en las habilidades de Lectura, escritura y Matemáticas evaluadas por el ExIES guarda una relación positiva con indicadores académicos futuros, como el promedio de calificaciones universitarias; donde las fuentes corresponden a las variables.

Tabla 72

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 (véase Apéndice G).

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 (Gerón, 2019; Burkov, 2019).

Para justificar variables, transformaciones y la descripción completa de hiperparámetros, así como los fundamentos teóricos y antecedentes, es necesario revisar el Apéndice G, sobre los antecedentes y método de este estudio.

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 73). 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 73

Desempeño comparativo de modelos predictivos sobre conjunto de prueba

Modelo	R^2 (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 (Hastie, Tibshirani, & Friedman, 2009; Géron, 2019).

También se realizaron pruebas con diferentes combinaciones entre variables, por ejemplo (véase Tabla 74), los puntajes de matemáticas (M) con el promedio de bachillerato (Promedio Bach), o la combinación entre comprensión lectora (L) y el promedio de bachillerato. Así, los hallazgos expresados en la Tabla 74 reafirman que la mayor capacidad predictiva se obtiene cuando se combinan las puntuaciones del ExIES con el promedio bachillerato. Por ejemplo, solo la variable sobre el promedio de bachillerato (modelo 4) alcanza valores de R^2 de 0.161682 (16.16%) en validación cruzada y 0.165921 (16.59%) en la prueba; a su vez, el modelo que usa solo el puntaje del examen (modelo 2) obtiene 0.085675 (8.56%) en validación cruzada y 0.113369 (11.33%) en la prueba. Sin embargo, al combinar el puntaje global del ExIES con el promedio bachillerato (modelo 5), el R^2 sube a 0.191985 (19.19%) en validación cruzada y 0.210097 (21%) en la prueba; aunque el mejor modelo es cuando se dividen las variables de las

áreas que conforman el ExIES (modelo 1). Asimismo, configuraciones que incluyen puntajes por separado como Lenguaje y Matemáticas (modelo 6) también alcanzan valores cercanos a 0.19-0.20.

Tabla 74

Comparación de configuraciones en regresión lineal sobre desempeño académico

Modelo	R^2 (validación cruzada)	R^2 (prueba)	Pendientes /Coeficientes
1. Regresión Lineal (Básico)	**0.193163**	**0.212871**	PuntajeL: 0.0135 PuntajeM: 0.0143 PuntajeE: 0.0253 PromedioBach_Sistemas: 0.0769
2. Regresión Lineal (Examen)	0.085675	0.113369	PuntajeL: 0.0152 PuntajeM: 0.0329 PuntajeE: 0.0364
3. Regresión Lineal (Puntaje Global)	0.083724	0.109880	PuntajeGlobal: 0.0645
4. Regresión Lineal (Promedio Bach)	0.161682	0.165921	PromedioBach_Sistemas: 0.0895
5. Regresión Lineal (Global-Bach)	0.191985	0.210097	PuntajeGlobal: 0.0407, PromedioBach_Sistemas: 0.0767
6. Regresión Lineal (Puntaje LM)	0.183586	0.197013	PuntajeL: 0.0235, PuntajeM: 0.0182 PromedioBach_Sistemas: 0.0799
7. Regresión Lineal (Puntaje LB)	0.178294	0.190947	PuntajeL: 0.0289 PromedioBach_Sistemas: 0.0844
8. Regresión Lineal (Puntaje MB)	0.173631	0.181467	PuntajeM: 0.0257 PromedioBach_Sistemas: 0.0818
9. Regresión Lineal (Puntaje EMB)	0.190391	0.208790	PuntajeE: 0.0308, PuntajeM: 0.0171 PromedioBach_Sistemas: 0.0772
10. Regresión Lineal (Puntaje EB)	0.185543	0.202761	PuntajeE: 0.0354 PromedioBach_Sistemas: 0.0812

Nota. Los valores en negritas representan el mejor desempeño en cada columna. Esta tabla presenta los resultados comparativos de diferentes configuraciones de modelos de regresión lineal aplicadas al análisis del desempeño académico. Se incluyen valores de R^2 para validación cruzada y pruebas, así como los coeficientes de pendiente de cada predictor. Los modelos varían en función de los predictores considerados, como puntajes específicos (L, M, E) y promedios de bachillerato por área (LM, LB, MB, EMB, EB). Global significa el puntaje final obtenido tras promediar L, M y E. Los coeficientes más altos indican una mayor contribución relativa de la variable al modelo.

Debido a que el otro modelo con buenos resultados fue el de Regresión Lineal Ridge, en la Tabla 75 también se presentan los resultados de las distintas combinaciones de variables, donde se observa un patrón muy similar al de la regresión lineal sin regularización. El mejor desempeño se obtiene cuando se mezclan las puntuaciones de lectura (L), matemáticas (M),

lengua escrita (E) y el promedio bachillerato (Promedio Bach), llegando a un R^2 de 0.212873 (21.28%) en el conjunto de prueba y un 0.193163 (19.31%) en validación cruzada. Con ello, se refuerza la idea de que la regularización L2 (típica de Ridge) no altera drásticamente el desempeño de base, pero aporta estabilidad y evita problemas de sobreajuste o multicolinealidad. Por otra parte, usar solo las puntuaciones del ExIES (modelo 2) o únicamente el promedio bachillerato (modelo 4) arroja valores de prueba en torno a 0.1133 (11.33%) y 0.1659 (16.59%) de R^2, respectivamente, confirmando la importancia de aunar ambos tipos de predictores.

Además, cuando se emplea únicamente la información del ExIES, el modelo que incorpora las tres áreas de forma independiente (modelo 2) alcanza un R^2 de prueba de 0.1134 (11.34%), mientras que el uso del puntaje global promedio (modelo 3) reduce ligeramente la varianza explicada a 0.1099 (10.99%). Aunque la diferencia es modesta, mantener las subpuntuaciones por separado aporta una fracción adicional de poder predictivo y mayor valor diagnóstico; por ello, se recomienda conservar las variables individuales en futuros análisis.

Tabla 75

Comparación de regresión Ridge con distintas combinaciones de predictores

Modelo	R^2 (validación cruzada)	R^2 (prueba)	Coeficientes/Pendientes
1. Regresión Ridge (Básico)	**0.193163**	**0.212873**	PuntajeL: 0.0135, PuntajeM: 0.0143, PuntajeE: 0.0128, PromedioBach_Sistemas: 0.0894
2. Regresión Ridge (Examen)	0.085675	0.113368	PuntajeL: 0.0152, PuntajeM: 0.0328, PuntajeE: 0.0126
3. Regresión Ridge (Puntaje Global)	0.083724	0.109878	PuntajeGlobal: 0.0645
4. Regresión Ridge (Promedio Bach)	0.161682	0.165924	PromedioBach_Sistemas: 0.0894
5. Regresión Ridge (Global-Bach)	0.191985	0.210099	PuntajeGlobal: 0.0407, PromedioBach_Sistemas: 0.0767
6. Regresión Ridge (Puntaje LM)	0.183586	0.197015	PuntajeL: 0.0235, PuntajeM: 0.0182, PromedioBach_Sistemas: 0.0767
7. Regresión Ridge (Puntaje LB)	0.178294	0.190948	PuntajeL: 0.0289, PromedioBach_Sistemas: 0.0844
8. Regresión Ridge (Puntaje MB)	0.173631	0.181469	PuntajeM': 0.0257, PromedioBach_Sistemas: 0.0818
9. Regresión Ridge (Puntaje EMB)	0.190391	0.208792	PuntajeE': 0.0308, PuntajeM: 0.0171, 'PromedioBach_Sistemas: 0.0767

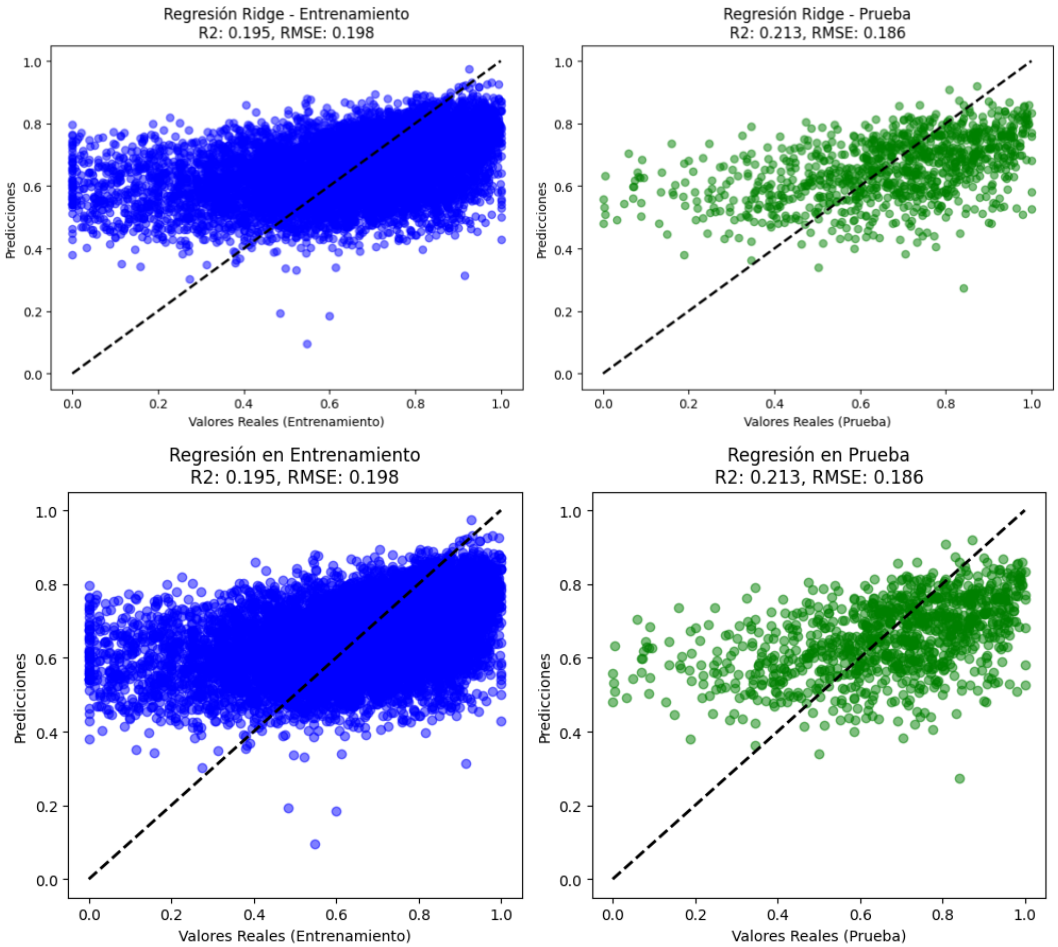
10. Regresión Ridge (Puntaje EB)	0.185543	0.202764	PuntajeE': 0.0354, PromedioBach_Sistemas: 0.0812

Nota. Los valores en negritas representan el mejor desempeño en cada columna. Todos los modelos corresponden a regresión Ridge (regularización L2), aplicados sobre variables escaladas mediante StandardScaler (media = 0, desviación estándar = 1). Los valores de R^2 se obtuvieron mediante validación cruzada (5 particiones) y conjunto de prueba (10% de los datos). Los coeficientes reportados corresponden a las pendientes de cada variable predictora. Elaboración propia.

Se contrastó la efectividad de la validación cruzada frente al muestreo aleatorio (Random Sampling) entre regresión lineal y ridge. En la Figura 43 presenta una comparación gráfica entre los valores reales y las predicciones generadas por el modelo de regresión Lineal y Ridge, tanto para el conjunto de entrenamiento como para el conjunto de prueba.

Figura 43

Dispersión de predicción en muestreo aleatorio versus validación cruzada



Nota. Gráficos elaborados a través de *Visual Studio* con código *Python*. Los resultados corresponden a modelos de regresión lineal y Ridge aplicados a datos normalizados mediante *StandardScaler* (media = 0, desviación estándar = 1). R^2 y *RMSE* calculados en conjunto de entrenamiento y prueba.

En ambos gráficos se observa una dispersión considerable alrededor de la línea diagonal, que representa el ajuste perfecto (valores predichos iguales a los reales). Además, se observa el conjunto de entrenamiento, el coeficiente de determinación ($R^2 = 0.195$) indica que cerca del 20% de la variabilidad en los datos es explicada por el modelo, con un error promedio de predicción (*RMSE*) de 0.198. Por otro lado, en el conjunto de prueba, el modelo muestra un desempeño ligeramente superior ($R^2 = 0.213$), lo que sugiere una adecuada capacidad de Generalización, con un *RMSE* más bajo (0.186). Estos resultados señalan que, aunque existe un grado moderado de incertidumbre en las predicciones, el modelo logra una aceptable capacidad predictiva para estimar el rendimiento académico estudiantil.

No solo se confirma la robustez del modelo de la regresión lineal y la regresión ridge básico, sino también su potencial al aprovechar las ventajas del aprendizaje automático aplicado en contextos educativos. La regularización L2, característica distintiva de ridge, mejora la estabilidad del modelo al reducir problemas de multicolinealidad y, al mismo tiempo, permite un manejo más efectivo de las complejidades inherentes en los datos educativos, minimizando el riesgo de sobreajuste, como lo destacan Hastie, Tibshirani y Friedman (2009), así como Gerón (2019).

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.

La coherencia se aprecia en la consistencia del desempeño entre OLS y Ridge ($R^2 \approx .21–.22$; RMSE≈.19) y en el uso sistemático de particiones 5-fold y escalamiento estándar, en línea con buenas prácticas de modelado (Hastie et al., 2009; Géron, 2019). La plausibilidad es alta porque los tamaños de efecto son comparables con la literatura internacional sobre validez predictiva en admisión (p. ej., SAT/EXANI) y con metaanálisis recientes (Ihlenfeldt & Rios, 2023; García, 2016; Tapasco et al., 2016), y porque el argumento se articula conforme al EBA

(Kane, 2013, 2015; Chapelle, 2021). Por todo lo anterior, como se observa en la Tabla 76, la valoración global es 91.66% (Alta), sustentada en relaciones predictivas claras y razonablemente estables para el primer año universitario (Hastie et al., 2009; Géron, 2019; Ihlenfeldt & Rios, 2023).

Tabla 76

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).

Inferencia de Utilización

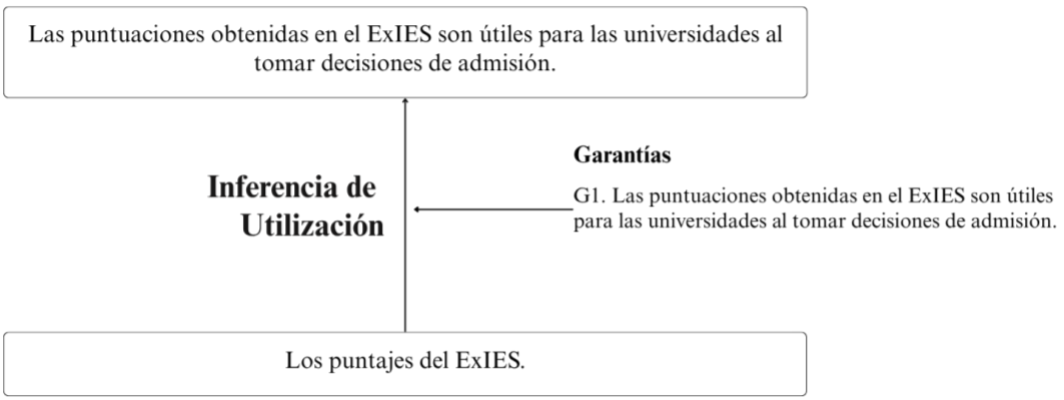
La evaluación de la inferencia de Utilización en el ExIES se fundamenta en el EBA de Kane (2006, 2013) y Chapelle (2021), considerando los Estándares 12.2, 12.1 y 11.4 (AERA et al., 2014), que establecen que los puntajes deben ser empleados de forma coherente con sus fines previstos y con utilidad práctica. El objetivo es evaluar el uso de los puntajes del ExIES en los procesos de selección y admisión institucionales, verificando si estos resultados se aplican de manera pertinente, justa y técnicamente fundamentada en las decisiones de ingreso universitario.

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 44 representa gráficamente el proceso de Utilización del ExIES.

Figura 44

Argumento de la inferencia de Utilización como parte de la Validez del Argumento



La Tabla 77, 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 77

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. 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.

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). Bajo el EBA, esa conexión debe demostrarse y no asumirse (Kane, 2013; Chapelle, 2021). Por ello, en la Tabla 78 se muestra que esta inferencia se valoró con un puntaje global de 58.3%.

Para iniciar a argumentar esta inferencia de forma adecuada, se podría: (a) recolectar evidencia de uso (actas, encuestas e entrevistas) con un diseño claro y trazable (Creswell & Creswell, 2022); (b) formalizar un protocolo UABC–desarrollador que deje por escrito cómo se fijan cortes y cómo se aplica la prelación, tal como recomiendan los Estándares para documentar usos apropiados (AERA et al., 2014); y (c) realizar un estudio de verificación semestral que contraste lo planificado vs. lo aplicado para validar el uso propuesto del puntaje (Kane, 2016). Estas acciones podrían mejorar la organización y gestión de la guía práctica del EBA de Kane para pasar de “está permitido usar el puntaje” a “probamos que se usa bien” (Cook, Brydges, Ginsburg, & Hatala, 2015).

Tabla 78

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)

Implicación de Consecuencias

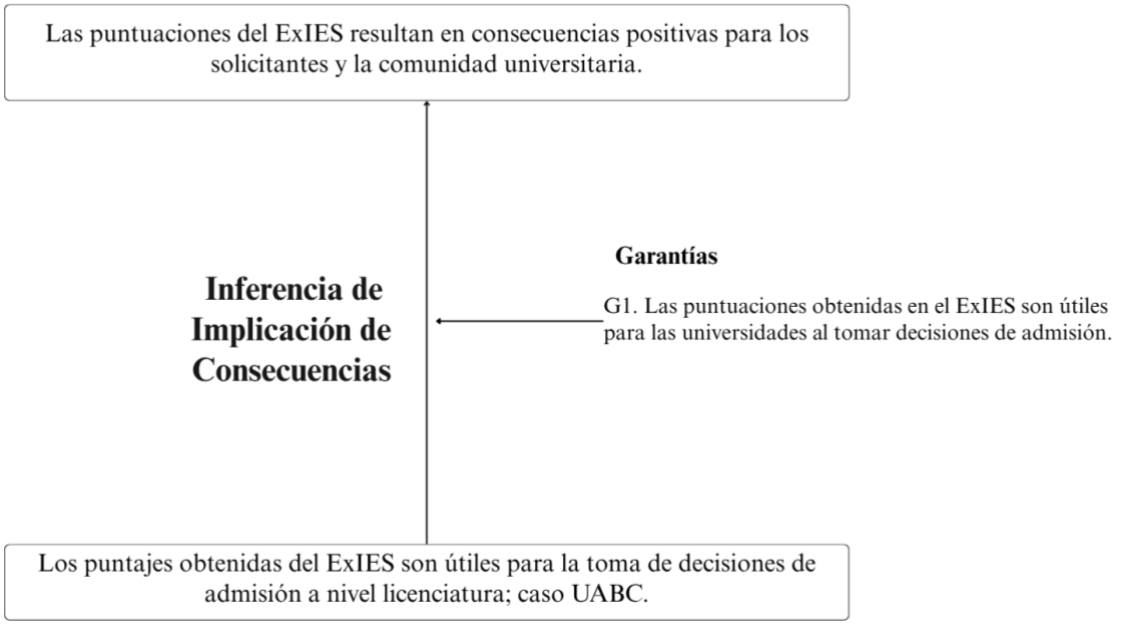
La evaluación de la inferencia de Implicación de Consecuencias en el ExIES se desarrolló siguiendo el marco del EBA (Kane, 2006, 2013) y bajo los Estándares 6.10, 13.1 y 13.6 (AERA, et al., 2014), que exigen documentar y valorar los efectos —esperados o imprevistos— del uso de una prueba en la toma de decisiones. Por lo que, el objetivo es evaluar los efectos de las decisiones basadas en el ExIES desde una perspectiva de imparcialidad y efectividad, identificando si las prácticas actuales de admisión producen consecuencias consistentes con un tratamiento comparable de los aspirantes y si contribuyen a beneficios institucionales observables. Como se revisó en el *Marco Teórico*, como los *Antecedentes*, esta inferencia implica un análisis ético y social de las decisiones basadas en la evaluación.

Definición de las garantías, supuestos y fuentes de la inferencia de Implicación de Consecuencias

La Figura 45 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 45

Argumento de Implicación de Consecuencias como parte de la Validez del Argumento



En este sentido, el propósito de la garantía 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. En ausencia de estudios de impacto, estas garantías se consideran hipótesis respaldadas provisionalmente por normativas institucionales e indicadores psicométricos, pero no por evidencia consecuencial directa. La Tabla 79 sintetiza las garantías y supuestos fundamentales, mostrando que las decisiones de admisión tomadas con base en los puntajes generan consecuencias positivas para los aspirantes y la comunidad académica.

Tabla 79

Estructura argumentativa para la inferencia de Implicación de Consecuencias

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.

En la práctica, este supuesto tiene implicaciones en la forma en que se define el corte de admisión. 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; se requiere vigilancia continua del DIF, estudios longitudinales de retención y eficiencia terminal y protocolos de evaluación de impacto para fortalecer esta inferencia; por mencionar algunos ejemplos de evidencias.

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 & Gómez Monárrez, 2025c) y estudios que sugieren predicción del desempeño cuando se combina el puntaje con el promedio de bachillerato (Pedroza Zúñiga & 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, & Connelly, 2008). Se necesita conectar decisiones de admisión con resultados observables como rendimiento y retención en el primer año para valorar consecuencias (Tinto, 1993; Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008). Así, el 50% que se

refleja en la Tabla 80, 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 (Tinto, 1993; Kuh et al., 2008).

En este caso se puede proceder con: (a) un seguimiento longitudinal de cohortes admitidas para observar calificaciones y permanencia (Tinto, 1993; Kuh et al., 2008); (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 (Zumbo & Chan, 2014; Sackett et al., 2008).

Tabla 80

Evaluación de la Inferencia de Implicación de Consecuencias

Supuesto	**Descripción**	**Claridad (1–4)**	**Coherencia (1–4)**	**Plausibilidad (1–4)**	**Puntaje global (3–12)**
S7.1.1 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)
S7.1.2 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	2	2	2	6 (50%, Baja)

	se reportan con el contexto necesario y advertencias sobre usos indebidos (Estándar 13.6).				
	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.

Valoración del Argumento Global

Como se mencionó, el último paso del EBA es la expresión del Argumento de Validez, es decir, la relación entre la evidencia, respaldos y recomendaciones; la cual se puede visualizar en el Apéndice H. No obstante, con el propósito de plantear los resultados de forma visual, la Tabla 81 muestra una valoración del Argumento Global (LVg) —que refiere a la valoración del Argumento como Unidad según Kane (2006, 2013)— de 78.7% (312/396 puntos), con puntaje 9.4/12, valor que indica solidez suficiente para sustentar decisiones de admisión de alto impacto dentro de un nivel de interpretación moderado; es decir, la evidencia es mayormente sólida; se aconseja un seguimiento continuo. La coherencia interna del razonamiento y la convergencia de múltiples fuentes de datos confirman que el examen mide, de manera coherente y técnicamente adecuada, las competencias de lengua escrita, matemáticas y lectura requeridas en el nivel universitario.

Tabla 81

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

Lo anterior se justifica porque las inferencias con interpretación alta fortalecen el argumento: Generalización (88.8%) y, sobre todo, Extrapolación (91.7%). A su vez, Evaluación (86.9%) y Explicación (80.5%) se ubican en un nivel moderado, aportando soporte técnico relevante sobre procedimientos de aplicación, puntuación y estructura interna. En conjunto, la coherencia del razonamiento y la convergencia de múltiples fuentes de datos confirman que el examen mide, de manera consistente y técnicamente adecuada, las competencias de lengua escrita, matemáticas y lectura requeridas en el nivel universitario. Tales hallazgos cumplen con los requerimientos de evidencia empírica y argumentativa señalados en los Estándares 1.13–1.15 y 7.0 (AERA et al., 2014) y respaldan la solidez técnica (Kane, 2013).

En contraste, las áreas con mayor margen de mejora se concentran en Definición de Dominio (84.8%, moderada) y, especialmente, en Utilización (58.3%, baja) e Implicación de Consecuencias (50%, baja). Estas dos últimas demandan fortalecer la trazabilidad del uso institucional del puntaje (p. ej., documentación del orden de prelación y su aplicación) y la evidencia sobre consecuencias (estudios de impacto, equidad y seguimiento longitudinal). Definición de Dominio ya supera el umbral de aceptación, pero requiere afinar la actualización

de contenidos y la alineación curricular. Con estas acciones, el argumento global podría elevarse desde el nivel moderado observado hacia un desempeño superior en próximas aplicaciones.

Discusión y conclusiones

Con base en los resultados del AIU, evaluados a partir de la Escala EBA, la evidencia reunida respalda la interpretación de las puntuaciones del ExIES como criterio para seleccionar aspirantes a programas de licenciatura en la UABC; lo cual responde a la pregunta de investigación presentada. Esta afirmación está sostenida por el resultado de los objetivos específicos, los cuales se alinearon intencionalmente con cada inferencia del AIU con el fin de valorar los supuestos según las fuentes de datos disponibles o elaboradas (como en el caso de Extrapolación).

Para dimensionar este juicio, conviene recordar que las pruebas transitan por procesos de validación largos y continuos (AERA et al., 2014; Kane, 2013; Markus & 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 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.

Lo último no implica, per se, debilidad en el AIU, sino el inicio de un ciclo argumentativo de acumulación de evidencia y mejora iterativa coherente con los Estándares (AERA et al., 2014) y con la visión contextual-pragmática de la validez (Zumbo & Chan, 2014); como también lo indican Kane (2013, 2015) y Chapelle (2021), al referirse a la relevancia del contexto específico (como la población, las decisiones, la normativa, el currículo, valores y consecuencias) más que a una propiedad abstracta de la prueba. En otras palabras, gracias al EBA es posible revisar aquellos aspectos que hacen falta considerar o mejorar para el ExIES, lo que marca el inicio del uso de una metodología en un proceso de mejora continua.

Y es que, en efecto, las inferencias de Utilización e Implicación de Consecuencias siguen siendo los eslabones débiles en la literatura, pues su documentación y estudio exige estudios longitudinales, trazabilidad de decisiones y análisis de equidad/impacto (Lavery et al., 2020). Las RSL de Dursun y Li (2021), Lavery et al. (2020), de Cook et al. (20139, como la revisión de los *Antecedentes* (Utilización con seis, e Implicación de Consecuencias con 4 estudios), confirman

esta sub-cobertura y la falta de heterogeneidad, o simplemente, la ausencia de fuentes de evidencia de este tipo. En este punto, dejando a un lado pruebas tan consolidadas como el SAT o ACT (College Board, 2023a; ACT, 2024), donde la justificación del uso, así como la claridad sobre la equidad que se encuentran en sus manuales y reportes técnicos (sin formar un AIU explícito); la ruta para el ExIES es clara: transparentar reglas de decisión (criterios de corte, prelación, excepciones); así como realizar seguimientos de cohorte longitudinales; y fortalecer la auditoría de equidad (p. ej., DIF por sexo, plantel, región). Por ende, el aporte de incluir estas inferencias es lograr visibilizar estas necesidades para el ExIES, donde un buen punto de partida es que la institución (p. ej. UABC) tiene una base normativa para poder hacerlo, es decir, la Ley Orgánica de la UABC (2010), el Estatuto General (UABC, 2019), y el Estatuto Escolar (UABC, 2021); ya que sin ellas no hay un principio rector o justificación.

En cuanto a las inferencias con mayor tradición psicométrica, Extrapolación, Generalización y Explicación, los hallazgos en el ExIES fueron altos y moderados. En términos conceptuales, estas inferencias se apoyan en marcos consolidados (confiabilidad y equiparación para Generalización; estructura interna y relaciones con otras variables para Explicación; asociación/predicción con criterios para Extrapolación), lo que explica su frecuencia en la literatura (Kane, 2013; Chapelle, 2021). En la RSL de Dursun y Li (2000–2018), por ejemplo, se observa precisamente esa preeminencia (Explicación ≈ 39; Generalización ≈ 31; Extrapolación ≈ 32), frente a menor atención a Utilización/Consecuencias y a Definición de dominio (Dursun & Li, 2021). Lo mismo sucedió en la revisión de la literatura (*Antecedentes*) donde hubo un total de 17 artículos de Generalización, 16 de Extrapolación y 12 de Explicación. Esto es consistente con el SAT y ACT, que publican de manera sostenida manuales técnicos, estudios de equiparación y de validez predictiva que se relacionan a estas tres inferencias—aunque no por inferencias—

(College Board, 2023a; ACT, 2024); en América Latina, INEP/ENEM, DEMRE-PAES/PDT e ICFES-Saber 11, también se difunden especificaciones y resultados periódicos que, con variaciones de detalle, soportan sobre todo esta tríada psicométrica (INEP, 2023; DEMRE, 2021; ICFES, 2024a, 2024b).

En el caso de esta triada, su cobertura permite una mayor facilidad para el análisis y evaluación (Durson & Li, 2021) de una prueba. Por ello, para la mejora continua del ExIES conviene: (a) reforzar la generalización mediante equiparación y anclaje IRT/Rasch y estudios de generalizabilidad para garantizar comparabilidad entre formas y momentos (Kolen & Brennan, 2014; Brennan, 2006); (b) consolidar la explicación con análisis factorial confirmatorio e invarianza, control sistemático de DIF y evidencia de procesos de respuesta (Sireci & Faulkner-Bond, 2014; Magis, Béland, Tuerlinckx, & De Boeck, 2010; Kane, 2013); (c) robustecer y seguir actualizando la extrapolación mediante modelos predictivos multivariados (validez incremental), o bien estudios longitudinales amplios, a cuatro años (Kane, 2013; Chapelle, 2021).

En cuanto a la inferencia Definición de Dominio y Evaluación—aunque son las bases estructurales—mantienen menor visibilidad en estudios publicados (Esfandiari et al., 2018; Rafatbakhsh & Ahmadi, 2022; Fechter et al., 2021; Gotch & French, 2020); corroborado en los *Antecedentes* con cinco artículos sobre Definición de Dominio y nueve en Evaluación. La Definición de Dominio es el cimiento y la raíz que entreteje el resto del AIU, al enlazar constructo-contenido-procesos y ordenar la acumulación de evidencias (Messick, 1989; Chapelle, 2021; AERA, APA, & NCME, 2014); sin embargo, la RSL de Dursun y Li (2021) y la de Lavery et al. (2020) —así como se revisó en los *Antecedentes*— muestran que suele reportarse poco (con frecuencia queda en manuales/guías internas). En el ExIES esto representa un logro de visibilidad, orden y estructura. Si bien la valoración de ambas es moderada, la Definición de

Dominio cuenta con la con trazabilidad ítem-subcontenido-NDC y base en el MCCEMS, y tiene pendiente mejorar las actualizaciones curriculares y su matriz de contenidos. Algo similar ocurre con Evaluación, donde conviene sistematizar los repositorios de capacitación y auditorías de condiciones en informes técnicos públicos, alineado a los Estándares (AERA et al., 2014). Tanto el SAT como el ACT, publican evidencia de contenido de la prueba; (College Board, 2023a; ACT, 2024); en el caso latinoamericano, el más acercado a este tipo de publicaciones sería INEP (2023); pero en ambos casos no organizados por inferencias, garantías y supuestos.

Ahora bien, el proceso de validación de una prueba es, en esencia y según la AERA et al. (2014), la recopilación y evaluación de la evidencia propuesta, la cual permite justificar las interpretaciones y usos de sus puntuaciones, lo que conlleva a tomar a esa prueba como clara, coherente y plausible para tomar decisiones. Sin embargo, como se ha comprobado en este documento, estas afirmaciones no son fáciles de realizar, según Lavery et al. (2020) conlleva una complejidad que deviene de los debates sobre el uso la lógica informal (organización del argumento a través del modelo de Toulmin) o formal (métodos psicométricos y estadísticos). Es por lo que el EBA de Kane (2006, 2013, 2020) y Chapelle (2021) ha sido una propuesta interesante para poner en práctica con el ExIES. Este gran argumento, basado en peldaños de inferencias y respaldos, ayuda a mejorar la organización —con claridad, coherencia y plausibilidad— de las afirmaciones.

Entonces, si bien las RSL (Dursun y Li, 2021; Lavery et al., 2020) coinciden en que el EBA ha ganado terreno como marco teórico útil, su adopción ha sido desigual y limitada. Se señalan tres causas principales: (a) existe debate epistemológico y técnico entre enfoques "estructurados" vs. "flexibles" para construir el argumento, lo que genera confusión y hace que el método parezca críptico para los no especialistas; (b) las guías formales (p. ej. los Estándares)

ofrecen principios generales pero poco detalle operativo sobre cómo construir o documentar paso a paso un Argumento de Validez, de modo que las prácticas concretas quedan a menudo a criterio del investigador; y (c) la investigación publicada muestra que muchos instrumentos carecen aún de evidencias completas (especialmente sobre dominio, usos y consecuencias), de modo que la comunidad no dispone todavía de modelos consolidados replicables.

Debido a estas dificultades, un acierto de esta tesis fue la integración de figuras y tablas para hacer visible el razonamiento; como también dar seguimiento a la propuesta de Chapelle (2021) desde su interpretación metodológica. Los diagramas del AIU y las figuras por inferencia (p. ej., Figura 27) funcionan como guía didáctica; las tablas de estructura argumentativa (p. ej., Tabla 37) declaran fuentes de datos y vínculos con garantías; lo cual también permite realizar recomendaciones pertinentes. Este repertorio dialoga con propuestas recientes que usan esquemas tipo Toulmin o matrices equivalentes (Choi, 2021, 2022; Fechter et al., 2021; Nomura et al., 2020; Dabrh et al., 2020; Rafatbakhsh & Ahmadi, 2022; Tavares et al., 2017; Yan & Staples, 2019; Mendoza & Knoch, 2018; Li, 2018). El resultado fue un diseño que no eleva artificialmente la carga de validación, sino que mejora la forma de presentar y rastrear —incluso transparentar—el respaldo de cada reclamo, favoreciendo la revisión y previniendo la sobre extensión del uso del puntaje (como bien anticipaban Gotch & French, 2020).

Además, como se ha ido anticipando, la decisión de trabajar explícitamente con las siete inferencias constituye el principal logro metodológico: permitió ver cómo cada enlace aporta evidencia distinta y cómo se encadenan lógicamente desde la observación hasta la decisión institucional, lo que da mayor sentido y responsabilidad ética a un EAI como el ExIES.

Experimentar el EBA con el ExIES —pensándolo como un «portafolio de evidencias» que se defendería en un tribunal: ¿qué usos se quieren autorizar? ¿qué consecuencias se prevén y

cómo se monitorean? — deja un precedente práctico y replicable que puede ayudar a otros procesos de validación para EAI, porque traduce una teoría robusta en pasos operativos y criterios de transparencia que otros equipos pueden adaptar; p. ej. el Apéndice C, el cual es una guía del EBA, sobre todo para la redacción de las inferencias, garantías, supuestos y el tipo de investigación a realizar por inferencia.

En este marco, la Escala EBA propuesta aporta una síntesis numérica de la fuerza del argumento, útil para priorizar acciones y detectar vacíos, sobre todo como herramienta de autoevaluación para la mejora continua, sin sustituir la interpretación cualitativa. Es importante señalar que alcanzar puntajes altos en todas las inferencias constituye un reto permanente, dado que las pruebas son entes vivos, sujetos a cambios y ajustes constantes. Asimismo, el empleo de esta escala se vio fortalecida por el trabajo colaborativo y paulatino con el equipo del ExIES, lo que permitió comprender con mayor profundidad el proceso de construcción del examen. Cabe señalar que, en publicaciones breves, puede omitirse la figura de argumentación por inferencia —por su carácter principalmente ilustrativo— y priorizar la tabla de estructura argumentativa, preservando así la trazabilidad esencial del EBA.

Es así como con estos criterios —claridad, coherencia y plausibilidad— los hallazgos del ExIES ofrecen una conclusión integrada, se puede decir que el Argumento Global (78.7%) se fortaleció por la inferencia Generalización (88.8%) y Extrapolación (91.7%), mientras que Utilización (58.3%) e Implicación de Consecuencias (50%) mostraron claramente líneas de mejora, por lo que las mejoras podrían ser más evidentes en estas dos últimas pero también es posible mejorar cada inferencia según el criterio. Y, aunque cada uno de los criterios se mantuvo en una interpretación moderada, la coherencia es el punto más fuerte (85.4%), seguido de la claridad (79.5%) y plausibilidad ligeramente menor (77.2%). En cuanto a los hallazgos por

criterio, según el EBA (Kane, 2015; 2020) y los Estándares (AERA et al., 2014), a mayor coherencia se sugiere que las evidencias tienden a encajar internamente y no presentan contradicciones relevantes (modelos psicométricos, equiparación, procesos de revisión); en el ExIES se presentan dos inferencias con alta coherencia: Evaluación y Extrapolación.

La claridad moderada, por su parte, indica que algunos supuestos y procedimientos están bien formulados y documentos, y algunos otros no están explicitados a detalle (p. ej., mapeos dominio e ítems, reglas operativas, bitácoras), lo que podría afectar la reproducibilidad y la evaluación externa; en este caso la Generalización y Extrapolación han sido muy claras, dejando líneas de mejora para el resto de las inferencias. Y, la plausibilidad permite señalar algunas lagunas en la fundamentación teórica o en la evidencia empírica que conecte los puntajes con usos y consecuencias reales (solo con la excepción de Extrapolación), es decir, la evidencia debe ser relevante y creíble (verosimilitud); en el ExIES algunas evidencias carecen de aspectos teóricos y operativos robustos —umbrales, reglas de decisión, ECD—, o imprecisiones que tienen relación con aspectos de forma.

Finalmente, la aplicación de los criterios de claridad, coherencia y plausibilidad resultó fundamental, ya que permitieron enfocar la evaluación en determinar si las evidencias estaban teóricamente respaldadas, si guardaban coherencia entre sí y si ofrecían una base convincente para sostener las inferencias; aspectos señalados en los Estándares como esenciales para valorar la validez de un examen (AERA et al., 2014). Sobre todo, considerando que estos conceptos se diluyen al presentar los hallazgos en estudios (Kane, 2015; Lavery et al., 2020). Estos resultados, permiten observar cómo va la garantía, la inferencia y con ello la prueba como ente global. Así, este ejercicio contribuye a la visibilidad del Argumento de Validez, en consonancia con la

propuesta de Kane (2013) sobre la importancia de la coherencia lógica entre las inferencias, y con el énfasis de Chapelle (2021) en la plausibilidad como criterio de juicio dentro del EBA.

No obstante, es importante resaltar que el juicio final se mantiene en las concatenaciones de los argumentos y las evidencias propuestas, por lo que, estos criterios ayudan a mejorar la visibilidad para la mejora continua. Es decir, no es procedente ni realista asignar un 100 % a la validez de una inferencia, la validez es una propiedad de grado y las inferencias en el EBA son presumptivas y sujetas a revisión continua.

Limitaciones

Aunque pudieran escaparse algunas limitaciones, acá se consideran dos generales y algunas limitaciones por etapa según el *Procedimiento*. Primero, la actualización continua del ExIES obliga a seleccionar y presentar lo más relevante en cada iteración; varias mejoras ya se están aplicando, pero este carácter dinámico exige mantener el argumento y sus resúmenes de evidencia como documentos vivos; sin embargo, debido al tiempo y a la extensión del trabajo, la comparación sistemática de evidencias entre ciclos no fue posible. Segundo, se requiere aún más profundización en estudios similares para mejorar la plausibilidad de la propia evaluación, por ejemplo, en técnicas y tipos de evidencia según la inferencia; es decir, seguir realizando revisiones sistemáticas, pero por tipo de inferencia independiente del EBA; lo cual no se profundizó debido al tiempo. En cuanto a las limitaciones por etapas se enlistan a continuación.

Etapa 1. definición del AIU: se partió de lo existente, al no existir argumentos anteriores, las referencias de la construcción fueron a partir de la literatura consultada de forma teórica, por lo que estos argumentos forman un precedente para continuar con modificaciones en pro de la mejora continua.

Etapa 2. garantías, supuestos y fuentes: al partir de las fuentes existentes, aunque bastas, naturalmente algunas no se lograron recuperar para fortalecer aún más o desarrollar otros supuestos.

Etapa 3. Desarrollo de fuentes de datos: se concentró la recolección en validez predictiva, que, aunque no hubo limitaciones en este caso particular, podrían haberse generado otros estudios, por temas de tiempo no se logró; por ejemplo, considerar un estudio a partir de la teoría de la Generalizabilidad para la inferencia de la Generalización; algunos otros se colocaron en las recomendaciones que podrían considerarse gracias según su plan.

Etapa 4. Desarrollo de respaldos: debido al tiempo y la limitación de la literatura, no se profundizó en diversos referentes teóricos según cada supuesto, por lo que la revisión continua, así como continuar con revisiones externas, puede ayudar a robustecer aún más el desarrollo de los respaldos.

Etapa 5. Valoración por inferencia (Escala EBA): aunque resultó un ejercicio relevante y una propuesta que puede ser útil en valoraciones futuras, requiere mayor discusión y evidencia para generalizar su aplicación.

Etapa 6. Argumento global: no hubo limitaciones específicas, pero al promediar, se pueden ocultar problemas puntuales, por lo que mostrar las reservas siempre es necesario (véase Apéndice H).

Etapa 7. Ciclo iterativo: no se compararon ciclos posteriores a 2023-2 por tiempo y alcance. Mantener un repositorio con versiones (qué cambió y por qué), fijar fechas de revalidación tras cambios importantes puede dar continuidad a esta limitación.

Conclusión

Según los objetivos específicos planteados, se confirma que en la evaluación del ExIES (2023-1) se obtuvieron argumentos claros, coherentes y plausibles de moderados a altos, es decir, con evidencia mayormente sólida (moderada) y libre de contradicciones (alta), ya que: 1) existe congruencia entre el contenido y el dominio definido según el plan curricular vigente (moderada, 84.8%); 2) se cumple con la administración y gestión para el desarrollo completo de la prueba, siendo un resumen fiel del desempeño en las tareas del examen (moderada, 86.9%); 3) las puntuaciones son consistentes a través de formas y áreas evaluadas (alta, 88.8%); 4) las puntuaciones respaldan el constructo pretendido (moderada, 80.5%); 5) el examen predice adecuadamente el desempeño del primer año universitario (alta, 91.7%); 6) el uso institucional requiere fortalecerse (baja, 58.3%); y 7) las consecuencias reportadas aún carecen de verificación sistemática (baja, 50%). En consecuencia, el LVg (moderada, 79%) indica que la evidencia global es suficiente para los usos previstos para la UABC, sobre todo tomando el contexto general de que el ExIES es una prueba joven, y que, de forma muy oportuna, está realizando un proceso de validación para continuar con su proceso de mejora continua a partir de las recomendaciones.

Recomendaciones

En general, se recomienda mantener y publicar anualmente el AIU como la matriz de contenido del ExIES, vinculándolo con las reglas de prelación y puntos de corte; así como estar más en contacto con los tomadores de decisiones sobre qué se les comunica a los sustentantes, es decir, el qué hacer una vez que se les entregan sus resultados; esto podría ser beneficioso para la Implicación de Consecuencias. Aunado a ello, para el equipo desarrollador se pueden revisar de forma detallada en el Apéndice H, donde se establecieron las recomendaciones por supuesto

basándose en la misma teoría propuesta por el ExIES con el fin de mantener la coherencia teórica.

Analizando lo descrito por la AERA et al. (2014), Kane (2013, 2015), Lane et al. (2016), Sireci & Faulkner-Bond (2014), Kolen & Brennan (2014) y Bond & Fox (2015), así como las recomendaciones puntuales del Apéndice H, se recomienda, de forma general, que el ExIES cuente con un sistema integrado y versionado de manuales que deje trazabilidad y continuidad: 1) Definición de dominio: manual de contenido y especificaciones fundamentados de forma teórica, donde se agregue una bitácora de cambios curriculares; 2) Evaluación: continuar con los manuales operativos (aplicación, seguridad e incidencias, capacitación, deshonestidad, guía del sustentante); 3) Generalización: manual de equiparación y calibración, y protocolo de confiabilidad; 4) Explicación: manual de análisis de constructo (AFC/AFE, modelos jerárquico/bifactor), protocolo de DIF e invariancia y manual de convergencia externa; 5) Extrapolación: manual de validación externa que especifique criterios y diseños para evidenciar relaciones con el mundo real (p. ej., GPA, créditos, retención, EXANI II); 6) Utilización: guía de interpretación de puntajes y cortes y manual de decisiones de admisión/prelación; y 7) Implicación de consecuencias: reglamento del comité de consecuencias y protocolo de seguimiento longitudinal y equidad.

Además, sería conveniente elaborar tres manuales transversales y generales que aplican a todas las inferencias y productos: el manual de estilo (normas de redacción, formatos y plantillas; convenciones terminológicas y de notación; citación y rotulación de tablas y figuras; ejemplos modelo y anti-ejemplos) y la guía o manual de bases de datos (diccionario de variables, sintaxis, llaves y versionado; reglas de depuración/imputación; resguardo y accesos; plantillas de reportes), y un manual o guía de procedimientos técnicos reproducibles para análisis (p. ej., flujo

de Rasch, correlaciones o validez convergente, AFC, DIF y equiparación, con software, parámetros y umbrales). Al ser generales, estos dos documentos alinean y estandarizan el trabajo de todas las áreas y garantizan que, independientemente de la inferencia, la evidencia sea comparable, verificable y actualizable; y con ello mejorar la claridad, coherencia y plausibilidad.

Con la implantación y versionado de estos manuales y guías, el ExIES quedaría alineado con los estándares internacionales más estrictos en la materia, entendidos principalmente como los Estándares (AERA, et al., 2014), el EBA (Kane, 2013, 2015; Chapelle, 2021), y las mejores prácticas operativas para equidad, confiabilidad, equiparación y uso de puntajes (Sireci & Faulkner-Bond, 2014; Brennan, 2006; Cronbach et al., 2004; Kolen & Brennan, 2014).

Sin olvidar que se puede consultar el Apéndice C, la cual funciona como Guía general para proponer las inferencias, garantías, supuestos y tipo de evidencias recomendadas. Asimismo, sería conveniente desarrollar un *dashboard* o tablero con los resultados del ExIES por inferencia y como una especie de portafolio de evidencias con tendencias más automatizadas y visualización por ciclo, podría ser un logro de transparencia hacia la comunidad educativa; o si se realiza internamente puede ser un buen auxiliar de transparencia entre los tomadores de decisiones; véase Apéndice I.

En cuanto a recomendaciones metodológicas, para adoptar el EBA se recomienda, en un primer momento, delimitar y publicar el AIU desde el diseño inicial: explicitar la afirmación final que se pretende defender, los usos previstos de la puntuación y la población meta; y describir la cadena de siete inferencias junto con sus garantías y supuestos clave. Conviene, además, vincular cada inferencia con los Estándares aplicables y preidentificar la evidencia a reunir (contenido, procesos de respuesta, estructura interna, relaciones con otras variables, consecuencias), así como las técnicas analíticas correspondientes. Para operacionalizar el plan, se

sugiere estructurar estudios modulares y publicables —por ejemplo, un estudio de confiabilidad para Generalización y otro de validez convergente para Explicación— y discutir sus resultados con la literatura específica antes de integrarlos.

En un segundo momento, se propone usar escalas de valoración transparentes (p. ej., la escala EBA de muy baja a alta) acompañadas de rúbricas explícitas para los criterios de claridad, coherencia y plausibilidad, justificando cada puntaje con ejemplos concretos de la evidencia presentada. Asimismo, resulta pertinente implementar revisiones iterativas respaldadas por metadatos: mantener un repositorio versionado de bases de datos, scripts y decisiones sobre ítems; registrar cambios en especificaciones, pesos y puntos de corte para documentar la evolución longitudinal del argumento; conformar comités de jueces internos y externos con criterios claros de selección y entrenamiento; y fomentar la alfabetización en validez entre tomadores de decisiones mediante talleres y guías breves que faciliten la lectura crítica de reportes técnicos y la participación informada en ajustes de la prueba. Con estas acciones, el EBA se vuelve operativo, auditable y transferible para EAI; véase también el Apéndice C.

Líneas de investigación futuras

Las futuras investigaciones sobre el ExIES podrían centrarse en a) seguimientos longitudinales que vinculen puntajes con trayectorias académicas y evaluar la validez incremental del examen frente a otros predictores; b) estudios de decisión y costo-beneficio para afinar puntos de corte y reglas de prelación; c) análisis de procesos de respuesta que comprueben la coherencia cognitiva de las tareas; d) evaluaciones de comparabilidad entre formas mediante teoría de la generalizabilidad; e) exploraciones de imparcialidad ampliada —DIF multigrupo—; y, f) encuestas y entrevistas a usuarios que documenten interpretación de puntajes y percepción de consecuencias.

En cuanto a otros estudios relacionados con el proceso de elaboración de pruebas y el EBA, 1) investigaciones que innoven en la línea del uso de pruebas con Inteligencia Artificial Generativa (IAG), por ejemplo, ya se llevó a cabo una comparación basada en Rasch entre ítems creados con y sin IA (Ruiz-Mendoza & Pedroza, 2025); 2) ahondar en la elaboración de textos con IAG y someterlo a jueceo los resultados entre humanos e IAG para su uso en EAI o pruebas con enfoque formativo; 3) usos de pruebas con enfoque sumativos como formativos, la integración de la IAG, su uso y resultados en Sistemas de Organización del Aprendizaje (LMS, por sus siglas en inglés); 4) aplicar el EBA desde un enfoque formativo.

Por último, y no menos importante, el EBA ofrece ese marco fértil para estudiar cómo la verdad de los resultados de un examen, como el ExIES, se va construyendo y legitimando en actos comunicativos (Rorty & Habermas, 2012), por lo que investigar sobre la inclusión de actores, dentro de la deliberación habermasiana, más medir su impacto en la aceptación social de cortes y decisiones podría ser relevante. Aunado a ello, se podría profundizar en la justificación pragmática como detalla Rorty (Rorty & Habermas, 2012), es decir, examinar cómo narrativas y las prácticas de transparencia influyen en la confianza y en la llama utilidad pública. Ya que, tanto las afirmaciones, garantías, respaldos y reservas, la calidad como los procesos de deliberación permiten que este enfoque sea público y estructurable, es decir, es la comunidad quien hace explícita la defensa y revisión de las inferencias dentro de un contexto determinado. La verdad, entonces, se justifica con el AIU donde resisten la crítica pública y guían los usos responsables. En este sentido, el EBA operacionaliza una convergencia práctica entre la validez habermasiana—la fuerza del mejor argumento en condiciones de diálogo—y la justificación rortyana—acuerdos contingentes que funcionan para una comunidad—: como seres humanos,

convergemos en verdades intersubjetivas que nos permiten avanzar en comunidades e instituciones, sobre todo en decisiones de alto impacto.

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.

Alkin, M. C. (2012). *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed.). SAGE Publications.

Alnahdi, G. H. (2015). Validity and predictive ability of the General Aptitude Test (GAT) for academic performance in Saudi higher education. *Educational Assessment, Evaluation and Accountability, 27*(3), 237–254. https://doi.org/10.1007/s11092-015-9212-7

Alotaibi, N. (2021). The relationship between college admission test scores, GPA, and academic performance in higher education: Evidence from Saudi Arabia. *Journal of Education and Learning, 10*(4), 123–132. https://doi.org/10.5539/jel.v10n4p123

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

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 Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2018). *Estándares para pruebas educativas y psicológicas* (Versión oficial en español). American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests* (2nd ed.). American Psychological Association.

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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

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]. (2023). *Anuarios estadísticos de educación superior. Ciclo escolar 2021–2022 y 2022–2023.* http://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

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

Atchison, D., Garet, M. S., Smith, T. M., & Song, M. (2022). The validity of measures of instructional alignment with state standards based on Surveys of Enacted Curriculum. *AERA Open,* 8. https://doi.org/10.1177/23328584221098761

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.

Backhoff, E., & Tirado, F. (1992). Diseño y validación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa, 4*(1), 1–15.

Bennett, C. T. (2022). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal, 59*(1), 180–216. https://doi.org/10.3102/00028312211003526

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

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev,* 54, 1937–1967 (2021). https://doi.org/10.1007/s10462-020-09896-5

Bernstein, B. (2000). *Pedagogy, symbolic control and identity: Theory, research, critiqu*e (rev. ed.). Rowman & Littlefield.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer. https://doi.org/10.1007/978-0-387-45528-0

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. (2006). The Attack of the Psychometricians. *Psychometrika*, 71(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

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

Bourdieu, P. (1979). *La distinction: Critique sociale du jugement*. Les Éditions de Minuit.

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.

Brookhart, S. M., & Nitko, A. J. (2018). *Educational assessment of students* (8th ed.). Pearson.

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

Ceneval. (2023). EXANI-II: Examen Nacional de Ingreso a la Educación Superior. https://www.ceneval.edu.mx/EXANI-ii

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing,* 29(1), 19–27. https://doi.org/10.1177/0265532211417211

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

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733116

Chapelle, C. A., & Voss, E. (2013). Evaluation of language testing through validation research. En A. Kunnan (Ed.), *Companion to language assessment* (pp. 1–13). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla110

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. https://doi.org/10.1177/0265532214565386

Chapelle, C. A., Enright, M., & Jamieson, J. (2008). Building a validity argument for the Test of *English as a Foreign Language*. Routledge.

Chapelle, C. A., Enright, M., & Jamieson, J. (2010). Does an argument-based approach to validation make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, 12(1), 50–66. https://doi.org/10.1080/15434303.2014.981334

Chiva, I., Perales, M. J., & Pérez Carbonell, A. (2009). Historia de la evaluación educativa. En M. Jornet & E. Leyva (Coords.), *Conceptos, metodología y profesionalización en la evaluación educativa* (pp. 43–70). INITE.

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

Cliff, A., & Montero, E. (2010). El balance entre excelencia y equidad en pruebas de admisión: Contribuciones de experiencias en Sudáfrica y Costa Rica. *Revista Iberoamericana de Evaluación Educativa*, 3(2), 8-28. http://hdl.handle.net/10486/661609

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

College Board. (2023b). Research supporting the use of the digital SAT. https://research.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

Cook, D. A., Zendejas, B., Hamstra, S. J., & Hatala, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, 19(2), 233–250. https://doi.org/10.1007/s10459-013-9458-4

Creswell, J. W., & Creswell, J. D. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches* (6ta ed.). SAGE.

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.

Dang, C. N., & Dang, T. N. Y. (2023). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese students in the UK. *RELC Journal, 54*(1), 84–98. https://doi.org/10.1177/0033688220985533

DEMRE–Universidad de Chile. (2021). *Informe de resultados PDT – Admisión 2021*. Departamento de Evaluación, Medición y Registro Educacional. https://demre.cl

Departamento de Evaluación, Medición y Registro Educacional [DEMRE]. (2020). *Prueba de Selección Universitaria (PSU)*. https://historico.demre.cl/investigacion/documentos/informes/2021-informe-resultados-admision-2021.pdf

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the Performance Assessment for California Teachers. *Journal of Teacher Education, 65*(5), 402–420. https://doi.org/10.1177/0022487114542517

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

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

Esfandiari, M. R., Riasati, M. J., Vaezian, H., & Rahimi, F. (2018). A quantitative analysis of TOEFL iBT using an interpretive model of test validity. *Language Testing in Asia, 8*(7). https://doi.org/10.1186/s40468-018-0062-7

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

Ferguson, K. J., Kreiter, C. D., Franklin, E., Haugen, T. H., & Dee, F. R. (2020). Investigating the validity of web-enabled mechanistic case diagramming scores to assess students' integration of foundational and clinical sciences. *Advances in Health Sciences Education, 24*, 695–709. https://doi.org/10.1007/s10459-019-09944-y

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

Furuta, J., Schofer, E., & Wick, S. (2021). The effects of high-stakes educational testing on enrollments in an era of hyper-expansion: Cross-national evidence, 1960–2010. *Social Forces, 99*(4), 1631–1657.

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). Guadalajara: 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.

Ghio, F. B., Bruzzone, M., Rojas-Torres, L., & Cupani, M. (2020). Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial. *Avances en Psicología Latinoamericana, 38*(1), 123–137. https://doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

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

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

Hidri, S. (2021). Linking the International English Language Competency Assessment suite of examinations to the Common European Framework of Reference. *Language Testing in Asia, 11*, Article 9. https://doi.org/10.1186/s40468-021-00123-8

Higdem, J.L., Kostal, J.W., Kuncel, N.R., Sackett, P.R., Shen, W., Beatty, A.S., & Kiger, T.B. (2016). The role of socioeconomic status in SAT-freshman grade relationships across gender and racial subgroups. *Educational Measurement: Issues and Practice*, 35(1), 21-28. https://doi.org/10.1111/emip.12100

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test 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

Huang, H.-T. D., Hung, S.-T. A., Chao, H.-Y., Chen, J.-H., Lin, T.-P., & Shih, C.-L. (2022). Developing and validating a computerized adaptive testing system for measuring the English proficiency of Taiwanese EFL university students. *Language Assessment Quarterly, 19*(2), 162–188. https://doi.org/10.1080/15434303.2021.1984490

ICFES. (2021). *Prueba Saber 11*. https://www.icfes.gov.co

ICFES. (2024a). *Informe nacional de resultados del examen Saber 11° – 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

Instituto Nacional de Estadística y Geografía [INEGI]. (2023). *Mapa de Baja California con división político-administrativa*. Cuéntame de México. https://cuentame.inegi.org.mx/imprime_tu_mapa

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2023). *Exame Nacional do Ensino Médio (ENEM)*. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

Instituto Nacional para la Evaluación de la Educación [INEE]. (2017). *La educación obligatoria en México: Informe 2017*. INEE.

International Baccalaureate Organization. (2021). *What is the IB?* https://www.ibo.org

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer.

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

Jornet, J. M., Perales, M. J., & González-Such, J. (2020). El concepto de validez de los procesos de evaluación de la docencia. *Revista Española de Pedagogía*, 78(276), 233-252. https://doi.org/10.22550/REP78-2-2020-01

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. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. En R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 489–552). Springer Science Business Media. https://doi.org/10.1007/978-3-319-58689-2_16

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

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. En K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). Routledge.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9ª ed.). Cengage Learning.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering.* Keele University and University of Durham, Technical Report. https://doi.org/10.48550/arXiv.2104.05148

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI),* 2, 1137–1145.

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

Kumazawa, T., Shizuka, T., Mochizuki, M., & Mizumoto, A. (2016). Validity argument for the VELC Test® score interpretations and uses. *Language Testing in Asia, 6*(1). https://doi.org/10.1186/s40468-015-0023-3

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing, 34*(4), 451–475. https://doi.org/10.1177/0265532217713951

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

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación En Educación Médica*, 3(9), 40-55. https://doi.org/10.1016/S2007-5057(14)72724-3

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

López-García, Y. & Willms, D. (2019). A validity study of the Evaluation Infantil Temprana (EIT). [Tesis doctoral]. The University of New Brunswick.

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

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

Madaus, G. F., & Stufflebean, D. (2000). Program evaluation: A historical overview. En D. L. Stufflebean, G. F. Madaus & T. Kellaghan (Eds.), *Evaluation in education and human services* (Vol. 49, pp. 3–22). Springer Publishing. https://doi.org/10.1007/0-306-47559-6_1

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.

Métrica Educativa. (2023). Nuestra historia. *Métrica educativa*. https://metrica.edu.mx/quienessomos/nuestra-historia/

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

Ministerio de Educación de Chile. (2021, 11 de febrero). *Prueba de Transición 2021 logra reducir brechas y se duplican puntajes nacionales* [Comunicado de prensa]. Mineduc. https://biobio.mineduc.cl/2021/02/11/prueba-de-transicion-2021-logra-reducir-brechas-y-se-duplican-puntajes-nacionales/

Ministerio de Educación Nacional y de la Juventud. (2023). Le Bac. *Ministerio de Educación Nacional y de la Juventud*. https://www.education.gouv.fr/reussir-au-lycee/le-baccalaureat-general-10457

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report No. 632). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of Maryland. https://cresst.org/wp-content/uploads/R632.pdf

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.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2020). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Palmer, S., & Rolf, J. (2023). Predicting academic success in first-year university students using GPA and prior academic performance. *Journal of Educational Psychology, 115*(2), 456–467. https://doi.org/10.1037/edu0000728

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). *Presentación 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). *Examen 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. (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., & 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., & 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.

Poole, P., Shulruf, B., Rudland, J., & Wilkinson, T. (2012). Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study. *Medical Education*, 46(2), 163-171. https://doi.org/10.1111/j.1365-2923.2011.04078.x

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.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Stewart, L. A. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10, Article 39. https://doi.org/10.1186/s13643-020-01542-z

Rojas, L. (2013). Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de colegio como variables control. Actualidades Investigativas en Educación, 13(1), 45-69.

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

Ruiz-Mendoza, K. K., & Pedroza-Zúñiga, L. H. (2025). Comparación basada en Rasch de ítems creados con y sin IA generativa. *Journal of Technology and Science Education*, 152, 479–494. https://doi.org/10.3926/jotse.3135

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7.a ed.). McGraw-Hill.

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

Saks, K. (2024). The effect of self-efficacy and self-set grade goals on academic outcomes. *Frontiers in Psychology*, 15, Article 1324007. https://doi.org/10.3389/fpsyg.2024.1324007

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

Sánchez-Mendiola, M., & Delgado-Maldonado, L. (2017). Exámenes de alto impacto: implicaciones educativas [High-stakes testing: Educational implications]. *Investigación en Educación Médica, 6*(21), 52–62. https://doi.org/10.1016/j.riem.2016.12.001

Santelices, M. V., & Wilson, M. (2015). The revised SAT score and its potential benefits for the admission of minority students to higher education. Education Policy Analysis Archives, 23(113). https://www.redalyc.org/pdf/2750/275041389104.pdf

Sartania, N., McClure, J. D., & Sweeting, T. J. (2014). UK Clinical Aptitude Test (UKCAT) scores as a predictor of medical school performance: A national prospective observational study. *BMC Medical Education, 14*(1), 116. https://doi.org/10.1186/1472-6920-14-116

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

Scriven, M. (1967). The methodology of evaluation. En R. W. Tyler, R. M. Gagné & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Rand McNally.

Scriven, M. (1991). *Evaluation thesaurus* (4ª ed.). SAGE.

Scriven, M. (2000). Evaluation ideologies. En D. L. Stufflebeam, G. F. Madaus & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 249–278). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47559-6_15

Scriven, M. (2008). A summative evaluation of RCT methodology & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5*(9), 11–24.

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

Sheppard, G., Williams, K. L., Metcalfe, B., Clark, M., Bromley, M., Pageau, P., Woo, M., Yi, Y., Devasahayam, A. J., & Dubrowski, A. (2023). Using Kane's framework to build an assessment tool for undergraduate medical students' clinical competency with point of care ultrasound. *BMC Medical Education, 23*, Article 43. https://doi.org/10.1186/s12909-023-04030-9

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

Sireci, S. G., 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

Soares, J. A. (2012). The Future of College Admissions: Discussion. *Educational Psychologist*, 47(1), 66–70. https://doi.org/10.1080/00461520.2011.638902

Stake, R. E. (2003). *Standards-based & responsive evaluation*. SAGE.

Stufflebeam, D. L. (1972). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5*(1), 19–25. https://eric.ed.gov/?id=EJ057431

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, & applications* (3ª ed.). Jossey-Bass.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications* (2ª ed.). Jossey-Bass.

Stufflebeam, D. L., & Zhang, G. (2017). *The CIPP evaluation model: How to evaluate for improvement and accountability*. The Guilford Press.

Tapasco, L., Martínez, G., & Restrepo, M. (2016). Análisis predictivo del rendimiento académico en educación superior a partir del examen SABER 11. *Revista de Educación Latinoamericana, 52*(4), 435–450. https://doi.org/10.1016/j.rel.2016.10.005

Tapasco, L., Martínez, G., & Restrepo, M. (2016). Análisis predictivo del rendimiento académico en educación superior a partir del examen SABER 11. Revista de Educación Latinoamericana, 52(4), 435–450. https://doi.org/10.1016/j.rel.2016.10.005

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

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6ª ed.). Prentice-Hall.

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).

Tristán López, A. (2013). *Análisis de Rasch para todos: Una guía simplificada para evaluadores educativos* (1a ed. rev.). Instituto de Evaluación e Ingeniería Avanzada.

U.S. National Library of Medicine. (s. f.). *PubMed – Búsqueda avanzada*. National Institutes of Health. https://pubmed.ncbi.nlm.nih.gov/advanced/

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, Coordinación General de Servicios Estudiantiles y Gestión Escolar. (2025). *Información para aspirantes a ingresar*. Recuperado el 5 de mayo de 2025, de http://cgsege.uabc.mx/web/cgsege/aspirantes-a-ingresar

Universidad Autónoma de Baja California, Coordinación General de Servicios Educativos. (2023). *Reporte de matrícula y programas de la UABC, ciclo 2023-1*. https://www.uabc.mx/

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

Universidad Autónoma de Baja California. (s. f.). Admisiones UABC. Recuperado el 30 de junio de 2025, de https://admisiones.uabc.mx

Urrutia, A., Reyes, C., González, M., & Sepúlveda, D. (2014). Validación de contenido de instrumentos para evaluación de competencias en el nivel superior: revisión en Chile. *Calidad en la Educación, 40*(2), 61–80. https://doi.org/10.4067/S0718-45652014000200003

Vidal, R. (2009). ¿ENLACE, EXANI, EXCALE o PISA? CENEVAL. https://es.scribd.com/document/602721563/Rafael-Vidal-2009-ENLACE-EXANI-EXCALI-O-PISA

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. (Trabajo original publicado en 1934).

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

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225. https://doi.org/10.1177/0265532214557113

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

Apéndices

Apéndice A

Método de la RSL sobre el EBA 2014-2024

A continuación, se detalla el proceso de revisión y selección de artículos de la RSL sobre el EBA de 2014 a 2024.

Objetivo

Identificar cómo se operacionaliza el EBA en estudios empíricos del área educativa mediante una RSL utilizando el modelo PRISMA.

Diseño y preguntas de investigación

Como se ha mencionado en la introducción, se ha optado por utilizar el modelo de Elementos Preferidos para Informes de Revisiones Sistemáticas y Metaanálisis (PRISMA, por sus siglas en inglés) para realizar esta RSL, considerando las características y actualizaciones del año 2020 (Page et al., 2021; Rethlefsen et al., 2021). La pregunta general que guía esta revisión es ¿cómo se lleva a cabo el método del EBA en estudios empíricos del área educativa?

P1. ¿Cuáles son las características de los artículos revisados?

P2. ¿Cómo se especifican los AIU?

P3. ¿Qué inferencias se utilizan en la evaluación del argumento?

P4. ¿Qué técnicas han utilizado por inferencia como respaldo?

P4. ¿Cómo se evalúa la validez global del argumento?

Estrategias de búsqueda

Previo a realizar la búsqueda sistemática se elaboraron fórmulas cortas a manera de exploración, tanto en español como en inglés. Las bases de datos utilizadas para este ejercicio fueron ERIC, Springer Link, Web of Science; elegidas por su relevancia global y su conexión

con el campo educativo. Con base en las RSL de Cook et al. (2014), Durson y Li (2021) y Lavery et al. (2020), cuyas revisiones concluyen en 2012 y 2017 respectivamente, se agregaron otros criterios de inclusión y exclusión tales como considerar sólo artículos publicados en el periodo 2014-2023 y que pertenezcan a revistas especializadas en áreas como Exámenes, Evaluaciones y Pruebas. Asimismo, se tomó en cuenta la revisión sistemática realizada por Durson y Li (2021), quienes se enfocaron específicamente en pruebas de idiomas, ya que este tipo de exámenes, aunque suelen asociarse a la lingüística aplicada, se consideran parte del área educativa porque evalúan habilidades fundamentales para el aprendizaje (Chapelle, 2021), como la Lectura, la expresión oral y escrita, y la competencia comunicativa.

Después se realizó una búsqueda a partir de operadores booleanos como AND y OR según la conveniencia de distintos términos. La fórmula final obtenida, en inglés y español, se define de la siguiente manera: author: (Chapelle) OR (Kane) abstract: (enfoque basado en argumentos) AND ((prueba) OR (examen) OR (prueba de aptitud)) AND descriptor: "validez" OR "validación". Esta fórmula se desarrolló en la página de PubMed (https://pubmed.ncbi.nlm.nih.gov/advanced/), a fin de que fuera funcional para diferentes bases de datos.

Tabla 82

Criterios de inclusión y exclusión

Criterio	Inclusión	Exclusión
Tipo de documento	Artículos en revistas especializadas (con ISBN o ISSN) en revistas indexadas nacionales e internacionales.	Libros o capítulos de libro; artículos de información; tesis de licenciatura, maestría y doctorado; actas de congresos; ponencias; noticias; blogs especializados; revisiones de libros
Subdisciplina	Exámenes, Pruebas y Evaluación (Springer Link)	otras subdisciplinas
Producción del autor(es)	Nacional e internacional	No aplica

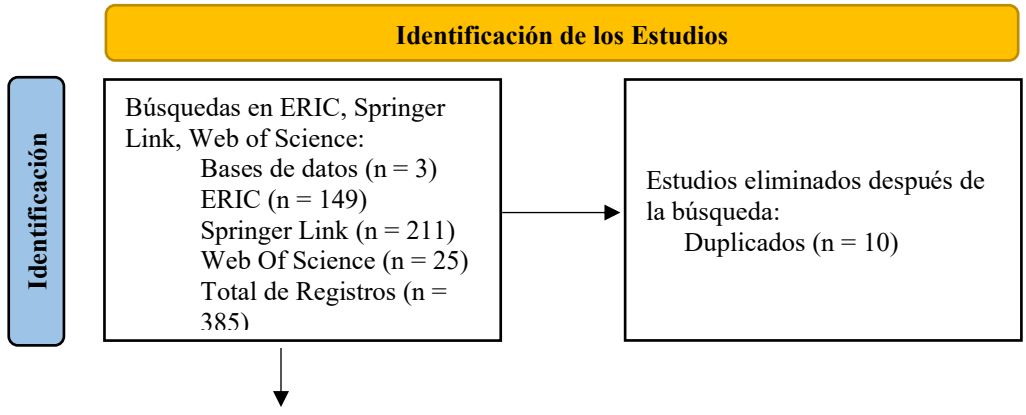
Periodo de publicación	2014-2023	Antes de 2014
Nivel educativo	Desde educación básica hasta posgrados	otros niveles educativos
Idioma	inglés, español	No aplica
Documento	Documento completo	Documento duplicado; sin resumen disponible
Disciplina	Educación	otras disciplinas
Accesibilidad	Acceso abierto/ privado	Ninguno

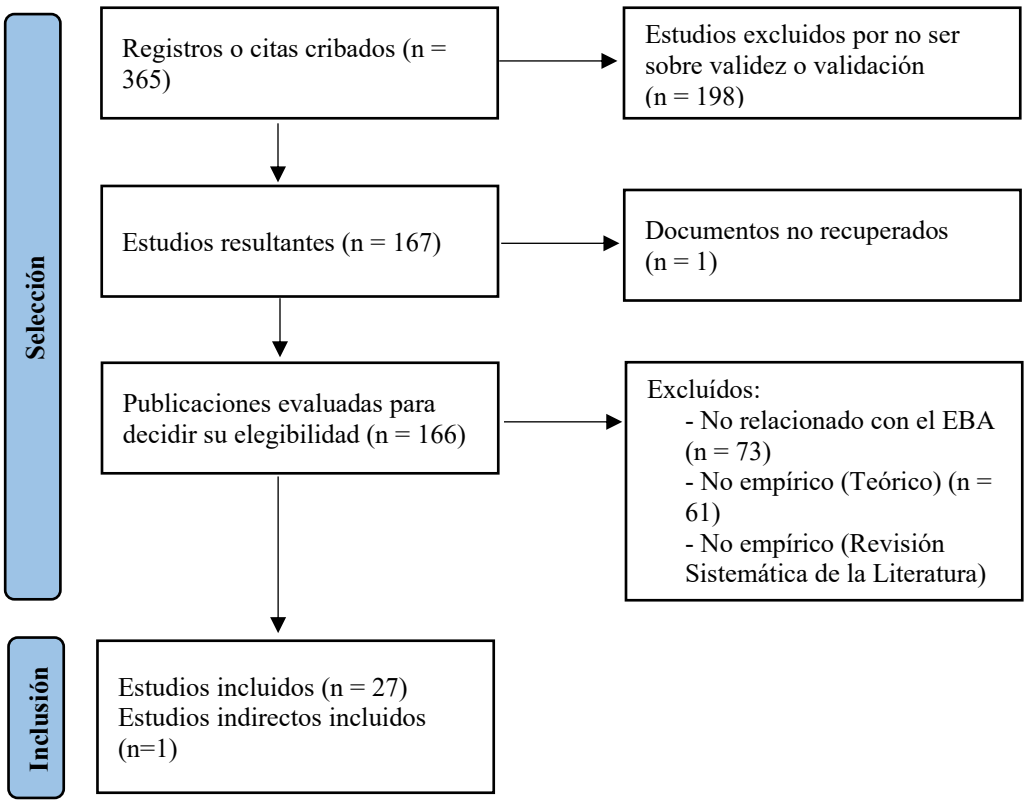
Revisión y selección

En la Figura 46 se observa el proceso de identificación, selección e inclusión de los artículos encontrados (Page et al., 2021). En este caso, hubo un total de 385 artículos provenientes de las bases de datos ya mencionadas. De los 385 se excluyeron 20 debido a la duplicación, quedando un total de 365 artículos; 198 no tuvieron relación con temas de validez o validación de la prueba o de exámenes; de los 167 restantes, un artículo no se recuperó. Los artículos restantes se sometieron a revisión como lo indica PRISMA (Page et al., 2021; Rethlefsen et al., 2021), 72 no tenían relación con el EBA, 61 eran teóricos o conceptuales, y 5 RSL. Con estas revisiones se ubicó la propuesta de Chapelle et al. (2015), por lo que se agregó al listado. El total de artículos incluidos fue de 28.

Figura 46

Diagrama de flujo de la RSL con PRISMA





Nota. Adaptado de Page et al. (2020). *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews*. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71. Distribuido bajo licencia CC BY 4.0.

Codificación y análisis de datos

En esta etapa se procedió a la lectura, revisión y análisis de los 28 artículos. Para la recopilación de los resultados se codificó sistemáticamente en una hoja de cálculo de Excel, donde se registraron los siguientes datos: Título de la publicación; Autor(es) y afiliación institucional; Fecha de publicación; Fuente de la publicación.

Apéndice B

Estructura Argumentativa con sus Fuentes de datos

Tabla 83

Estructura argumentativa con sus fuentes de datos

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

Inferencia	Garantía	Suposiciones	Fuentes de datos
Definición de Dominio El ExIES representa adecuadamente el contenido y las habilidades que se pretenden medir.	G1.1. Los contenidos del ExIES se seleccionaron para reflejar adecuadamente las áreas y habilidades identificadas como esenciales para el propósito del examen. G1.2. El ExIES asegura que los ítems de la prueba representan adecuadamente el dominio establecido en las especificaciones 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 le Evaluación de ítems del Nuevo Examen de Selección de aspirantes a ingresar a la Universidad Autónoma de Baja California (Caso & 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)
		S.1.2.1 Los ítems de la prueba se adhieren a las especificaciones que establecen las proporciones apropiadas de habilidades,	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

		conceptos y niveles de habilidad cognitiva requeridos.	de Resultados Rasch y estadísticas ítem–forma (2023s)
		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.	F1.2.1.2 Especificaciones de Lectura, Lengua Escrita y Matemática (Pedroza Zúñiga et al.,2023n, 2023o, 2023p);
		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.	F1.2.2-3 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)
Evaluación Las puntuaciones del ExIES son un resumen preciso del desempeño relevante en las tareas del examen.	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 ~~e~~l manejo del examen durante la 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

		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.	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 basado en sus respuestas. S2.2.2 Se realiza un procedimiento operativo después de la estimación Rasch por cada área.	F2.2.1.1 Técnica Rasch, Reporte Técnico 2023-1 (Pedroza Zúñiga et al., 2024a) 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)
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.	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)
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.	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 & 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)
Extrapolación Las habilidades evaluadas por el ExIES predicen	G5.1. El rendimiento en las habilidades de Lectura, escritura y Matemáticas evaluado por el ExIES se relaciona con	S5.1.1 Las puntuaciones del ExIES se correlacionan positivamente con indicadores de desempeño académico en la universidad para comprender y utilizar la Lengua	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***

adecuadamente el rendimiento futuro del estudiante en sus estudios de licenciatura.	indicadores futuros de rendimiento.	Escrita, las Matemáticas y su habilidad en Lectura, y otros indicadores relacionados con estas áreas.	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)
Utilización Las puntuaciones obtenidas en el ExIES son útiles para las universidades al tomar decisiones de admisión.	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)
Implicación de Consecuencias Las puntuaciones del ExIES resultan en consecuencias positivas para los solicitantes y la comunidad universitaria.	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. S7.1.2 Los puntajes del ExIES permiten clasificar de manera imparcial y útil a los aspirantes para asignar prelación de ingreso universitario y, combinados con el promedio de bachillerato, predicen con suficiencia el rendimiento del primer año.	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) F7.1.2.1 Resultados del análisis del DIF por sexo (Pedroza Zúñiga et al., 2025c) F7.1.2.2 Resultados de la inferencia de Extrapolación.

Nota. Elaboración propia, retomando la estructura de Chapelle (2021) y con tipos de evidencias (AERA et al.) posibles a utilizar según la inferencia y garantía.

Apéndice C

Guía para proponer Inferencias, Garantías, Supuestos y Tipos de Evidencia para la Validación de EAI según el EBA

La siguiente tabla sintetiza las siete inferencias del Enfoque Basado en Argumentos (EBA) para la validación de exámenes —propuesto por Chapelle (2021)— y organiza, para cada una, las garantías, los supuestos que las sustentan y los tipos de evidencia recomendados. Su propósito es ofrecer una guía operativa que permita a desarrolladores, investigadores y responsables de pruebas de alto impacto identificar qué condiciones deben verificarse y qué datos deben recogerse para construir un Argumento de Validez completo y coherente. La numeración jerárquica (G x.y / S x.y.z) facilita el rastreo lógico entre inferencias, garantías y supuestos, mientras que la última columna ilustra las fuentes empíricas típicas que respaldarían cada afirmación.

Tabla 84

Guía para proponer Inferencias, Garantías, Supuestos y Tipos de Evidencia para la Validación de Exámenes de Alto Impacto según el Enfoque Basado en Argumentos (Chapelle, 2021)

Inferencia (definición breve)	Garantías	Supuestos por garantía	Tipos de evidencia recomendada por Chapelle (2021)

1. Definición de dominio: El examen cubre de forma exhaustiva las habilidades y contenidos objetivos.	G1.1 Representatividad del contenido G1.2 Correspondencia con especificaciones G1.3 Relevancia y pertinencia G1.4 Integridad del desarrollo	*G1.1* → S1.1.1 Análisis de tareas concluyente; S1.1.2 Marco curricular aceptado; S1.1.3 Equilibrio de áreas. *G1.2* → S1.2.1 Tabla de especificaciones cubre el dominio; S1.2.2 Ítems redactados y revisados contra la tabla; S1.2.3 Actualización periódica. *G1.3* → S1.3.1 Revisión experta de sesgo; S1.3.2 Eliminación de contenido irrelevante. *G1.4* → S1.4.1 Elaboradores calificados; S1.4.2 Revisión multicapa; S1.4.3 Pilotaje psicométrico.	Análisis de tareas, marcos curriculares, tablas de especificaciones, paneles de jueces, manuales técnicos, informes de pilotaje.
2. Evaluación: Las puntuaciones resumen con precisión el desempeño observado.	G2.1 Administración estandarizada G2.2 Calificación exacta y objetiva G2.3 Seguridad e integridad G2.4 Selección de ítems basada en datos	*G2.1* → S2.1.1 Manuales para aplicadores; S2.1.2 Capacitación certificada; S2.1.3 Condiciones comparables. *G2.2* → S2.2.1 Protocolos de calificación validados; S2.2.2 Control-calidad (doble scoring); S2.2.3 Tecnología auditada. *G2.3* → S2.3.1 Procedimientos anti-fraude; S2.3.2 Monitoreo continuo. *G2.4* → S2.4.1 Banco calibrado (IRT/Rasch); S2.4.2 Rotación controlada.	Manuales de aplicación, registros de capacitación, estudios de confiabilidad inter-evaluador, reportes de seguridad, estadísticas del banco de ítems.
3. Generalización: Las puntuaciones serían similares en formas paralelas, poblaciones y ocasiones distintas.	G3.1 Consistencia interna G3.2 Equivalencia de formas G3.3 Estabilidad temporal G3.4 Precisión de estimación	*G3.1* → S3.1.1 Coeficientes α/ω aceptables. *G3.2* → S3.2.1 Formas equiparadas; S3.2.2 Muestra representativa en pilotaje. *G3.3* → S3.3.1 Diseño test–retest o G-Theory. *G3.4* → S3.4.1 SEM dentro de rangos.	Coeficientes de confiabilidad, estudios de equiparación, análisis test–retest, informes de error estándar e información IRT.

4. Explicación: Las puntuaciones reflejan el constructo teórico sin varianza irrelevante.	G4.1 Estructura interna acorde al modelo G4.2 Relaciones con otras variables G4.3 Equidad entre subgrupos G4.4 Evidencia cognitiva de procesos	*G4.1* → S4.1.1 Ajuste AFC/IRT satisfactorio; S4.1.2 Cargas y residuales adecuados. *G4.2* → S4.2.1 Correlaciones convergentes; S4.2.2 Correlaciones discriminantes bajas. *G4.3* → S4.3.1 Ausencia de DIF; S4.3.2 Sin sesgo cultural-lingüístico. *G4.4* → S4.4.1 Procesos (think-aloud, eye-tracking) alineados al constructo.	Resultados AFC/IRT, estudios de validez convergente/discriminante, análisis DIF, protocolos cognitivos y de procesos.
5. Extrapolación: El examen predice el desempeño futuro en el dominio real.	G5.1 Relación predictiva global G5.2 Clasificación y cortes fiables G5.3 Invarianza predictiva	*G5.1* → S5.1.1 Correlaciones/R^2 significativos con GPA u otros criterios. *G5.2* → S5.2.1 Curvas ROC aceptables; S5.2.2 Índice de consistencia de decisiones ≥ .80. *G5.3* → S5.3.1 Equidad de la relación predictiva entre subgrupos.	Estudios de regresión y correlación predictiva, análisis ROC, validación multigrupo de la ecuación predictiva.
6. Utilización: Las puntuaciones se usan de forma apropiada y beneficiosa.	G6.1 Comprensión por los usuarios G6.2 Políticas alineadas G6.3 Retroalimentación accionable G6.4 Supervisión ética y transparencia	*G6.1* → S6.1.1 Informes claros; S6.1.2 Recursos de interpretación. *G6.2* → S6.2.1 Regulaciones que definen usos; S6.2.2 Verificación periódica. *G6.3* → S6.3.1 Información diagnóstica relevante. *G6.4* → S6.4.1 Comité de gobernanza y auditorías externas.	Encuestas de usuarios, revisiones de políticas, estudios de impacto en admisión, actas de comités y auditorías sobre uso.
7. Implicación de consecuencias: El uso de las puntuaciones genera efectos positivos y minimiza los negativos.	G7.1 Mejora en decisiones G7.2 Wash-back positivo	*G7.1* → S7.1.1 Selección o certificación más acertada. *G7.2* → S7.2.1 Prácticas de enseñanza alineadas al constructo.	Estudios longitudinales de admisión y desempeño, investigaciones de wash-back, análisis de impacto, cronogramas y planes de mejora.

	G7.3 Impacto distributivo imparcial	*G7.3* → S7.3.1 Monitoreo y mitigación de impacto diferencial.
	G7.4 Mejora continua del examen	*G7.4* → S7.4.1 Ciclo sistemático de evaluación-revisión.

Nota. Elaboración propia basada en *Argument-based validation in testing and assessment* (Chapelle, 2021). *G* = garantía; *S* = supuesto; AFC = análisis factorial confirmatorio; DIF = diferencial en el funcionamiento de los ítems; IRT = Teoría de Respuesta al Ítem; ROC = receiver-operating characteristic; SEM = error estándar de medida; α/ω = coeficientes de confiabilidad (alfa de Cronbach y omega).

Apéndice D

Cambios en subcontenidos del ExIES

Tablas comparativas de cambios por área y versiones en las especificaciones

Tabla 85

Cambios en el dominio en subcontenidos: Lectura

Contenido	Subcontenido v2022	Subcontenido v2023
Información e ideas	Comprensión del propósito	Comprensión del propósito
	Determinación de significados explícitos	Determinación de significados explícitos
	Determinación de significados implícitos	Determinación de significados implícitos
	Uso de razonamiento analógico	Uso de razonamiento analógico
	Uso de respaldos o evidencias	Interpretación de palabras y frases en contexto
	Determinación de temas centrales	(Eliminado en v2023)
	Determinación de temas implícitos	(Eliminado en v2023)
	Análisis de relaciones entre ideas	(Eliminado en v2023)
	Interpretación de palabras y frases en contexto	(Eliminado en v2023)
Formas discursivas	Evaluación del tono textual	Evaluación del estilo
	Evaluación del estilo	Evaluación de argumentos explícitos
	Evaluación de argumentos	Evaluación de argumentos implícitos
	Coherencia lógica de las oraciones	Evaluación de coherencia entre ideas
Intertextualidad	Evaluación de textos múltiples	Análisis de textos múltiples
	Análisis de información cuantitativa	Análisis de información cuantitativa

Tabla 86

Cambios en el dominio en subcontenidos: Lengua Escrita

Contenido	Sub-contenido (Versión 2022)	Sub-contenido (Versión 2023)
Expresión de ideas escritas	Desarrollo: Revisión y edición de un texto	Uso de frases o palabras en oraciones
	Organización: Mejora de lógica y cohesión	Uso de conectores del español
	Uso efectivo del lenguaje: Precisión, estilo, tono	Uso de oraciones subordinadas
		Economía del lenguaje
		Uso efectivo de la pragmática
		Uso efectivo de la semántica: antónimos
		Uso efectivo de la semántica: sinónimos
Cumplimiento de reglas del español escrito	Estructura de oraciones: Corrección de problemas	Oraciones subordinadas
	Convenciones de uso: Gramática y uso	Concordancia entre sustantivo y adjetivo
		Concordancia entre sujeto y verbo

	Convenciones de puntuación: Adhesión a normas	Clases de palabras: adverbios
		Convenciones de puntuación: punto
		Convenciones de puntuación: coma
		Convenciones de puntuación: comillas
		Convenciones de puntuación: exclamación
		Convenciones de puntuación: interrogación
		Convenciones de puntuación: uso de paréntesis
		Ortografía: uso de la tilde

Tabla 87

Cambios en el dominio en subcontenidos: Lectura

Contenido	Sub-contenido (Versión ExIES 2020)	Sub-contenido (Versión ExIES 2024)
Herramientas algebraicas	Planteamiento de una ecuación lineal con una variable mediante un contexto	Plantear una ecuación lineal con una variable mediante un contexto.
	Solución de inecuaciones lineales con una variable	Solucionar inecuaciones lineales con una variable.
	Construcción de una función lineal que represente la relación lineal entre dos variables	Representar la relación entre dos variables para la construcción de una función lineal.
	Resolución de un sistema de ecuaciones lineales con tres variables.	Resolver un sistema de ecuaciones lineales con tres variables.
	Resolución de un sistema de ecuaciones lineales con dos variables	Sub-contenido eliminado dado que esto ya se evalúa dentro de otro
	Resolución de ecuaciones lineales en una variable	Resolver ecuaciones lineales con una variable.
	Resolución de sistemas de ecuaciones lineales con dos variables	Resolver sistemas de ecuaciones lineales con dos variables.
	Interpretación de las características de una función lineal dentro de un contexto.	Interpretar las características de una función lineal dentro de un contexto.
	Relación entre la representación gráfica y algebraica de una función lineal.	Relacionar la representación gráfica y algebraica de una función lineal.
Problemas, probabilidad y análisis de datos	Resolución de problemas utilizando índices, tasas, relaciones proporcionales y dibujos a escalas mediante uno o varios pasos.	Emplear índices, tasas, relaciones proporcionales o dibujos a escala para resolver problemas en uno o varios pasos.
	Resolución de problemas utilizando porcentajes con uno o varios pasos.	Utilizar porcentajes para resolver problemas en uno o varios pasos.
	Resolución de problemas utilizando diferentes magnitudes, y diferentes sistemas de unidades.	Equiparar diferentes unidades de medida para la solución de un problema.
	Análisis de variables involucradas en los diferentes modelos lineales, cuadráticos y exponenciales.	Sub-contenido eliminado dado que esto ya se evalúa dentro de otro
	Identificación de las características claves de un gráfico utilizando la relación entre las dos variables.	Identificar las características clave de un gráfico mediante la relación de dos variables.
	Calcular frecuencias relativas y probabilidades (sumativa y multiplicativa)	Calcular frecuencias relativas y probabilidades.

	Realizar inferencias a partir de los datos de una muestra.	Inferir a partir de los datos de una muestra (no utilizar tablas).
	Obtención de las medidas de tendencia central de datos y medidas de dispersión.	Obtener medidas de tendencia central y dispersión.
	Analizar reportes para hacer inferencias a partir de una tabla estadística.	Inferir a partir de los datos en una tabla.
Matemáticas avanzadas	Resolver problemas mediante funciones cuadráticas o exponenciales.	Resolver problemas mediante funciones cuadráticas o exponenciales.
	Traducirá contextos de lenguaje escrito a lenguaje algebraico.	Traducir del lenguaje escrito al lenguaje algebraico.
	Convertir expresiones algebraicas con exponentes racionales a radicales y viceversa.	Convertir expresiones algebraicas con exponentes racionales a radicales (o viceversa).
	Convertir ecuaciones de la forma ordinaria a la forma general y viceversa.	Convertir ecuaciones de la forma ordinaria a la forma general (o viceversa).
	Resolver ecuaciones cuadráticas.	Resolver ecuaciones cuadráticas.
	Simplificar operaciones aritméticas con polinomios.	Simplificar operaciones aritméticas con polinomios.
	Resolver ecuaciones radicales y racionales en una variable.	Resolver ecuaciones radicales y racionales en una variable.
	Resolución de un sistema de ecuaciones lineal y cuadrática	Resolver un sistema de ecuaciones lineal y cuadrática.
	Simplificación de expresiones algebraicas (fracciones) complejas	Simplificar expresiones algebraicas complejas (fracciones).
	Interpretación de parámetros, constantes o variables de una expresión no lineal en términos de un contexto dado	Interpretar algún parámetro, constante o variable de una expresión no lineal en términos de un contexto dado.
	Comprensión y determinación de ceros y factores de polinomios para la elaboración de gráficos	Determinar los ceros o factores de un polinomio por medio de un gráfico.
	Análisis de variables de expresiones algebraicas y su relación directa con el sistema de representación gráfico (sistemas de ecuaciones, descripción verbal del comportamiento gráfico, determinación de puntos importantes de una gráfica).	Analizar las variables de expresiones algebraicas y su relación directa con el sistema de representación gráfico (sistemas de ecuaciones, descripción verbal del comportamiento gráfico, determinación de puntos importantes de una gráfica).
	Uso de notación de funciones e interpretación del significado de dichas notaciones (evaluación de funciones).	Utilizar notación de funciones para describir un contexto.
	Comprensión de las formas general, estándar o canónica de expresiones algebraicas para identificar parámetros de interés (vértice, ordenada en el origen)	Identificar parámetros de interés de una expresión algebraica escrita en su forma general, estándar o canónica.
Temas adicionales en Matemáticas	Resolución de problemas que incluyan el cálculo de áreas o volúmenes de figuras geométricas	Calcular áreas o volúmenes de figuras geométricas.
	Uso de proporción trigonométrica y el Teorema de Pitágoras para resolver distintos problemas que consideran triángulos rectángulos.	Utilizar proporciones trigonométricas o Teorema de Pitágoras en triángulos rectángulos.
	Resolución de triángulos oblicuángulos mediante ley de senos y ley de cosenos	Emplear la ley de senos o ley de cosenos para determinar las medidas de un triángulo oblicuángulo.

	Conversión entre grados y radianes y uso de radianes para determinar la longitud del arco; usar funciones trigonométricas en escala de radianes	Determinar la longitud de arco.
	Aplicación de teoremas sobre círculos para encontrar la longitud del arco, medidas de ángulos, longitud de la cuerda y áreas de un sector	Aplicar teoremas sobre círculos para encontrar medidas de ángulos, longitud de una cuerda o el área de un sector.
	Uso de conceptos y teoremas sobre congruencia y similitud para resolver problemas sobre líneas, ángulos y triángulos	Emplear teoremas sobre congruencia y similitud para resolver problemas sobre líneas, ángulos o triángulos.
	Uso de la relación entre similitud, triángulo-rectángulo y proporciones trigonométricas; usar la relación entre seno y coseno de ángulos complementarios	Utilizar la similitud de triángulos-rectángulos y proporciones trigonométricas.
		Utilizar la relación entre seno y coseno de ángulos complementarios en triángulos rectángulos.
	Elaboración o uso de una ecuación en dos variables para resolver problemas sobre un círculo en un plano cartesiano	Emplear la ecuación de un círculo en el plano cartesiano dentro de un contexto.

Apéndice E

Ejemplo de Presentación de los Análisis Psicométricos ExIES 2023-1

Tabla 19

Análisis psicométricos de la Subversión 1: Lectura

Ítem	Dificultad	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Corr. Punto-biserial	Discriminación
1	0.46	1.01	0.2	1.01	0.1	0.25	0.95
2	0.35	1.01	0.1	0.99	0	0.21	0.99
3	0.44	0.97	-0.4	0.95	-0.5	0.31	1.11
4	0.49	1.01	0.2	1.01	0.1	0.26	0.96
5	0.49	1.06	0.9	1.07	0.8	0.19	0.72
6	0.49	0.99	-0.2	0.98	-0.2	0.29	1.07
7	0.43	1.11	1.3	1.17	1.4	0.1	0.66
8	0.54	1.09	1.2	1.12	1.2	0.14	0.65
9	0.45	1.01	0.1	1	0	0.25	0.97
10	0.55	1.15	2	1.18	1.8	0.06	0.44
11	0.51	1.01	0.2	1.01	0.1	0.25	0.95
12	0.63	1.15	1.2	1.29	1.5	0	0.74
13	0.47	0.9	-1.6	0.88	-1.5	0.41	1.45
14	0.48	0.93	-1.2	0.91	-1.2	0.37	1.35
15	0.5	1.04	0.7	1.06	0.8	0.21	0.77
16	0.39	1.02	0.1	1.05	0.3	0.21	0.96
17	0.41	0.94	-0.6	0.91	-0.7	0.34	1.15
18	0.49	1.06	1	1.07	0.8	0.18	0.7
19	0.43	0.98	-0.2	0.98	-0.2	0.28	1.05
20	0.38	0.9	-0.9	0.83	-1.1	0.39	1.18
21	0.56	0.98	-0.3	1	0	0.28	1.05
22	0.44	0.96	-0.5	0.95	-0.5	0.31	1.13
23	0.36	1	0	1.04	0.2	0.22	0.99
24	0.42	0.9	-1.2	0.87	-1.1	0.39	1.27
25	0.59	1.25	2.5	1.36	2.5	-0.1	0.38
26	0.56	1.05	0.7	1.07	0.7	0.18	0.81
27	0.39	1.02	0.2	1.07	0.4	0.21	0.95
28	0.35	0.89	-0.8	0.79	-1.1	0.39	1.16
29	0.53	1.1	1.4	1.11	1.3	0.13	0.59
30	0.53	1.01	0.2	1.02	0.3	0.25	0.94
31	0.46	0.89	-1.6	0.86	-1.6	0.42	1.42
32	0.63	1.08	0.6	1.18	0.9	0.1	0.85
33	0.51	0.98	-0.3	0.98	-0.3	0.29	1.08
34	0.49	1	0	1	0	0.27	1

35	0.49	0.97	-0.5	0.97	-0.4	0.31	1.14
36	0.53	0.96	-0.7	0.95	-0.6	0.33	1.19
37	0.49	0.91	-1.4	0.89	-1.4	0.39	1.43
38	0.59	0.96	-0.5	0.96	-0.3	0.3	1.09
39	0.53	0.86	-2.3	0.85	-1.9	0.46	1.6
40	0.53	1.1	1.5	1.11	1.3	0.13	0.55
41	0.51	1.04	0.6	1.05	0.6	0.21	0.8
42	0.44	0.91	-1.2	0.88	-1.3	0.39	1.31
43	0.44	0.94	-0.8	0.92	-0.8	0.35	1.21
44	0.4	0.95	-0.5	0.93	-0.5	0.31	1.1
45	0.59	1.14	1.5	1.23	1.6	0.04	0.65
46	0.55	1	-0.1	0.99	-0.1	0.27	1.02
47	0.65	0.96	-0.3	0.99	0	0.26	1.04
48	0.58	1.02	0.2	1.03	0.2	0.23	0.95
49	0.36	0.9	-0.8	0.81	-1.1	0.38	1.16
50	0.45	0.86	-2	0.82	-2	0.47	1.52

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

Apéndice F

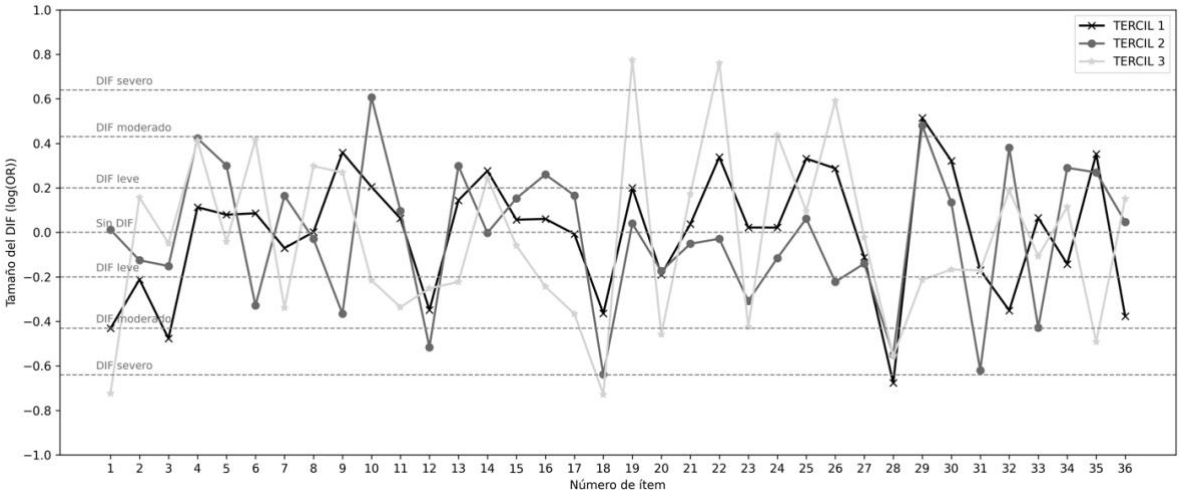
Análisis DIF por terciles

Lectura

En la Figura 47 se desglosa el análisis por nivel de rendimiento (tercil bajo, medio y alto). El gráfico evidencia un mayor número de ítems con log (OR) que alcanza niveles moderados o severos en algunos estratos de habilidad. Por ejemplo, en el tercil inferior surgen hasta 4 ítems con DIF relevante (3 moderados, 1 severo), mientras en el tercil superior aparecen 8 (5 moderados y 3 severos). Estas cifras destacan la importancia de examinar la estabilidad del DIF en distintas bandas de desempeño, pues ítems que eran "equilibrados" en el promedio global se tornan sesgados en un subgrupo específico.

Figura 47

DIF por tercil de habilidad y sexo en Lectura (Forma A)



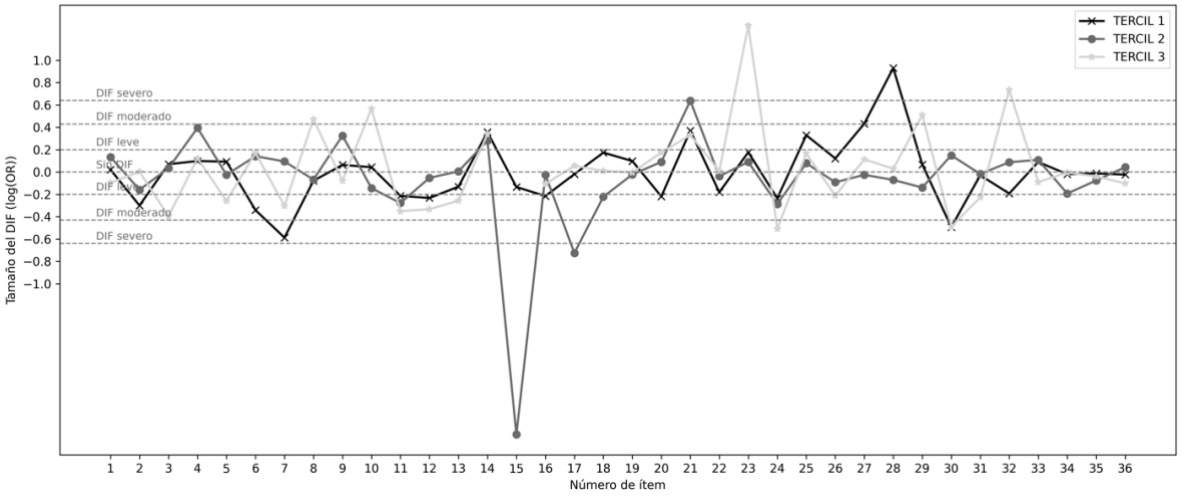
Nota. Reproducido de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 6, Figura 2).

La Figura 48 indica que también en la Forma C se incrementa el número de ítems con DIF moderado o severo al segmentar por tercil de habilidad. En el tercil inferior se identifican 4 ítems relevantes (3 moderados y 1 severo), en el tercil medio aparecen 3 ítems (1 moderado y 2

severos) y, en el tercil superior, otros 7 (5 moderados y 2 severos). Así, varios ítems, aun siendo "neutros" a nivel global, resultan más sencillos para un sexo en un nivel de habilidad particular.

Figura 48

DIF por tercil de habilidad y sexo en Lectura (Forma C)



Nota. Reproducido de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 7, Figura 4).

En la Tabla 88 se concentra la relación de todos los ítems —en ambas formas (A y C)— que superan los umbrales de ±0.43 o ±0.64 en cualquiera de los terciles. Cada ítem se codifica con símbolos que indican a qué grupo favorece (hombres o mujeres) y la intensidad de dicho DIF (moderado o severo). Por ejemplo, en la Forma A, el ítem 28 presenta un DIF severo (--) en el tercil 1 y, simultáneamente, un DIF moderado (-) o leve en otros niveles. El caso del ítem 18 también sobresale, pues exhibe DIF moderado o severo en más de un tercil. Estos resultados invitan a la revisión de aspectos como la redacción, el contenido o la exigencia cognitiva, para asegurar que la prueba sea equitativa en todos los rangos de desempeño.

Tabla 88

Ítems con DIF moderado o severo en Lectura (formas A y C)

Forma	Item	Tercil 1	Tercil 2	Tercil 3	Forma	Item	Tercil 1	Tercil 2	Tercil 3

A	1	-		--	**C**	7	-		
	3	-				8			+
	10		+			10			+
	12		-			15		--	--
	18		-	--		17		--	
	19			++		21		+	
	20			-		23			++
	22			++		24			-
	24			+		27	+		
	26			+		28	++		
	28	--	-	-		29			+
	29	+	+			30	-		-
	31		-			32			++
	35			-					

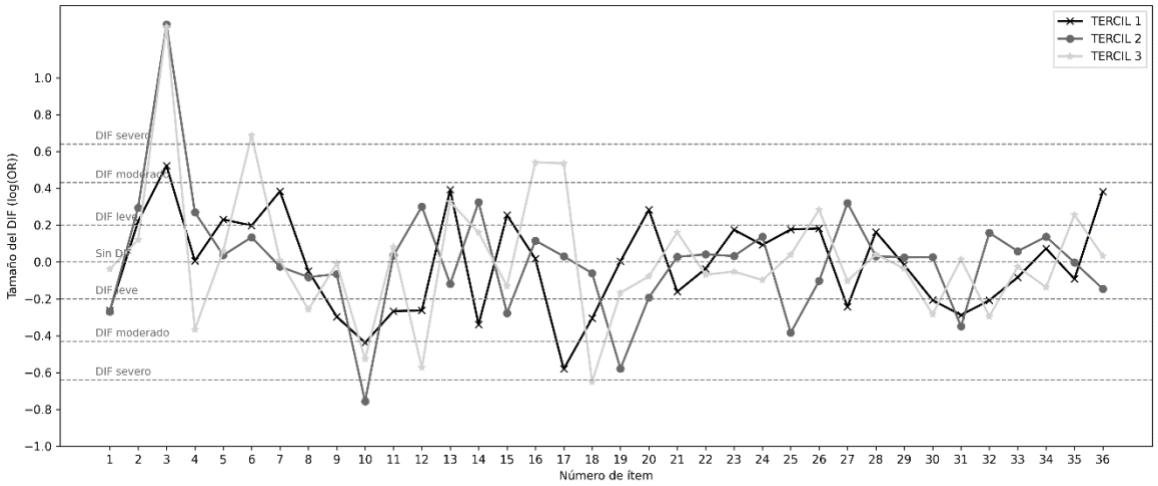
Nota. + = DIF moderado a favor de hombres, ++ = DIF severo a favor de hombres, − = DIF moderado a favor de mujeres, −− = DIF severo a favor de mujeres. Adaptada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (P Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 8, Tabla 1).

Lengua Escrita

Tal como en el caso anterior, véase ahora la Figura 49, se expone el análisis del DIF segmentado por terciles de habilidad para Lengua Escrita en la Forma A. A diferencia de lo reportado en la Figura 5, donde los ítems presentaban un comportamiento similar entre hombres y mujeres de manera global, el análisis por niveles de rendimiento revela diferencias más notorias. En el tercil inferior, tres ítems (3, 10 y 17) presentaron DIF moderado; en el tercil medio, uno mostró DIF moderado (19) y dos alcanzaron un DIF severo (3 y 10); mientras que en el tercil superior se identificaron cuatro ítems con DIF moderado (10, 12, 16 y 17) y dos con DIF severo (3 y 6). Destaca especialmente el ítem 3, que en el análisis global se ubicaba cerca del límite de DIF moderado, pero que al analizarse por nivel de habilidad evidencia un sesgo mucho más marcado, siendo el de mayor magnitud entre los ítems evaluados en esta figura.

Figura 49

DIF por tercil de habilidad y sexo en Lengua Escrita (Forma A) del ExIES 2023-2

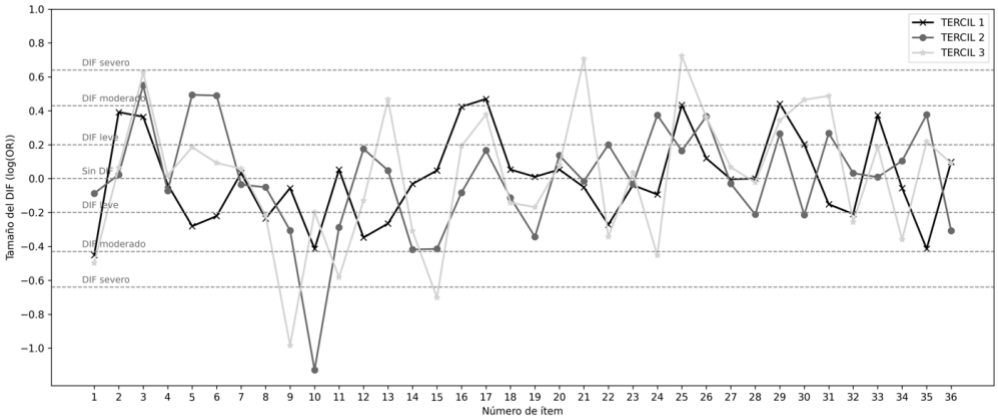


Nota. Reproducido de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 10, Figura 7).

Por otro lado, la Figura 50 muestra el análisis DIF por sexo para Lengua Escrita en la Forma C. Visualmente, la mayoría de los ítems se agrupan cerca del eje central (log (OR) = 0), reflejando un predominio de DIF leve. Ninguno de los ítems supera los umbrales establecidos para DIF moderado o severo según los criterios de Zumbo (1999). En consecuencia, se concluye que el comportamiento de los ítems fue homogéneo entre hombres y mujeres, sin diferencias relevantes que pudieran afectar la equidad de la evaluación.

Figura 50

DIF por tercil de habilidad y sexo en Lengua Escrita (Forma C)



Nota. Reproducido de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 11, Figura 8).

Este listado, de la Tabla 89, resume los ítems de ambas formas que exhiben valores fuera del rango de DIF leve en cualquiera de los tres terciles. Por ejemplo, el ítem 3 (Forma A) aparece con +, ++, ++, lo que sugiere que en diferentes rangos de habilidad mantiene un sesgo a favor de hombres y que este se intensifica según el tercil. Asimismo, en la Forma C se registran casos como el ítem 9, con DIF severo (--) en terciles medios y altos, reforzando la posible ventaja para el grupo femenino en esos niveles de desempeño.

Tabla 89

Ítems con DIF moderado o severo en Lengua Escrita (formas A y C)

Forma	Item	Tercil 1	Tercil 2	Tercil 3	Forma	Item	Tercil 1	Tercil 2	Tercil 3
A	3	+	++	++	C	1	-		-
	6			++		3		+	+
	10	-	--	-		5		+	
	12			-		6		+	
	16			+		9			--
	17	-		+		10		--	
	18			--		11			-
	19		-			13			+
						15			--
						17	+		
						21			++
						24			-
						25	+		++
						29	+		
						30			+
						31			+

Nota. + = DIF moderado a favor de hombres, ++ = DIF severo a favor de hombres, − = DIF moderado a favor de mujeres, −− = DIF severo a favor de mujeres. Adaptada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 12, Tabla 2).

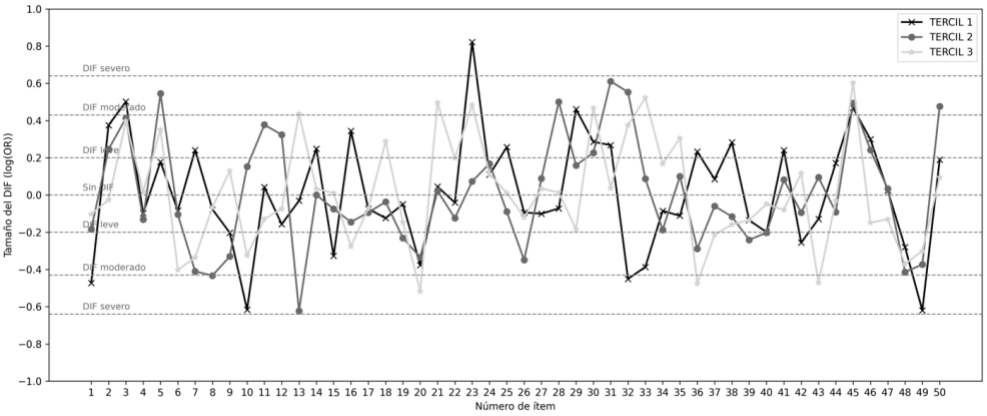
Matemáticas

Sin embargo, la situación se matiza al revisar el comportamiento en terciles de habilidad. De hecho, en el tercil inferior aparecen 8 ítems con DIF (7 moderados y 1 severo), mientras en el tercil medio hay 8 moderados, y en el tercil superior se identifican 9 (varios moderados, algunos severos). Por ejemplo, el ítem 23 pasa de ser no significativo al tener un DIF severo (++) en el

tercil bajo y un DIF moderado (+) en el tercil alto, lo que apunta a una inestabilidad sustancial según el nivel de logro del examinado.

Figura 51

DIF por tercil de habilidad y sexo en Matemáticas (Forma A)

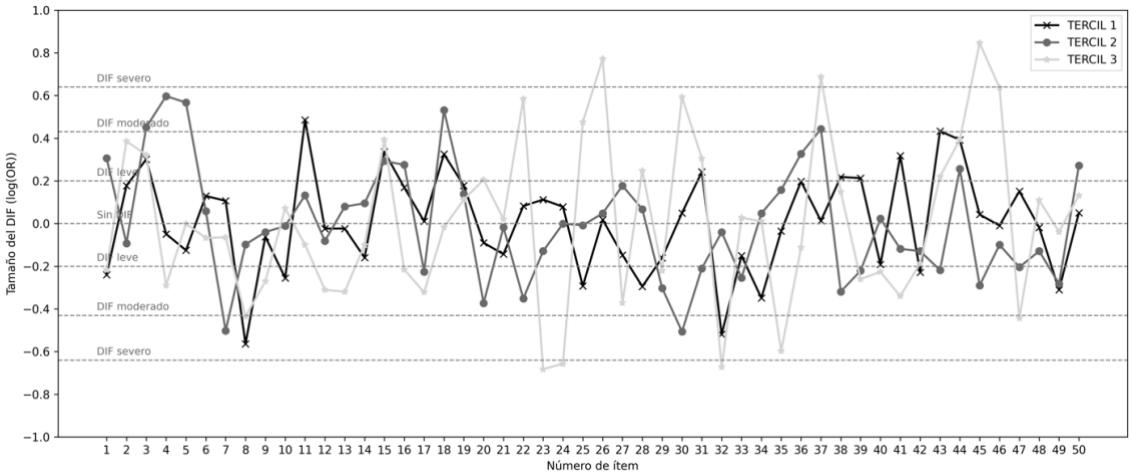


Nota. Reproducido de Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2 (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 14, Figura 10).

Asimismo, en la Figura 52, sobre el funcionamiento diferencial por tercil en la Forma C, al dividir la muestra en tercil bajo, medio y alto, surgen diferencias relevantes: el tercil inferior registra 4 moderados, el tercil medio 7 (todos moderados), y el tercil superior llega a 12 (7 moderados y 5 severos). Especial atención merecen los ítems 23, 24, 26, 37 y 45, todos ellos con DIF severo (--) o (++) en el tercil superior, poniendo en evidencia un posible impacto sobre la imparcialidad de la medición para sustentantes de alto rendimiento en un sexo determinado.

Figura 52

DIF por tercil de habilidad y sexo en Matemáticas (Forma C)



Nota. Reproducido de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 12, Figura 15).

Finalmente, la Tabla 90 reúne los ítems con log (*OR*) superior a ±0.43 (moderado) o ±0.64 (severo) en uno o más terciles, destacando a qué grupo favorecen. En la Forma A, ítems como el #23 saltan a la vista por mantener un carácter severo (++) en el tercil 1 y moderado (+) en el 3, mientras que el ítem #45 aparece repetidamente con + (moderado) en todos los terciles. De manera análoga, en la Forma C resaltan los ítems 23 y 24 con DIF severo (--) a favor de mujeres en el tercil superior. Estos datos concretos proveen guías claras para intervenir en la construcción de los ítems y asegurar que la prueba no penalice o beneficie injustamente a ninguno de los dos sexos.

Tabla 90

Ítems con DIF moderado o severo en Matemáticas (formas A y C)

Forma	Item	Tercil 1	Tercil 2	Tercil 3	Forma	Item	Tercil 1	Tercil 2	Tercil 3
	1	-				3		+	
	3	+				4		+	
	5		+			5		+	
A	8		-		**C**	7		-	
	10	-				8	-		-
	13		-	+		11	+		
	20			-		18		+	

	21			+		22			+
	23	++		+		23			--
	28		+			24			--
	29	+				25			+
	30			+		26			++
	31		+			30		-	+
	32	-	+			32	-		--
	33			+		35			-
	36			-		37		+	++
	43			-		43	+		
	45	+	+	+		45			++
	49	-				46			+
	50		+			47			-

Nota. + = DIF moderado a favor de hombres, ++ = DIF severo a favor de hombres, − = DIF moderado a favor de mujeres, −− = DIF severo a favor de mujeres. Adaptada de *Funcionamiento diferencial del ítem (DIF): Examen de ingreso a la educación superior (ExIES) 2023-2* (Pedroza Zúñiga & Gómez Monárrez, 2025c, p. 16, Tabla 3).

Apéndice G

Estudio sobre Modelos de Regresión: Método

Antecedentes

Para comprender mejor las capacidades predictivas de diferentes modelos en el ámbito del rendimiento académico, se realizó una Revisión Sistemática de la Literatura (RSL) para explorar estudios previos relacionados con la validez predictiva de las pruebas de admisión, además se incluyó estudios relevantes sobre aprendizaje automático, regresión y redes neuronales, siguiendo el enfoque metodológico propuesto por García Peñalvo (2019). La revisión sistematizada que se llevó a cabo integró metodologías y lineamientos de Kitchenham y Charters (2007), los lineamientos PRISMA de Page et al. (2021) que incluyó el análisis de estudios relevantes sobre aprendizaje automático, regresión y redes neuronales. Con el fin de encontrar artículos sobre Machine Learning, se separaron dos fórmulas: “TITLE-ABS-KEY ("predictive validity" OR "academic performance" OR "university achievement") AND ("entrance exams" OR "admission tests") AND ("Ridge regression" OR "XGBoost") AND ("GPA" OR "high school grades")”; y “("validez predictiva" OR "evidencias de validez predictiva" OR "predictive validity") AND ("pruebas de admisión a la universidad" OR "college entrance examinations" OR "examen de ingreso a nivel superior" OR "higher education admission test" OR "college admission test")”.

Tabla 91

Comparativa de estudios y modelos predictivos

Artículo	Modelo Utilizado	Variables Predictoras	Coeficiente de Determinación (R^2)	Resultados Clave

Palmer & Rolf (2023)	Regresión múltiple, bivariado	GPA, exámenes de ingreso, calificaciones previas	0.268	Calificaciones de farmacología mejor predictor de éxito en cursos del primer semestre.
Alotaibi, N. (2021)	Regresión	GPA, Test de admisión (STEP), Tipo de escuela	0.015	La prueba de admisión fue el mejor predictor de rendimiento académico.
García (2016)	HLR (Hierarchical Linear Regression)	PAA (COLLEGE BOARD), EXANI II	0.195 - 0.230	El EXANI II presentó un mayor coeficiente de determinación comparado con PAA
Alnahdi (2015)	Regresión	GAT	0.08	Moderada capacidad predictiva del rendimiento académico
Santelices y Wilson (2015)	Regresión	SAT Verbal (Subtest)	0.098 - 0.215	Diferentes coeficientes de determinación para grupos étnicos (AAS, HS, WS)
Sartania et al. (2014)	Regresión	UKCAT	0.063	Moderada capacidad predictiva del rendimiento académico en la escuela médica
Poole et al. (2012)	Regresión múltiple	UMAT	0.097	Moderada capacidad predictiva para el promedio del año 2
Rojas (2013)	Regresión múltiple	PAA	0.243	Capacidad predictiva moderada sobre el rendimiento académico (Promedio segundo año)
Higdem et al. (2016)	Regresión	SAT	0.172	Capacidad predictiva moderada para el rendimiento académico
García (2016)	Regresión	PAA (COLLEGE BOARD), EXANI II	0.195 - 0.230	Coeficiente de determinación para el rendimiento académico de estudiantes universitarios
Tapasco et al. (2016)	Regresión múltiple	SABERONCE	0.148 - 0.214	Capacidad predictiva moderada del rendimiento académico
Shulruf et al. (2012)	Regresión	UMAT1, UMAT2, UMAT3	0.05	Capacidad predictiva moderada para el rendimiento académico en el segundo año académico

Los resultados de la tabla comparativa muestran que los modelos de regresión, como la regresión múltiple y la regresión lineal, presentaron capacidades predictivas variables con coeficientes de determinación que iban desde 0.05 hasta 0.7865. Los modelos de aprendizaje automático, como Random Forest, demostraron una mejor capacidad predictiva con un coeficiente de determinación de hasta 0.80. Estos hallazgos sugieren que los enfoques basados en aprendizaje automático, aunque no siempre superiores, pueden ofrecer ventajas significativas en términos de precisión al predecir el rendimiento académico, en comparación con los métodos tradicionales de regresión. Además, estudios recientes han demostrado que la combinación de diferentes herramientas de evaluación, como exámenes estandarizados y promedios de calificaciones, mejora la capacidad predictiva, lo cual es esencial para optimizar los procesos de admisión y seguimiento académico de los estudiantes.

En este sentido, el contexto que nos acoge es el de la Universidad Autónoma de Baja California (UABC), la cual aplica el Examen de Ingreso a la Educación Superior (ExIES) para procesos de admisión a la universidad y cuya validez como predictor del rendimiento académico podría ampliarse si se utilizara en conjunto con otras métricas, como el promedio bachillerato. Aunque actualmente la UABC emplea exclusivamente el ExIES como criterio principal de admisión, los resultados de este análisis sugieren que integrar el promedio bachillerato podría fortalecer su capacidad predictiva, en línea con lo señalado por García (2016), quien encontró que el EXANI II combinado con otras métricas, como el promedio de calificaciones previas, mejora sustancialmente la predicción del desempeño académico en la educación superior.

Asimismo, Alotaibi (2021) destaca que los promedios académicos y las pruebas de admisión son predictores moderadamente efectivos, pero que su combinación puede ser clave para identificar estudiantes con mayor precisión, especialmente en el primer año universitario,

donde los desafíos académicos son mayores. De manera similar, Santelices y Wilson (2015) y Tapasco et al. (2016) encontraron que las pruebas estandarizadas, como el SAT y el SABERONCE, tienen una capacidad predictiva limitada cuando se utilizan de manera aislada, pero su poder explicativo aumenta considerablemente cuando se complementan con promedios escolares. Estos hallazgos refuerzan la importancia de integrar múltiples indicadores académicos en el análisis predictivo para proporcionar una evaluación más completa y efectiva de los estudiantes.

El contexto de la UABC, en el cual el ExIES es el único criterio de selección, presenta una oportunidad para explorar nuevas combinaciones de 7ivariables predictoras, tal como se ha demostrado en estudios previos. Por ejemplo, Rojas (2013) evidenció que el uso del PAA junto con el promedio bachillerato permitió predecir con mayor precisión el rendimiento académico de estudiantes universitarios en su segundo año. Este enfoque es particularmente relevante dado que el primer año es crucial para el éxito académico a largo plazo, como lo señalan Shulruf et al. (2012), quienes subrayaron que los predictores académicos, especialmente aquellos que integran tanto habilidades estandarizadas como antecedentes académicos previos, son esenciales para diseñar intervenciones efectivas y mejorar la retención estudiantil.

En este sentido, este análisis propone la posibilidad de combinar los resultados del ExIES con el promedio bachillerato como parte de un enfoque predictivo integral, que no solo ayude a mejorar los procesos de admisión, sino que también permita identificar con mayor precisión a los estudiantes en riesgo de bajo rendimiento. Tal propuesta no solo se alinea con los hallazgos de estudios como el de García (2016) y Tapasco et al. (2016), sino que también responde a la necesidad de desarrollar modelos predictivos más robustos que capten las complejas dinámicas que influyen en el éxito académico durante el primer año universitario.

Método

El objetivo de este estudio es evaluar la capacidad predictiva de diferentes modelos de regresión para inferir el rendimiento académico en el primer año universitario, a partir de los predictores: los resultados del ExIES y el promedio bachillerato. El fin es alcanzar un coeficiente de determinación R^2 de al menos 0.15, lo cual indicaría que al menos el 15% de la variabilidad en el rendimiento académico es explicada por los predictores seleccionados.

Datos y Variables

Se empleó una base de datos con 10,184 registros, proporcionada por la UABC, así como por la Dirección de la EMS Baja California. Las variables predictoras fueron los puntajes de Lenguaje (PuntajeL), Matemáticas (PuntajeM), Lengua Escrita (PuntajeE), el puntaje global del ExIES y el promedio final de bachillerato. La variable criterio, Año1, corresponde al promedio obtenido por cada estudiante en su primer año universitario. En la Tabla 92 se describe el conjunto de datos y las transformaciones realizadas para prepararlos para el análisis.

Tabla 92

Conjunto de datos y transformaciones

Variable	Tipo	Descripción	Transformación Realizada
PuntajeL	Numérica	Puntaje obtenido en la sección de Lenguaje del ExIES.	Escalado con StandardScaler.
PuntajeM	Numérica	Puntaje obtenido en la sección de Matemáticas del ExIES.	Escalado con StandardScaler.
PuntajeE	Numérica	Puntaje obtenido en la sección de Lengua Escrita del ExIES.	Escalado con StandardScaler.
PromedioBach_Sistemas	Numérica	Promedio final del bachillerato.	Escalado con StandardScaler.
Año1 (etiqueta)	Numérica	Promedio obtenido en el primer año universitario.	Variable objetivo del modelo. Sin transformación, conserva formato decimal.

En la Tabla 93 se detallan los diferentes modelos de regresión lineal empleados en el análisis, diferenciados por las combinaciones de variables predictoras utilizadas. Estos modelos fueron diseñados con el propósito de evaluar el aporte individual y conjunto de las distintas variables en la predicción del promedio del primer año universitario (Año1). Se incluyeron variantes que combinan puntajes del ExIES (Lenguaje, Matemáticas y Lengua Escrita), el puntaje global del examen, y el promedio bachillerato, así como versiones que consideran únicamente una o dos de estas Fuentes de datos. Esta estrategia permite comparar el poder explicativo de cada conjunto de variables y determinar qué combinaciones ofrecen un mejor desempeño predictivo.

Tabla 93

Modelos empleados según las variables incluidas

Modelo	Variables Incluidas
Regresión Lineal (Básico)	PuntajeL, PuntajeM, PuntajeE, PromedioBach_Sistemas
Regresión Lineal (Examen)	PuntajeL, PuntajeM, PuntajeE
Regresión Lineal (Puntaje Global)	PuntajeGlobal
Regresión Lineal (Promedio Bach)	PromedioBach_Sistemas
Regresión Lineal (Global-Bach)	PuntajeGlobal, PromedioBach_Sistemas
Regresión Lineal (Puntaje LM)	PuntajeL, PuntajeM, PromedioBach_Sistemas
Regresión Lineal (Puntaje LB)	PuntajeL, PromedioBach_Sistemas
Regresión Lineal (Puntaje MB)	PuntajeM, PromedioBach_Sistemas
Regresión Lineal (Puntaje EMB)	PuntajeE, PuntajeM, PromedioBach_Sistemas
Regresión Lineal (Puntaje EB)	PuntajeE,

Técnicas

Para evaluar la inferencia de Extrapolación, declarado en la Tabla 72, se emplearon modelos de regresión lineal, tanto simples como múltiples, debido a su capacidad para identificar relaciones predictivas entre las variables independientes (puntajes del ExIES, promedio bachillerato y otras variables contextuales) y la variable dependiente (promedio académico del primer año universitario). Estas técnicas son ampliamente utilizadas en estudios de validez predictiva porque permiten estimar el grado en que los resultados de una prueba explican el

desempeño académico futuro (Montgomery et al., 2012). Se utilizaron las siguientes técnicas específicas:

- Regresión lineal simple: Para evaluar el impacto individual de cada variable predictora en el rendimiento académico del primer año universitario.
- Validación cruzada: Se implementó validación cruzada con 5 particiones para garantizar la robustez y Generalización de los modelos, minimizando el riesgo de sobreajuste (James et al., 2021).
- Coeficiente de determinación *(R^2):* Para medir el porcentaje de la varianza en la variable dependiente explicado por los predictores.
- Error cuadrático medio *(MSE) y raíz del error cuadrático medio (RMSE):* Para evaluar la precisión de las predicciones realizadas por los modelos.
- Análisis de multicolinealidad*:* Se calculó el Variance Inflation Factor (VIF) para asegurar que las variables independientes no presentaran una correlación excesiva entre sí, evitando problemas en las estimaciones de los coeficientes.

Además de evaluar el efecto de un puntaje global, se consideró la conveniencia de tratar por separado los puntajes de las distintas secciones (Lenguaje, Matemáticas y Lengua Escrita). Autores como Cronbach y Meehl (1955) y Kane (2002) subrayan que la validez de constructo se fortalece cuando cada dimensión evaluada corresponde a un dominio de habilidades diferenciado, lo que facilita interpretar la información de manera más específica. Asimismo, en la literatura de predicción del rendimiento académico universitario, se han documentado beneficios de analizar subpuntuaciones de exámenes de admisión o diagnósticos (p. ej., Saks, 2024; Santelices & Wilson, 2015; Sartania et al., 2014; Tapasco et al., 2016), pues cada área puede contribuir de manera distinta al desempeño posterior según la carrera o contexto.

Preprocesamiento de datos

1. Conversión y manejo de datos faltantes: Las columnas de puntajes (PuntajeL, PuntajeM, PuntajeE) y el promedio bachillerato fueron convertidas a formato numérico y se imputaron valores faltantes con la media, lo que permitió mantener el tamaño del dataset y evitar la pérdida de registros. Según Gerón (2019), la imputación es preferible a la eliminación de filas cuando la proporción de valores nulos es baja.
2. Análisis de distribuciones y normalización: Se utilizaron histogramas y curvas KDE para inspeccionar la distribución de cada variable (véase Figura 55). Esto permitió identificar algunas distribuciones sesgadas, pero no se aplicaron transformaciones adicionales (e.g., logarítmicas), ya que la regresión lineal no requiere normalidad estricta.
3. Escalado de variables: Dado que los algoritmos de regresión lineal son sensibles a las magnitudes de las variables, todas las columnas numéricas fueron escaladas utilizando StandardScaler. Esto asegura que los coeficientes del modelo sean interpretables y estén en la misma escala.
4. Manejo de outliers: Se identificaron valores extremos mediante el Z-score (>3 desviaciones estándar) y se eliminaron para evitar sesgos en el modelo, ya que los outliers pueden distorsionar las predicciones y reducir la capacidad generalizadora del modelo.
5. Revisión de multicolinealidad: Se calculó el Variance Inflation Factor (VIF) para identificar correlaciones altas entre las variables predictoras (véase Tabla 95). Aunque se encontró cierta correlación entre los puntajes del ExIES, su inclusión conjunta fue justificada por la capacidad explicativa adicional que aportan al modelo.

Tabla 94

Proceso de preprocesamiento aplicado a los datos del ExIES

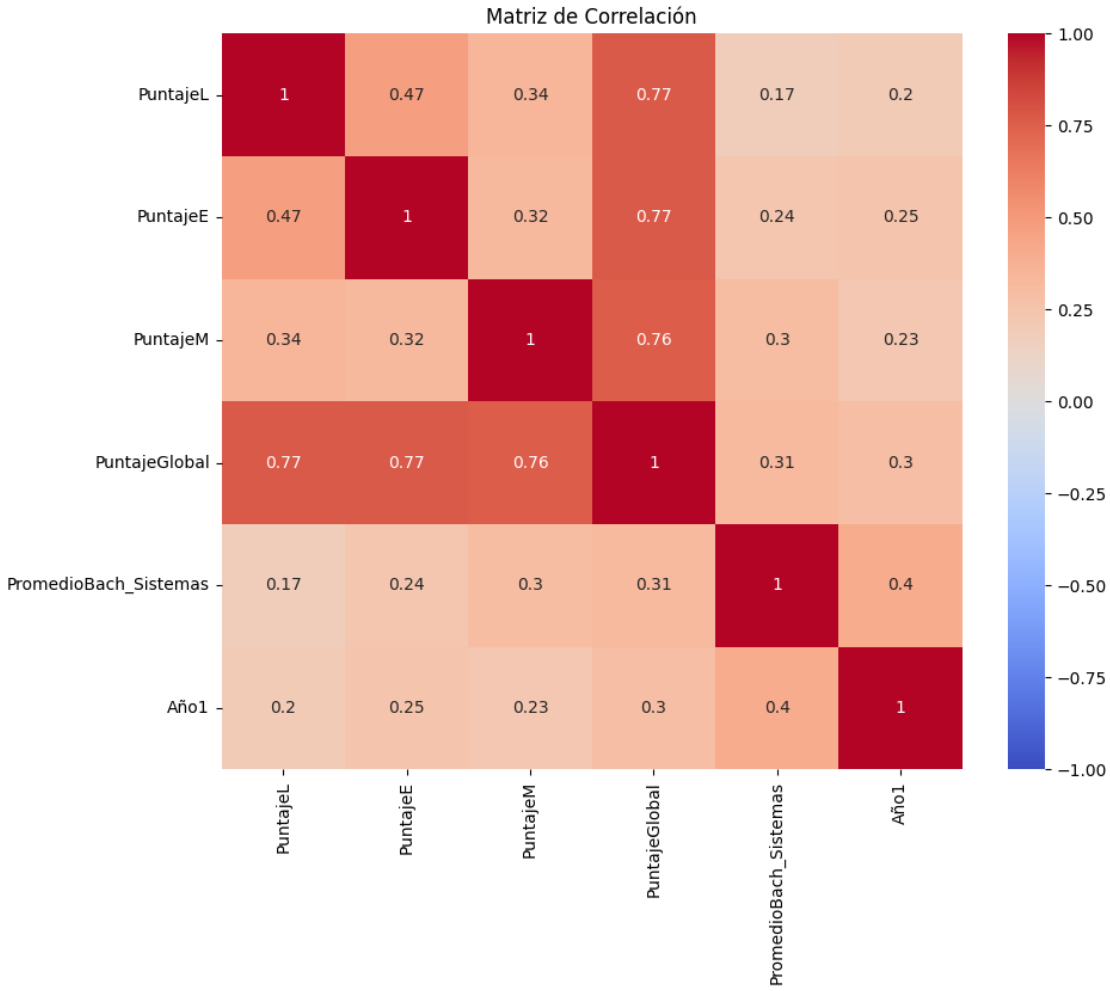
Etapa	Variable(s)	Descripción del problema	Acción realizada	Justificación
Conversión de tipos de datos	PuntajeL, PuntajeM, PuntajeE, PuntajeFinal	Estas columnas estaban clasificadas como *object* en lugar de numéricas.	Convertidas a tipo float utilizando pd.to_numeric.	Garantiza que las variables puedan ser utilizadas en operaciones Matemáticas y modelos predictivos.
Valores faltantes	PuntajeL, PuntajeM, PuntajeE, PromedioBach_Sistemas	Presencia de valores nulos en columnas predictoras clave.	Imputación con la media para cada columna.	Evita la pérdida de registros al eliminar valores faltantes y preserva la distribución general de los datos.
Escalado de variables	PuntajeL, PuntajeM, PuntajeE, PromedioBach_Sistemas	Diferencias en las escalas de las variables numéricas que podrían desbalancear el impacto de cada predictor.	Escalado en el rango [0, 1] utilizando StandardScaler de sklearn.preprocessing.	Garantiza que todas las variables tengan un impacto proporcional en el modelo, evitando que valores con escalas mayores dominen la optimización.
Eliminación de outliers	PuntajeL, PuntajeM, PuntajeE, PromedioBach_Sistemas	Valores extremos identificados como fuera de 3 desviaciones estándar, lo cual afecta la precisión del modelo.	Eliminación utilizando el Z-score.	Los outliers extremos pueden influir desproporcionadamente en el ajuste del modelo, reduciendo su capacidad de Generalización.
Revisión de multicolinealidad	PuntajeL, PuntajeM, PuntajeE	Alta correlación entre las variables predictoras podría generar redundancia y problemas numéricos.	Cálculo del Variance Inflation Factor (VIF) para detectar colinealidad.	Reducir la redundancia entre predictores mejora la estabilidad y precisión del modelo, especialmente en regresión lineal.

El análisis de multicolinealidad mediante el cálculo del *VIF* confirmó que las variables seleccionadas no presentan problemas significativos de colinealidad. La Figura 53 presenta una Tabla de correlaciones en la que los colores reflejan la intensidad y dirección de las relaciones entre las variables; los tonos más cercanos al rojo indican correlaciones positivas más fuertes, mientras que los tonos cercanos al azul describen correlaciones negativas; en tanto, valores próximos a cero evidencian escasa relación.

Sumado a lo anterior, la Tabla 95 evidencia que todas las variables presentan valores de VIF claramente inferiores a 3, lo cual se ubica muy por debajo del umbral de 10 considerado como indicativo de multicolinealidad alta. En la literatura metodológica general (Gerón, 2019; Burkov, 2019) se sugiere que un Valor de VIF que supere ciertos umbrales (por ejemplo, 5 o 10) puede indicar colinealidad problemática. En concreto, las variables asociadas a los puntajes de ingreso (PuntajeL, PuntajeE y PuntajeM) muestran valores cercanos a 1.3, lo que sugiere que cada una contribuye de manera diferenciada a la explicación de la variable dependiente. Asimismo, el promedio bachillerato exhibe un *VIF* cercano a 1.1, reforzando la conclusión de que no existen problemas de colinealidad significativos. De esta forma, si bien podrían existir correlaciones moderadas entre los predictores, no alcanzan un nivel que comprometa la estabilidad del modelo; en consecuencia, cada variable mantiene su relevancia y aporta información única para la predicción.

Figura 53

Matriz de correlación entre variables predictoras utilizadas en el modelo



Nota. La figura muestra coeficientes de correlación de Pearson entre las variables predictoras. Los valores cercanos a 1 indican una correlación positiva fuerte; los cercanos a 0 indican una relación débil o nula. El gradiente de color representa la magnitud de la correlación. Elaboración propia.

Tabla 95

Valores del índice de inflación de la varianza para variables predictoras del modelo

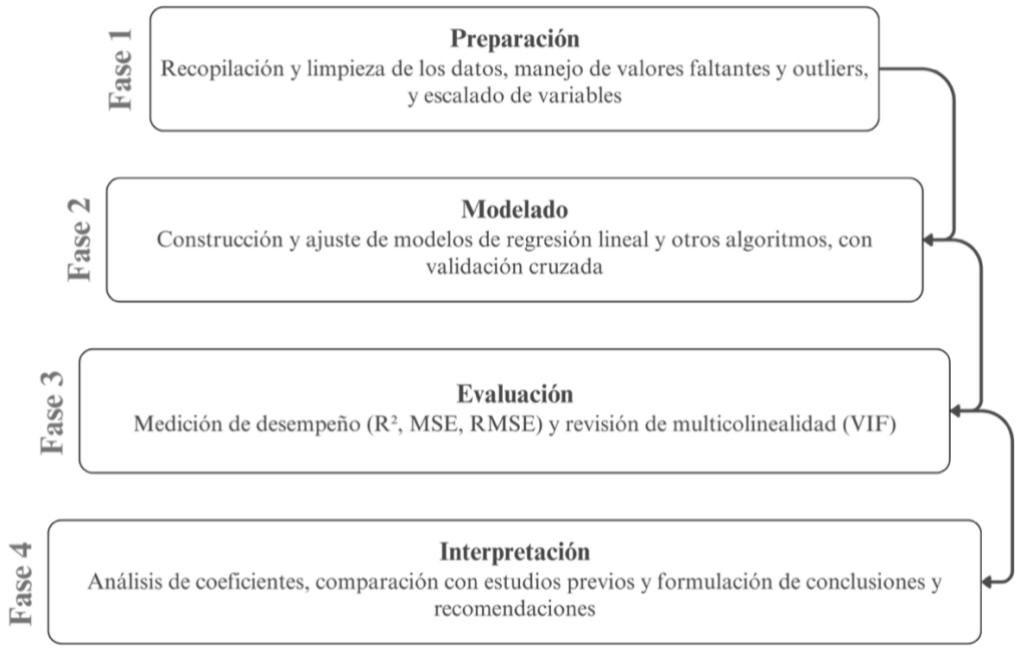
Variable	VIF	Interpretación
PuntajeL	1.357079	Baja colinealidad. La variable está bien condicionada para ser usada en el modelo.
PuntajeE	1.359918	Baja colinealidad. La variable está bien condicionada para ser usada en el modelo.
PuntajeM	1.246814	Baja colinealidad. La variable está bien condicionada para ser usada en el modelo.
PromedioBach_Sistemas	1.125217	Baja colinealidad. No se observa multicolinealidad significativa.

Nota. VIF (Variance Inflation Factor) indica el grado de colinealidad entre variables predictoras. Valores por debajo de 5 suelen interpretarse como evidencia de baja multicolinealidad. Elaboración propia.

El procedimiento siguió cuatro fases secuenciales (Figura 54): Preparación—limpieza, detección de outliers y normalización—; Modelado—ajuste de regresiones lineales simples y múltiples, además de Ridge, con validación cruzada 5-fold y partición 90 %/10 %—; Evaluación—cálculo de R^2, MSE, RMSE y comprobación de colinealidad (todos los VIF < 3)—; e Interpretación, basado en las recomendaciones de Gerón (2019) y Burkov (2019), contrastando resultados con la literatura (Fechter et al., 2021; Ferguson et al., 2020; Abu Dabrh et al., 2020; Tavares et al., 2017).

Figura 54

Procedimiento de análisis predictivo para la inferencia de extrapolación del ExIES



Nota. El proceso comprende la preparación, modelado, evaluación y la interpretación de resultados, siguiendo estándares internacionales de análisis educativo (Gerón, 2019; Burkov, 2019).

Técnicas de análisis

En cuanto a técnicas de análisis, emplearon: regresión lineal para estimar el efecto conjunto de los predictores; Ridge para atenuar posibles redundancias; y, con fines exploratorios, modelos de ensamble (Random Forest, Gradient Boosting, XGBoost) y un MLP. Todos se

ejecutaron en Python (pandas, numpy, scikit-learn). Las combinaciones exactas de predictores y las justificaciones detalladas de cada algoritmo aparecen en el Apéndice H. Los hiperparámetros explorados mediante grid-search y el flujo completo de preprocesamiento (Tablas E3-E4; Figuras E1-E2) se consignan igualmente en el Apéndice para quien desee replicar el estudio.

Modelado

La selección de los modelos (véase Tabla 96) Regresión Lineal, Ridge, Lasso, Random Forest, Gradient Boosting y XGBoost responde a la necesidad de balancear interpretabilidad, capacidad predictiva y adecuación a las características de los datos educativos analizados. Según las recomendaciones de Gerón (2019) y Burkov (2019), la Regresión Lineal fue incluida por su simplicidad y porque permite analizar relaciones lineales entre las variables predictoras y el rendimiento académico (Año1). Este modelo es particularmente útil en contextos educativos, donde la transparencia y la explicación de los efectos de cada predictor son esenciales. Ridge y Lasso, como variantes regularizadas de la regresión lineal, se eligieron para abordar posibles problemas de multicolinealidad entre las variables predictoras y realizar selección automática de características (Lasso), lo que puede ayudar a identificar las variables más relevantes para explicar el rendimiento académico.

Por otro lado, Random Forest y Gradient Boosting fueron seleccionados por su capacidad para modelar relaciones no lineales y capturar interacciones complejas entre las variables predictoras. En particular, Random Forest es robusto frente al ruido y menos propenso al sobreajuste, mientras que Gradient Boosting mejora iterativamente el rendimiento ajustándose a los errores residuales. Finalmente, XGBoost se incluye como una variante altamente eficiente de Gradient Boosting que combina un excelente desempeño predictivo con herramientas de regularización para prevenir el sobreajuste, haciéndolo ideal para el tamaño y la naturaleza de los

datos. Aunque opcional, el modelo de Red Neuronal (MLP) se considera para capturar relaciones altamente complejas entre las variables, aunque su naturaleza de caja negra y su alta demanda computacional limitan su aplicabilidad en contextos donde la interpretabilidad es prioritaria. En la Tabla 96 se describe el proceso y las evaluaciones realizadas.

Tabla 96

Descripción de los modelos predictivos utilizados y justificación metodológica

Aspecto	Descripción	Justificación
Modelo utilizado	Regresión Lineal, Ridge, Lasso, Random Forest, Gradient Boosting, Red Neuronal (MLP), XGBoost, todos en versión básica.	Facilita la comparación entre enfoques lineales, regularizados y de ensamble para identificar el modelo más adecuado.
Conjuntos de predictores	Varias combinaciones de variables predictoras (X_básico, X_completo, X_examen, etc.).	Permite evaluar el aporte de diferentes grupos de variables al rendimiento del modelo.
Escalado de variables	Normalización de todas las variables predictoras utilizando StandardScaler.	Evita problemas de escala y asegura la comparabilidad de los coeficientes del modelo.
Validación cruzada	Validación cruzada con 5 particiones.	Reduce el riesgo de sobreajuste y mejora la generalización del modelo.
División de datos	90% entrenamiento, 10% prueba.	Permite la evaluación objetiva del desempeño en datos no vistos.
Métricas de evaluación	Coeficiente de determinación (R^2), MSE y RMSE.	Permite cuantificar la capacidad explicativa y predictiva de los modelos probados.

Nota. Elaboración propia basada en datos obtenidos de la base de datos de promedios de bachillerato (EMS BC, 2024), base de datos del ExIES (Pedroza Zúñiga et al.,2024) y base de datos del promedio del primer año universitario (UABC, 2024). Todos los modelos fueron aplicados con sus hiperparámetros predeterminados; las variables predictoras fueron normalizadas mediante StandardScaler. Se empleó validación cruzada con 5 particiones y la división de datos fue de 90% para entrenamiento y 10% para prueba.

Cabe señalar que, en esta etapa del análisis, todos los modelos se emplearon con sus parámetros predeterminados o por defecto, es decir, sin ajuste específico de hiperparámetros. Esta decisión metodológica permite establecer una comparación inicial entre los modelos bajo condiciones estándar, priorizando la robustez y la reproducibilidad de los resultados. El ajuste fino de hiperparámetros (por ejemplo, búsqueda de la mejor combinación de alpha, número de

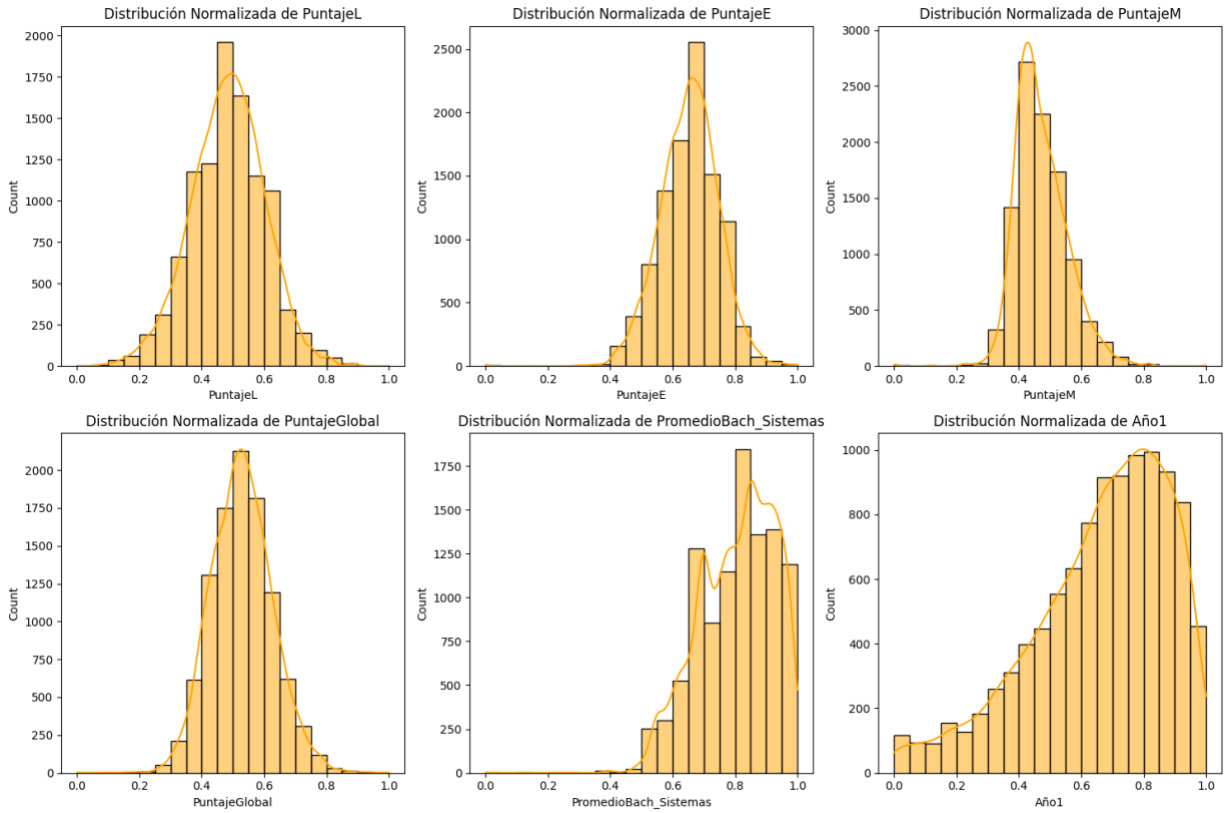
árboles o capas ocultas) se plantea como una posible línea futura de optimización, una vez identificados los modelos con mejor desempeño preliminar.

Ajuste de hiperparámetros

Se efectuó un ajuste sistemático de hiperparámetros mediante grid-search y validación cruzada para todos los modelos regularizados y de ensamble considerados (Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost y MLP). En términos generales, se exploraron valores graduales de penalización (*alpha*) en las regresiones Ridge y Lasso; diferentes profundidades, tamaños de bosque y criterios de división en Random Forest; combinaciones de número de iteraciones, tasa de aprendizaje y profundidad de árbol en los métodos boosting; y diversas arquitecturas y funciones de activación en la red neuronal. Tras contrastar cada configuración, se observó que los parámetros por defecto ofrecían el mejor compromiso entre simplicidad e incremento marginal de desempeño, de modo que los resultados reportados se basan en esas configuraciones base.

Figura 55

Histogramas de distribución de las variables predictoras tras normalización



Nota. Elaboración propia. Los histogramas muestran la distribución de variables predictoras después del proceso de normalización. La escala va de 0 a 1. La curva sobrepuesta representa una estimación de densidad. Las variables fueron normalizadas utilizando *StandardScaler*.

Hiperparámetros. Asimismo, se realizó un ajuste y comparación de los principales hiperparámetros en los modelos regularizados y de ensamble (Gerón, 2019; Burkov, 2019); para este propósito, se empleó una búsqueda en cuadrícula (grid search) combinada con validación cruzada. A continuación, se describen los hiperparámetros más relevantes evaluados para cada modelo:

***Ridge Regression*.** Los principales hiperparámetros ajustados fueron:

- *alpha*: Controla la intensidad de la regularización L2. Se evaluaron los valores 0.01, 0.1, 1.0, 10 y 100. Un valor bajo permite mayor ajuste al conjunto de entrenamiento, mientras que un valor alto penaliza la magnitud de los coeficientes para evitar el sobreajuste.

- *fit_intercept*: Determina si se incluye el término independiente en el modelo. Se evaluaron las opciones *True* y *False*.
- *solver*: Define el algoritmo de optimización. Se probaron los métodos *auto*, *svd* y *cholesky* para comparar su rendimiento computacional.

Lasso Regression. Los tres hiperparámetros principales fueron:

- *alpha*: Controla la regularización L1. Se evaluaron valores entre 0.001 y 10, lo que influye directamente en la selección automática de variables relevantes.
- *max_iter*: Número máximo de iteraciones del algoritmo. Se probaron valores de 1000 y 5000 para garantizar la convergencia en conjuntos de datos más grandes.
- *selection*: Estrategia de actualización de coeficientes. Se probaron los modos *cyclic* y *random*.

Random Forest. Los hiperparámetros esenciales evaluados fueron:

- *n_estimators*: Número de árboles en el bosque. Se evaluaron 100, 200 y 500. Más árboles suelen mejorar la precisión, pero aumentan el tiempo de cómputo.
- *max_depth*: Profundidad máxima de los árboles. Se probaron valores de 5, 10, 20, 50 y *None*, con el fin de evitar sobreajuste.
- *min_samples_split*: Mínimo de muestras necesarias para dividir un nodo. Se consideraron valores de 2, 5 y 10.

Gradient Boosting. Los tres hiperparámetros más ajustados fueron:

- *n_estimators*: Número total de iteraciones o árboles secuenciales. Se probaron valores entre 100 y 300.
- *learning_rate*: Tasa de aprendizaje del modelo. Se evaluaron 0.01, 0.05, 0.1 y 0.2. Un valor más bajo requiere más árboles, pero mejora la generalización.

- *max_depth*: Controla la complejidad de los árboles base. Se probaron profundidades entre 3 y 10.

XGBoost. Se ajustaron los siguientes hiperparámetros:

- *n_estimators*: Número total de árboles, evaluando 100, 200 y 300.
- *learning_rate*: Tasa de aprendizaje, con valores de 0.01 a 0.2.
- *max_depth*: Profundidad máxima de los árboles, para controlar el equilibrio entre precisión y sobreajuste.

Red Neuronal (MLPRegressor). Los hiperparámetros más relevantes fueron:

- *hidden_layer_sizes*: Configuración de capas ocultas. Se evaluaron estructuras como (50,), (100,), y (100, 50).
- *activation*: Función de activación en capas ocultas. Se probaron *relu*, *tanh* y *logistic*.
- *alpha*: Coeficiente de regularización L2. Se evaluaron valores como 0.0001, 0.001 y 0.01.

Tras la exploración de estos hiperparámetros, se concluyó que los valores predeterminados ofrecieron un equilibrio adecuado entre simplicidad del modelo y rendimiento predictivo. Por esta razón, los resultados presentados en este estudio corresponden principalmente a dichas configuraciones base.

Recursos de software y librerías

Retomando a Gerón (2019) y Burkov (2019), el análisis fue implementado en Python mediante el uso de bibliotecas especializadas para ciencia de datos y modelado estadístico. Las principales bibliotecas utilizadas fueron:

- pandas: Para la manipulación, transformación y limpieza de los conjuntos de datos, facilitando la gestión eficiente de estructuras tipo DataFrame.

- scikit-learn: Para la construcción de modelos de regresión lineal, aplicación de validación cruzada, cálculo de métricas de desempeño (R^2, *RMSE*, *MSE*) y escalado de variables mediante técnicas como StandardScaler.
- numpy: Para realizar operaciones Matemáticas avanzadas y manejar estructuras numéricas de alto rendimiento.

Apéndice H

Tabla Completa del Argumento de Validez del ExIES a partir del EBA: Recomendaciones y Reservas

A continuación, se presentan las recomendaciones y reservas por supuesto. Se colocaron las siglas C (claridad), Co (coherencia) y P (plausibilidad), para establecer qué ayuda a subsanar la recomendación.

Tabla 97

Tabla Completa del Argumento de Validez del ExIES a partir del EBA

Supuesto	Evidencia	Recomendaciones	Reservas
Inferencia de Definición de Dominio			
S1.1.1 Contenido definido desde competencias EMS	El ExIES mide Lectura, Lengua Escrita y Matemáticas con base en competencias oficiales de EMS y marcos internacionales (Acuerdos 442/444; DeSeCo/PISA; Manual y Guía) (SEP, 2008a, 2008b; Caso & Díaz, 2016; Caso et al., 2017).	Construir un mapa de trazabilidad (competencia→especificación→ítem) con criterios de inclusión/exclusión y nivel cognitivo homologado (Bloom; Anderson & Krathwohl). Sustentar con Estándares (AERA, 2014) y guías de validez de contenido (Sireci, 1998; Sireci & Faulkner-Bond, 2014; Lane et al., 2016). Foco: C y P.	Sin mapa ni justificación explícita, se debilita la representatividad del contenido (Kane, 2013; Chapelle, 2021).
S1.1.3 Revisión/actualización continua del dominio	El contenido se afinó entre 2022-2 y 2023-1 (más detalle en Matemáticas) (Pedroza Zúñiga et al., 2022; 2024a).	Llevar bitácora de cambios (qué cambió, por qué, quién) y documentar alineación con MCCEMS y Acuerdos 09/08/23 y 21/08/25 (NEM). (C, Co)	Posible desfase curricular y pérdida de validez de contenido (Kane, 2013).
S1.2.1 Ítems se adhieren a especificaciones y NDC	Los ítems siguen lo planeado (temas y niveles de demanda cognitiva; ver glosario) según manuales por área (Pedroza Zúñiga et al., 2023a–c; 2023n–p).	Comparar planificado vs. observado por forma/subversión y seguir revisando con rúbricas para revisar ítems (Haladyna & Rodríguez, 2013; AERA et al., 2014). (Co, P)	La adherencia puede degradarse con el tiempo.
S1.2.2 Estrategias de revisión (jueceo)	Hay jueceo con decisiones registradas (aceptar/modificar/descartar) (Pedroza Zúñiga et al., 2023d–f; 2023q).	Calcular CVR/CVI y dejar rastro de todas las decisiones (Lawshe, 1975; Lynn, 1986; AERA, 2014). (C, P)	Con criterios implícitos, la validez de contenido se cuestiona.
S1.2.3 Revisión continua y trazabilidad	Se monitorea con matrices NDC–subcontenido–ítem y con historial Rasch (Pedroza Zúñiga et al., 2023s; 2024c).	Reporte periódico de desviaciones comparándolo con la matriz inicial, reglas de retiro y edición con umbrales y responsables (AERA et al., 2014; Lane et al., 2016). (Co, P)	Pueden persistir ítems obsoletos, afectando la validez.

S1.3.2 Prevención de sesgos antes de publicar	Hay lineamientos anti-sesgo en el jueceo (Pedroza Zúñiga et al., 2023d–f; 2024a).	Manual breve de sesgos frecuentes con ejemplos del ExIES; fijar umbrales ETS para DIF (MH/log-OR) y rutas de acción (Holland & Thayer, 1988; Zieky, 1993; Magis et al., 2010; AERA et al., 2014). (C, Co, P)	Si el DIF solo "se ve" sin reglas, se afecta la imparcialidad.
S1.4.1 Autores/jueces calificados y capacitados	Selección con criterios, capacitación y métricas por autor (Pedroza Zúñiga et al., 2023g–h).	Formación continua con retroalimentación individual basada en métricas (tasa de aceptación, psicometría de sus ítems) (Haladyna & Rodríguez, 2013; Brookhart & Nitko, 2018). (Co)	Sin retroalimentación, baja la calidad de ítems.
Inferencia de Evaluación			
S2.1.1 Personal formado y estandarizado	Manuales/roles para aplicadores y supervisores (Pedroza Zúñiga et al., 2023i–k).	Evaluar la capacitación por sede y archivar micro-informes (AERA et al., 2014; Brennan, 2006). (P)	Crece la variabilidad entre sedes.
S2.1.2 Materiales para sustentantes	Guía con contenidos, ejemplos y manejo del tiempo (Pedroza Zúñiga et al., 2023l).	Medir cobertura y uso (descargas, distribución) y realizar simulacros estandarizados (AERA et al., 2014; Brookhart, 2013). (C, P)	Sin evidencia de alcance, no se prueba la equidad informativa.
S2.1.3 Seguridad y resguardo	Protocolos y registro de incidencias (Pedroza Zúñiga et al., 2023m; 2023r).	Sistema de incidencias (severidad, causas, acciones) y auditorías aleatorias. (Co)	Sin ciclo prevención-registro-retroalimentación, se riesga la aplicación.
S2.1.4 Reglas anti-deshonestidad	Conductas prohibidas y sanciones explícitas (Pedroza Zúñiga et al., 2023i, 2023j, 2023l).	Comunicación redundante (póster, lectura inicial, recordatorios) más un cuestionario breve de comprensión; guardar acuses. (C, Co)	Disuasión insuficiente si no se verifica la comprensión.
S2.1.5 Mismas condiciones de aplicación	Instrucciones para aulas, materiales y tiempos (Pedroza Zúñiga et al., 2023i–j).	Listas de cotejo por sede y auditorías; usar G-Theory cuando haya variaciones (Brennan, 2006; Shepard, 2006). (P)	Se compromete la estandarización.
S2.2.1 Rasch estima habilidad con calidad	El puntaje se calcula con Rasch, que transforma respuestas en una escala comparable y revisa ajustes de ítems/personas (Pedroza Zúñiga et al., 2024a; Bond & Fox, 2015).	Dibujar el flujo de análisis (depuración→escalamiento→equiparación→reporte) y justificar umbrales (Bond & Fox, 2015; Kolen & Brennan, 2014; AERA et al., 2014). (C)	Sin resumen unificado, decisiones dispersas.
S2.3.1 Ítems se pilotean antes de operar	Pilotaje 2022-2 (N≈2,210) y monitoreo semestral (Pedroza Zúñiga et al., 2022; 2024a).	Fijar criterios banco→operación; vigilar drift (cambios de parámetros) con anclas y equiparación, y reportarlo (Kolen & Brennan, 2014; AERA et al., 2014). (Co, P)	Sin monitoreo de drift, baja la comparabilidad entre semestres.
Inferencia de Generalización			
S3.1.1 Consistencia interna adecuada	Confiabilidad aceptable-alta: Lectura≈.73–.74; Lengua≈.77; Mate≈.82–.84; global≈.88–.89 (Pedroza Zúñiga et al., 2024a, 2024b).	Reporte semestral de α por área y, cuando haya varias fuentes de error, análisis con G-Theory (Nunnally & Bernstein, 1994; Brennan, 2006). (C, P)	Mirar solo α puede sobreestimar la estabilidad.

Inferencia de Explicación			
S4.1.1 Número de factores esperado	AFC sugiere tres factores; RMSEA aceptable; CFI/TLI < .90 (ajuste moderado) (Pedroza Zúñiga et al., 2024a).	Identificar ítems con cargas bajas/negativas, probar modelos jerárquicos o bifactor y justificar elección (Hu & Bentler, 1999; Zumbo & Chan, 2014; Lane et al., 2016). (P)	Si CFI/TLI siguen < .90, se cuestiona la estructura.
S4.1.2 Correlaciones positivas con global	Correlaciones altas entre verbales y global; con Matemáticas negativas en esta muestra (Pedroza Zúñiga et al., 2024a).	Revisar transversalidad y cargas cruzadas; explorar factor general o residuos correlacionados (Cronbach & Meehl, 1955; Kane, 2013; Zumbo & Chan, 2014). (C, P)	Sin explicación, interpretar con cautela la relación entre dominios.
S4.1.3 Cargas acordes al modelo	Hay cargas moderadas/débiles por forma/área (Pedroza Zúñiga et al., 2024a).	Lista de ítems débiles → editar–pilotear–reanalizar; si procede, estudios de procesos de respuesta y documentarlo (Zumbo & Chan, 2014; Kane & Mislevy, 2017). (Co, P)	Cargas débiles contaminan la interpretación de factores.
S4.2.1 Convergencia global (EXANI-II)	Correlación fuerte ExIES–EXANI II ($r \approx .76$; $p < .001$) (Pedroza Zúñiga et al., 2024a; Morales et al., 2015).	Documentar emparejamiento muestral, replicar en nuevas cohortes, reportar IC y restricción de rango (AERA et al., 2014; Sackett et al., 2008; Schober et al., 2018). (C, P)	Sin método claro, la convergencia se cuestiona.
S4.2.2 Convergencia por dominio	Correlaciones por dominio moderadas-fuertes (pilotaje 2022-2) (Pedroza Zúñiga & Gómez Monárrez, 2025a, 2025b).	Más estudios de correspondencia por dominio con mapeos documentados (AERA et al., 2014; Sireci & Faulkner-Bond, 2014). (Co, P)	Sin mapeo, la convergencia por dominio se debilita.
S4.3.1 Imparcialidad (DIF)	84 % de ítems sin DIF; 24 % con DIF moderado/severo por terciles; 6 % severo recurrente (2023-2) (Pedroza Zúñiga & Gómez Monárrez, 2025c).	Fijar umbrales ETS y rutas de acción (retirar C, revisar B, monitorear A) y añadir invariancia factorial (Holland & Thayer, 1988; Zieky, 1993; Magis et al., 2010; AERA et al., 2014). (Co, P)	No atender el 16 % con DIF compromete la equidad.
Inferencia de Utilización			
S6.1.1 La UABC usa la orientación del ExIES para prelación	La normativa permite usar exámenes y ordenar por prelación (UABC, 2010; 2019; 2021).	Protocolo UABC–ExIES para cortes y prelación; actas semestrales; guías de interpretación para aspirantes (AERA et al., 2014; Kane, 2013; Sackett et al., 2008). (C, Co)	Sin documentación, se debilita la justificación del uso.
S6.1.2 Efectividad del puntaje para seleccionar	Confiabilidad alta y convergencia externa apoyan el uso del puntaje (Pedroza Zúñiga et al., 2024a, 2024b).	Revisiones admitidos vs. rechazados, seguimiento del desempeño por cortes y subgrupos; reportes de equidad (AERA et al., 2014; Kolen & Brennan, 2014). (P)	Si se depende solo del puntaje, puede ser inequitativo/ineficiente.
Inferencia de Implicación de Consecuencias			
S7.1.1 Responsabilidad conjunta y prelación justa	Estándares y normativas respaldan el uso; hace falta formalizar una mesa conjunta UABC–ExIES (AERA, 2014; UABC, 2010, 2019, 2021; Pedroza Zúñiga et al., 2024a).	Establecer un Comité UABC–ExIES que cada semestre publique indicadores clave (admisiones por subgrupo, retención, GPA) y actas con decisiones para ajustar cortes y políticas según la evidencia. (Messick, 1989; Kane, 2015). (C, Co, P)	Sin comité ni métricas, se debilita la responsabilidad y la legitimidad.

S7.1.2 Clasificación imparcial y útil	DIF global leve (84 % sin DIF), validez predictiva moderada ($R^2 \approx .22$) y confiabilidad alta (Pedroza Zúñiga & Gómez Monárrez, 2025c; Pedroza Zúñiga et al., 2024a).	Monitorear y corregir continuamente ítems y cortes: usar umbrales ETS para DIF, reportar SEM e impacto por subgrupo, y ajustar o retirar ítems/cortes que generen desigualdad (AERA et al., 2014). (Co, P)	Sin correcciones y monitoreo por subgrupos, la clasificación pierde imparcialidad y utilidad.

Nota. La tabla resume el Argumento de Validez del Examen de Ingreso a la Educación Superior (ExIES) elaborado a partir del Enfoque Basado en Argumentos (EBA). Las columnas detallan, respectivamente, la Inferencia (G = garantía), las Suposiciones (S) que la sustentan, la Evidencia empírica o documental que las respalda, las Recomendaciones para fortalecer la validez y las Reservas que podrían debilitarla. Abreviaturas: EMS = Educación Media Superior; SEP = Secretaría de Educación Pública; OCDE = Organización para la Cooperación y el Desarrollo Económicos; DIF = Differential Item Functioning (funcionamiento diferencial del ítem); TRI = Teoría de Respuesta al Ítem; AFC = Análisis Factorial Confirmatorio. Las referencias citadas (p. ej., Pedroza Zúñiga et al., 2024a) corresponden a manuales, reportes técnicos o literatura especializada disponible en la bibliografía principal del estudio. Asimismo, para interpretar mejor la tabla se agrega un glosario:

1. Niveles de demanda cognitiva (NDC): grado de esfuerzo mental que exige un ítem (recordar, comprender, aplicar, analizar, evaluar, crear); ayuda a evitar que la prueba mida solo memoria (Bloom; Anderson & Krathwohl).
2. Rúbricas para revisar ítems (Haladyna & Rodríguez, 2013; AERA, 2014):
    1. una sola idea por ítem; 2) lenguaje claro y sin ambigüedades; 3) clave única e indiscutible; 4) distractores plausibles y homogéneos; 5) sin pistas gramaticales ni de longitud; 6) evitar negaciones y absolutos (“siempre”, “nunca”); 7) la dificultad debe provenir del contenido, no de la redacción; 8) alineación con la competencia y el NDC; 9) sensibilidad cultural/lingüística (*fairness*); 10) estímulos necesarios y bien editados; 11) no usar “todas/ninguna de las anteriores”; 12) revisar sesgo potencial (género, región, tecnicismos innecesarios).
3. CVR (*Content Validity Ratio*; Lawshe, 1975): razón que indica cuántos jueces consideran un ítem esencial; se compara con valores críticos según el número de jueces para decidir conservar, corregir o descartar.
4. CVI (*Content Validity Index*; Lynn, 1986): proporción de jueces que califican el ítem como relevante y claro (I-CVI por ítem; S-CVI por escala) para depurar redacción y pertinencia.
5. Rasch (Bond & Fox, 2015): modelo que convierte respuestas en una escala continua al comparar habilidad de la persona con dificultad del ítem; revisa ajuste (infit/outfit) para asegurar medición estable y comparable.
6. Equiparación (Kolen & Brennan, 2014): hace que diferentes formas del examen tengan igual significado de puntaje mediante ítems ancla y verificación de *drift* (cambios de parámetros).
7. Confiabilidad α: indica consistencia interna por área y global; puede complementarse con Teoría de la Generalizabilidad para separar fuentes de error (sede, forma, día).
8. AFC, RMSEA, CFI y TLI (Hu & Bentler, 1999; Zumbo & Chan, 2014): indicadores para comprobar si los datos siguen el modelo (p. ej., tres factores); RMSEA cercano a 0 y CFI/TLI cercanos a 1 sugieren mejor ajuste y respaldan la validez de constructo.
9. DIF (Holland & Thayer, 1988; Zieky, 1993; Magis et al., 2010): verifica si un ítem favorece a un grupo a igual habilidad; la clasificación ETS (A/B/C) guía decisiones (A = ok; B = revisar/pilotear; C = retirar) para garantizar imparcialidad.
10. Invariancia factorial: confirma que la estructura del examen es equivalente entre grupos y habilita comparaciones justas.
11. Error estándar de medida (SEM) e intervalos de clasificación: comunican la incertidumbre del puntaje y apoyan decisiones prudentes sobre cortes (AERA, 2014).

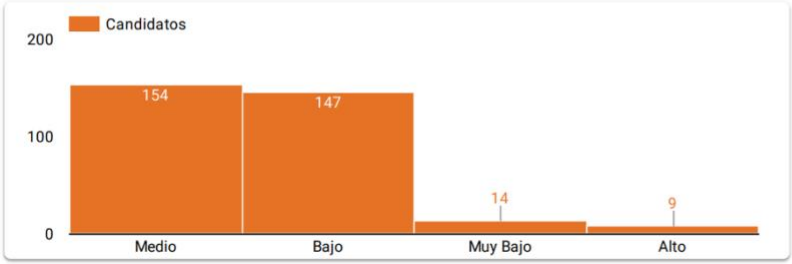
Apéndice I

Ejemplo de tablero para la divulgación de resultados del ExIES hacia una administración más automatizada

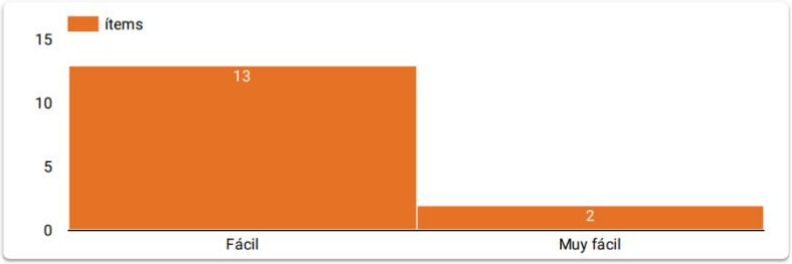
Selecciona un periodo

Resultados del Modelo Rasch

Distribución de la Habilidad de los Candidatos



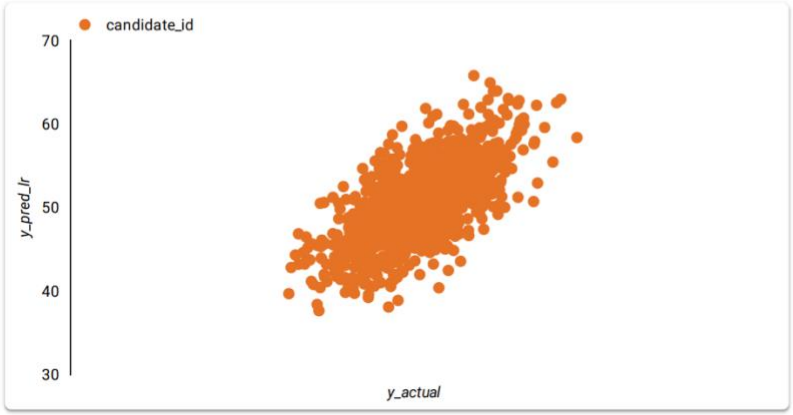
Distribución de la Dificultad de ítems

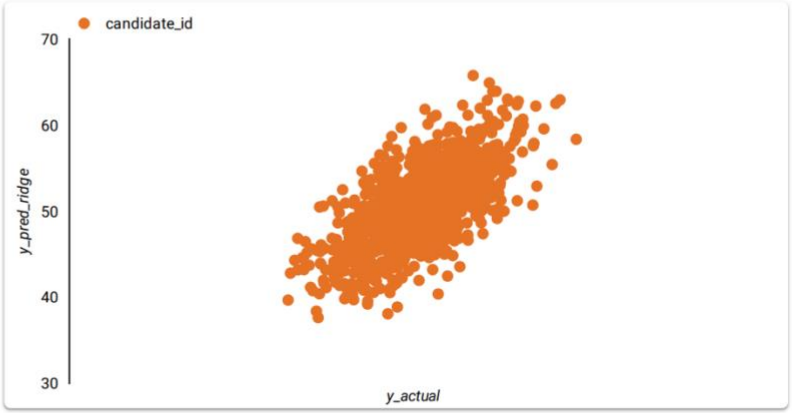


	Dificultad de ítems (Betha)	ítem	Cantidad
1.	Muy fácil	LEN_1	1
2.	Muy fácil	COM_2	1
		Total	**2**

Selecciona un periodo

Resultados de Regresión Lineal y Ridge





Variable	Coef_Ridge	Coef_Lineal
sexo_Masculino	-4,89	-4,89
nivel_socioeconomico_Medio	-3,84	-3,85
nivel_socioeconomico_Bajo	-6,99	-7
promedio_prueba	0,29	0,29
promedio_bachillerato	1,96	1,96

Modelo	R2
Ridge	0,49
Lineal	0,49