

Concepto Técnico

Propiedades psicométricas de instrumentos de medición psicológica para uso por parte de los profesionales adscritos al Consultorio Psicoforenses de la Universidad Santo Tomás: Una revisión sistemática rápida

Lauren Daniela Cuellar Valderrama

Sofía Luna Amado

Jhoana Alejandra Villalba Sayo

Consultorio Psicoforenses

Facultad de Psicología

Universidad Santo Tomás

2025

Esta investigación se realizó como requisito para optar al título de Psicología en la Universidad Santo Tomás, bajo la dirección del docente Luis Alberto Quiroga Baquero. Las autoras declaran no tener ningún conflicto de intereses.



UNIVERSIDAD
SANTO TOMÁS

Resumen

Objetivo: Analizar la evidencia empírica sobre las propiedades psicométricas de 13 instrumentos de evaluación psicológica utilizados en el Consultorio Psicoforense de la Universidad Santo Tomás, considerando su aplicabilidad en contextos forenses iberoamericanos. **Método:** Se realizó una revisión sistemática rápida siguiendo los estándares PRISMA. Se consultaron doce bases de datos (EBSCO, Emerald, Oxford Academic, PubMed, Redalyc, Sage, ScienceDirect, SciELO, Scopus, SpringerLink, Taylor & Francis, Web of Science) y búsqueda manual en referencias, seleccionando artículos publicados entre 2000 y 2024 en español e inglés. Tres revisoras independientes evaluaron títulos, resúmenes y textos completos. Se extrajeron datos sobre validez, confiabilidad y calidad metodológica según criterios COSMIN. Los instrumentos analizados fueron: MMPI-2, MMPI-2-RF, PAI, SIMS, NEUROPSI, BANFE, TECA, SCL-90-R, SENA, CUIDA, TAMAI y LSB-50. Del instrumento MCMI-IV no se encontraron registros. **Resultados:** Se incluyeron 44 estudios. Los instrumentos con mayor consistencia interna fueron NEUROPSI ($\alpha \approx .91$), SIMS ($\alpha \approx .87$) y TAMAI ($\alpha \approx .86$). El TECA, LSB-50, MMPI-2-RF, SCL-90 y PAI mostraron evidencia robusta de validez estructural mediante análisis factorial confirmatorio. La mayoría de los estudios presentaron calidad metodológica moderada según COSMIN. Se identificó ausencia crítica de: (a) estudios de invarianza factorial en siete instrumentos, (b) validaciones en población forense específica, y (c) baremos normativos colombianos actualizados. **Conclusiones:** Los instrumentos analizados poseen propiedades psicométricas adecuadas en contextos iberoamericanos generales, pero requieren validaciones específicas en población colombiana, estudios de invarianza transcultural y desarrollo de normas locales para garantizar interpretaciones válidas en peritajes psicológicos.

Palabras clave: evaluación psicológica forense, propiedades psicométricas, validación de instrumentos, COSMIN, contexto iberoamericano

Propiedades psicométricas de instrumentos de medición psicológica para uso por parte de los profesionales adscritos al Consultorio Psicoforense de la Universidad Santo Tomás: Una revisión sistemática rápida

En el ámbito jurídico-forense, la evaluación psicológica constituye un elemento fundamental para la toma de decisiones legales y la protección de los derechos de quienes están involucrados en procesos judiciales. A pesar de la importancia que reviste la psicometría en este contexto, existe una notable contradicción: mientras el sistema legal exige un alto nivel de rigor técnico y científico en la valoración de los casos y los medios probatorios, la aplicación de pruebas psicométricas por parte de los psicólogos jurídicos es escasa, especialmente en Latinoamérica. Esta situación evidencia una brecha significativa entre los avances científicos en evaluación psicológica y su implementación efectiva en la práctica forense (Espinosa-Becerra & Quiroga-Baquero, 2022; Nijdam-Jones et al., 2020).

El objetivo del presente estudio es analizar la evidencia empírica acerca de las propiedades psicométricas de 13 instrumentos de medición psicológica de uso por parte de los profesionales adscritos al Consultorio Psicoforense de la Universidad Santo Tomás, con el fin de justificar la necesidad de contar con herramientas de alta calidad técnica que respondan a las exigencias del contexto jurídico. Propuesto así, se busca identificar, sistematizar y evaluar críticamente la evidencia disponible sobre estos instrumentos, prestando especial atención a su validez, fiabilidad y aplicabilidad en distintos contextos culturales, particularmente en Iberoamérica.

El uso de pruebas psicológicas en contextos legales resulta clave para la toma de decisiones en diversas áreas como la evaluación de responsabilidad penal, la competencia procesal, la imputabilidad, la predicción del riesgo de reincidencia, la custodia infantil, entre otras (Neal et al., 2020). En estos escenarios, el margen de error en una evaluación psicológica puede tener repercusiones graves, afectando el derecho a un juicio justo o incluso determinando la sentencia de una persona.

La implementación de instrumentos de medición adecuados resulta crucial en la toma de decisiones jurídicas. Sin estas herramientas, las decisiones se basarían únicamente en la interpretación subjetiva del evaluador y del juzgador, lo que podría conducir a errores graves en

la administración de justicia. El uso de pruebas con un alto estándar técnico permite minimizar los sesgos subjetivos del evaluador y aporta objetividad a los informes periciales. En este sentido, las metodologías de evaluación psicológica forense multimétodo y multifuente basadas en la aplicación de entrevistas estructuradas, escalas de medición y pruebas estandarizadas aumenta la credibilidad de los dictámenes periciales y fortalece su impacto en el ámbito legal (Nijdam-Jones et al., 2020).

Al respecto, Neal et al. (2020) advierten que el uso de pruebas psicológicas es esencial para evitar la introducción de lo que denominan "*ciencia basura*" en los tribunales. Este término se refiere a la aceptación de métodos sin base empírica sólida que pueden distorsionar significativamente las decisiones legales. Por esta razón, la psicología forense debe apoyarse en instrumentos adaptados, validados y estandarizados que permitan diferenciar con precisión entre individuos que presentan características psicológicas genuinas y aquellos que pueden estar simulando o exagerando síntomas para obtener beneficios en el contexto legal (Espinosa-Becerra et al., 2020).

Frente a la necesidad de rigor metodológico, la psicología forense establece diversas directrices que asignan responsabilidades y competencias específicas a los profesionales del área. De acuerdo con la American Psychological Association (APA, 2013), los especialistas en psicología forense se comprometen a mantener la precisión, honestidad y veracidad en la investigación, la enseñanza y la práctica profesional. Asimismo, se esfuerzan por resistir a presiones externas y conflictos de intereses que puedan influir en la prestación de servicios a la administración de justicia de manera engañosa o inexacta.

La APA (2013) señala que, al ofrecer una opinión experta en la que pueda confiar quien toma las decisiones, los profesionales forenses deben procurar precisión, imparcialidad, equidad e independencia. Esto implica utilizar diseños de investigación y métodos científicos adecuados para evaluar de forma justa las cuestiones en curso y resistir presiones partidistas que puedan inducir a reportar resultados de manera sesgada o engañosa.

Además, se alienta a los profesionales forenses a identificar, dar a conocer y abordar los conflictos de intereses, reales o aparentes, con el fin de mantener la confianza pública, cumplir con sus obligaciones profesionales y preservar la responsabilidad, imparcialidad y rendición de cuentas (APA, 2010; Norma 3.06 del *Ethical Principles of Psychologists and Code of Conduct* [EPPCC]). También se reconoce la importancia de conocer las normas, leyes, reglas y

precedentes legales y profesionales que regulan su participación en procedimientos legales y del impacto de sus servicios en los destinatarios (APA, 2010; Norma 2.01 EPPCC).

La Guía 10.02 de las *Specialty Guidelines for Forensic Psychology* (APA, 2013) advierte específicamente que los profesionales deben emplear instrumentos validados para la población evaluada y señalar las limitaciones cuando esta validez no esté confirmada, algo que frecuentemente no ocurre de forma sistemática en la práctica. Asimismo, la APA (2006) enfatiza la importancia de la práctica basada en la evidencia (*Evidence-Based Practice in Psychology* [EBPP]) como enfoque fundamental para garantizar la efectividad de las intervenciones psicológicas.

Es importante destacar que la evaluación en contextos forenses difiere significativamente de la evaluación en contextos clínicos y terapéuticos. Los profesionales forenses deben tener en cuenta estas diferencias al realizar exámenes, considerando cuidadosamente las ventajas y desventajas de emplear procedimientos de evaluación establecidos (AERA et al., 2014). Además, dadas las implicaciones legales, se esfuerzan por garantizar la integridad y seguridad de los materiales y resultados de las pruebas (APA, 2013). En este sentido, en el ámbito jurídico donde las decisiones pueden tener implicaciones trascendentales para la vida de las personas evaluadas, es imperativo que los instrumentos utilizados posean un elevado rigor técnico (Jiménez-Ardila et al., 2021).

A continuación, se describen las propiedades psicométricas fundamentales que deben cumplir los instrumentos de evaluación en contextos forenses. Dos conceptos esenciales en el campo de la psicometría forense son la validez y la confiabilidad (fiabilidad). La validez, se refiere al grado en que un instrumento mide realmente el constructo que pretende evaluar, mientras que la confiabilidad hace referencia a la consistencia o estabilidad de las mediciones obtenidas con el instrumento (Lövgren et al., 2022).

La APA (2013) menciona que, en contextos forenses, esto resulta crucial para garantizar que las inferencias realizadas a partir de las puntuaciones sean precisas y relevantes para las decisiones judiciales. La validez abarca diferentes aspectos como: (a) la evidencia de validez basada en el contenido (representatividad de los ítems); (b) la evidencia de validez basada en la estructura interna y en el constructo (relación con el marco teórico); (c) evidencia de validez basada en su relación con otras variables (capacidad predictiva, convergente, divergente o discriminante).

Por su parte, Neal et al. (2020) señalan que el estado actual de la evaluación psicológica forense ha mejorado considerablemente respecto al pasado, gracias a la evolución del campo y sus fundamentos científicos. No obstante, advierten que aún existe una variación significativa en la calidad de las evaluaciones, la cual podría optimizarse mediante acciones coordinadas entre los psicólogos y los sistemas judiciales.

En este sentido, según Guarnera et al. (2017) resulta indispensable garantizar altos niveles de confiabilidad en los instrumentos utilizados, dado que esta propiedad refleja la minimización del error de medición, un aspecto crucial en evaluaciones que pueden tener consecuencias legales de gran impacto. La confiabilidad puede evaluarse a través de diferentes procedimientos, como: (a) la consistencia temporal (test-retest), (b) la consistencia interna (coherencia entre ítems), y (c) el acuerdo entre evaluadores, entre otras técnicas. Así mismo, comprender estos principios no solo permite analizar críticamente la calidad de los instrumentos psicométricos, sino también avanzar hacia el diseño y la aplicación de herramientas cada vez más precisas, válidas y éticamente responsables en el ámbito forense.

En concordancia con lo anterior, Archer et al. (2016) establecen criterios específicos que definen los instrumentos forenses con adecuado respaldo empírico, destacando dos características fundamentales. La primera es contar con una adecuada estandarización, lo que implica definir claramente las condiciones para su aplicación, puntuación e interpretación, usualmente detalladas en un manual. Este documento también debe especificar las competencias necesarias del evaluador y los métodos de calificación disponibles. La finalidad de la estandarización es minimizar los errores y asegurar que los resultados reflejen con precisión las características evaluadas, evitando influencias externas como el examinador o el entorno de aplicación. La segunda característica es la evidencia de fiabilidad y validez, avalada por publicaciones científicas revisadas por pares. Esto incluye información sobre la tasa de error del instrumento, expresada mediante el error estándar de medición, sensibilidad o especificidad en contextos forenses. Asimismo, se consideran aspectos como la consistencia en el tiempo (fiabilidad test-retest) y la coherencia interna (coeficiente alfa) para pruebas que evalúan características homogéneas.

Uno de los principales desafíos de la psicometría en contextos forenses es la presencia de sesgos de respuesta que pueden distorsionar significativamente los resultados de las evaluaciones. En estos escenarios, las personas evaluadas frecuentemente tienen incentivos para

manipular su presentación personal de dos formas principales. La simulación entendida como la exageración deliberada de síntomas con el fin de obtener beneficios legales (como declaración de inimputabilidad o atenuantes), y la disimulación, referida a la minimización de síntomas o problemas para evitar sanciones o consecuencias negativas. Asimismo, la deseabilidad social, entendida como la tendencia a responder de forma socialmente aceptable, se convierte en un obstáculo particularmente relevante cuando se utilizan instrumentos de autoinforme. Tal como señalan Sanz et al. (2018), este sesgo "implica la tendencia a presentarse de una manera demasiado favorable, tanto negando la existencia de características negativas como atribuyéndose cualidades positivas", y está relacionado con otros constructos como el autoengaño y el fingimiento. Por ello, es esencial que los instrumentos utilizados en contextos forenses incluyan mecanismos específicos para detectar y controlar estos estilos de respuesta, garantizando así la validez de las inferencias diagnósticas y periciales.

Un factor que ha obstaculizado la implementación efectiva de pruebas psicométricas en el contexto forense es la tendencia generalizada a emplear instrumentos diseñados originalmente para la psicología clínica, sin las necesarias adaptaciones para la evaluación de riesgo y victimología en entornos forenses. Estudios como el de Nijdam-Jones et al. (2020) han identificado barreras específicas en la implementación de estas herramientas en Latinoamérica, entre las que destacan la falta de capacitación especializada en psicometría forense, el acceso limitado a pruebas adaptadas al contexto sociocultural latinoamericano, la escasez de recursos para la adquisición de instrumentos estandarizados, la resistencia al cambio en las prácticas evaluativas tradicionales, entre otras.

De esta manera, la aplicación de herramientas no validadas para la población específica puede llevar a errores graves en la interpretación de los resultados y afectar significativamente la credibilidad del informe pericial ante las instancias judiciales. La investigación científica ofrece un respaldo sólido a la importancia de utilizar instrumentos psicométricos estandarizados en contextos forenses. A continuación, se presenta una síntesis de los estudios más relevantes que demuestran la efectividad de estas herramientas.

Lövgren et al. (2022) realizaron un estudio exhaustivo en Noruega donde analizaron 500 informes de evaluaciones forenses, encontrando una correlación significativa entre el uso de instrumentos psicométricos validados y la precisión de las conclusiones diagnósticas y forenses. Este hallazgo demuestra que la aplicación sistemática de pruebas científicas en la evaluación de

criminalidad no solo mejora la calidad del diagnóstico, sino que también fortalece la objetividad del informe forense presentado ante los tribunales.

Por su parte, Lally (2003) investigó la evidencia empírica disponible y realizó una encuesta sistemática a 64 expertos en psicología forense sobre la frecuencia y aceptabilidad de diversas pruebas psicológicas utilizadas en evaluaciones forenses. Los resultados identificaron instrumentos con amplio respaldo para diferentes propósitos evaluativos. Por ejemplo, para evaluar el estado mental en el momento del delito, identificaron el MMPI-2 (Inventario Multifásico de Personalidad de Minnesota-2), el PAI (Inventario de Evaluación de la Personalidad) y el R-CRAS (Escala de Evaluación de Responsabilidad Criminal de Rogers). Para evaluar el riesgo de violencia, resaltaron el PCL-R (Lista de Verificación de Psicopatía de Hare) y el HCR-20 (Evaluación de Riesgo de Violencia).

Asimismo, la literatura especializada ha identificado varios instrumentos diseñados específicamente para responder a las necesidades del ámbito jurídico-forense. El MacCAT-CA (Herramienta de Evaluación de Competencia de MacArthur para el Juicio Criminal) es un instrumento diseñado para evaluar la competencia de los individuos dentro del sistema de justicia penal. Este instrumento consta de tres escalas: Comprensión, Razonamiento y Apresiasi, y utiliza un caso hipotético sobre dos individuos involucrados en un delito para evaluar la capacidad del examinado para procesar información legalmente relevante (Poythress et al., 1999). El R-CRAS (Escala de Evaluación de Responsabilidad Criminal de Rogers) es un modelo estructurado de decisión diseñado para cuantificar variables psicológicas pertinentes a la “evaluación retrospectiva penal” en el momento del delito. Este instrumento fue validado específicamente para evaluar el estándar de locura del Instituto Americano de Derecho (ALI), que exige la evaluación del deterioro cognitivo y volitivo del acusado (Rogers, 1984).

Por otra parte, en torno a los avances metodológicos en este campo, los *métodos actuariales* basados en análisis estadísticos de datos empíricos han demostrado ser consistentemente superiores a los métodos clínicos tradicionales (basados únicamente en el juicio profesional) en diversos contextos forenses. Esta superioridad se debe a que permiten una evaluación más objetiva, sistemática y fundamentada en evidencia científica (Córdoba-Viveros et al., 2024).

A su vez, el *método de juicio profesional estructurado* (JPE) constituye un enfoque de evaluación de riesgo que integra la rigurosidad científica con la experiencia clínica, superando

las limitaciones tanto del juicio clínico no estructurado como de los métodos puramente actuariales. Este modelo se fundamenta en la valoración sistemática de factores de riesgo empíricamente validados, organizados en instrumentos estandarizados que guían al evaluador a través de dominios específicos (históricos, clínicos y de gestión del riesgo), sin reducir la complejidad del caso a una puntuación numérica. A diferencia de los instrumentos actuariales, el JPE permite al profesional integrar factores idiográficos, considerar la presencia, relevancia e implicaciones de cada variable en el contexto particular del evaluado, y formular escenarios de riesgo prospectivos que informan estrategias de gestión e intervención personalizadas. La decisión final sobre el nivel de riesgo (bajo, moderado o alto) emerge de un análisis comprehensivo que combina la evidencia empírica con el razonamiento clínico fundamentado, resultando en evaluaciones más precisas, transparentes y útiles para la toma de decisiones en contextos forenses, penitenciarios y de salud mental (Bartol & Bartol, 2015).

La comunidad de psicólogos forenses ha desarrollado un creciente consenso sobre la necesidad de incorporar herramientas psicométricas estandarizadas en sus evaluaciones, reconociendo su contribución en la detección eficaz de la simulación y otros sesgos de respuesta, el aumento de la fiabilidad interevaluadores, la mejora en la precisión de los diagnósticos en entornos legales y la mayor transparencia y objetividad de los informes periciales. Sin embargo, a pesar de este reconocimiento, la literatura actual evidencia vacíos importantes, especialmente en el contexto latinoamericano. Diversos estudios, como el de Nijdam-Jones et al. (2020), han señalado que, aunque estas herramientas fortalecen la calidad y fiabilidad de los informes periciales, su aplicación en la región continúa siendo limitada debido a factores como la escasa capacitación especializada y la falta de pruebas adaptadas culturalmente.

Las principales lagunas identificadas en la literatura actual incluyen:

- a. Carencia de investigaciones contextualizadas: Existe una escasez de estudios empíricos que analicen en profundidad cómo las barreras identificadas afectan específicamente la práctica evaluativa en diferentes países latinoamericanos.
- b. Falta de sistematización: Hay poca información organizada sobre la eficacia real de los instrumentos utilizados en contextos forenses latinoamericanos y su impacto concreto en las decisiones judiciales.

- c. Limitada adaptación cultural: La mayoría de los instrumentos disponibles han sido desarrollados en contextos anglosajones, con escasas adaptaciones que consideren las particularidades culturales, lingüísticas y normativas de Latinoamérica.
- d. Ausencia de consenso profesional: No existe un acuerdo generalizado entre los profesionales de la psicología forense latinoamericana sobre cuáles son los instrumentos más adecuados para diferentes tipos de evaluaciones.

El presente estudio se plantea como una respuesta necesaria ante estos vacíos, buscando recopilar, analizar críticamente y articular el conocimiento disponible para ofrecer una visión más completa y contextualizada de los desafíos actuales en la psicología forense de nuestra región. Propuesto así, esta investigación contribuirá a la psicología jurídica y forense en varios niveles: (a) A nivel teórico-conceptual, proporcionará una síntesis actualizada del estado del conocimiento sobre propiedades psicométricas de instrumentos forenses, clarificando conceptos fundamentales y estableciendo un marco de referencia unificado. (b) A nivel metodológico buscará identificar las mejores prácticas en la selección, aplicación e interpretación de instrumentos psicométricos en contextos forenses, ofreciendo lineamientos claros para los profesionales. (c) A nivel práctico-profesional, contribuirá a cerrar la brecha entre los avances científicos en evaluación psicológica y su implementación en la práctica forense cotidiana, facilitando la toma de decisiones informadas por parte de los psicólogos. (d) A nivel académico-formativo, servirá como referencia para la actualización de programas de formación en psicología jurídica y forense, enfatizando la importancia de la capacitación en psicometría.

Para superar los desafíos identificados, esta investigación promoverá acciones concretas como la capacitación continua en psicometría forense, la actualización de protocolos de evaluación para incluir pruebas estructuradas y el fortalecimiento de la colaboración entre psicólogos y el sistema legal. Solo mediante estos esfuerzos coordinados se podrá garantizar que las evaluaciones forenses sean más objetivas, científicamente fundamentadas y útiles para la toma de decisiones en los tribunales (Nijdam-Jones et al., 2020).

La presente revisión sistemática se justifica como un aporte fundamental para consolidar el conocimiento existente y superar las limitaciones previamente identificadas por autores como Archer et al. (2010), Guarnera et al. (2017), Lally (2003), Lövgren et al. (2022) y Neal et al. (2020), quienes han destacado la importancia crítica de emplear herramientas psicométricas

validadas para evitar errores graves en la administración de justicia, campo donde el margen de error puede tener consecuencias trascendentales para la vida de las personas evaluadas.

Método

Diseño

La presente investigación utilizó un diseño de síntesis de evidencia a través de una revisión sistemática rápida basada en los estándares establecidos en la Guía The Campbell Collaboration (2019) y en la Declaración PRISMA (Page et al., 2021; Tricco et al., 2018). Este tipo de revisión es una versión abreviada de las revisiones sistemáticas tradicionales y permitió identificar, recopilar, analizar y resumir evidencia científica de forma rigurosa, pero en un tiempo más corto sin perder calidad en el proceso (Page et al., 2021).

Se recurre a este tipo de revisión cuando se necesita contar con información actualizada y confiable para apoyar decisiones en contextos profesionales específicos. En este caso, el objetivo es ofrecer una síntesis técnica sobre las propiedades psicométricas como la validez, la confiabilidad y la estandarización de 13 instrumentos comúnmente utilizados en evaluaciones psicológicas forenses, a saber:

- Minnesota Multiphasic Personality Inventory-2 (MMPI-2)
- Minnesota Multiphasic Personality Inventory-2, Restructured Form (MMPI-2-RF)
- Personality Assessment Inventory (PAI)
- Millon Clinical Multiaxial Inventory-IV (MCMI-IV)
- Symptom Checklist-90-Revised (SCL-90-R)
- Listado de Síntomas Breve (LSB-50)
- Structured Inventory of Malingered Symptomatology (SIMS)
- NEUROPSI: Atención y memoria
- Test de Empatía Cognitiva y Afectiva (TECA)
- Sistema de Evaluación de Niños y Adolescentes (SENA)
- Cuestionario para la Evaluación Integral de la Idoneidad para la Adopción (CUIDA)
- Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI)
- Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales (BANFE)

La elección de este enfoque respondió a la necesidad de contar con evidencia clara, organizada y comprobable sobre la calidad técnica de algunas de las pruebas empleadas en el Consultorio Psicoforense de la Universidad Santo Tomás, haciendo un uso eficiente del tiempo y los recursos, pero sin sacrificar la rigurosidad del análisis. El motivo principal para optar por una revisión rápida fue la urgencia de fortalecer prácticas fundamentadas en evidencia dentro del ámbito forense, dado que los resultados de estas evaluaciones influyen directamente en decisiones jurídicas como la asignación de custodia, la determinación de imputabilidad, la evaluación del daño psicológico o los casos de violencia intrafamiliar (Arce et al., 2009; Lobo et al. 2016). Así, esta metodología permite abordar de manera técnica y ordenada un conjunto de instrumentos de medición clave, combinando evidencia científica con criterios de relevancia, validez y adecuación al contexto de aplicación.

Criterios de elegibilidad

Para asegurar la calidad de la información recopilada, la Tabla 1 muestra los criterios de elegibilidad implementados:

Tabla 1

Criterios de inclusión y exclusión

Categoría	Criterios de inclusión	Criterios de exclusión
Periodo de publicación	2000 – 2024	Estudios anteriores al año 2000 o posteriores a diciembre de 2024
Idioma	Español e inglés	Otros idiomas
Cobertura geográfica	Preferencia por estudios realizados en Iberoamérica	Ninguno explícito, pero se excluye si no aportan al contexto iberoamericano
Tipo de publicación	Artículos científicos revisados por pares, informes técnicos de validación, publicaciones institucionales	Documentos sin respaldo metodológico o con carácter exclusivamente teórico
Población objetivo	Adultos (sin distinción de sexo, identidad de género u orientación sexual)	Estudios en poblaciones exclusivamente infantiles, adolescentes o sin relación con contextos forenses
Tipo de estudio	Investigaciones instrumentales, estudios cuantitativos sobre propiedades psicométricas	Estudios sin resultados psicométricos
Instrumentos evaluados	MMPI-2, MMPI-2-RF, PAI, MCMI-IV, SIMS, NEUROPSI, BANFE, TECA, SCL-90, SENA, CUIDA, TAMAI, LSB-50	Estudios que no incluyan alguno de los instrumentos seleccionados

Fuentes de información

Para asegurar una recopilación rigurosa y completa de la información relacionada con las propiedades psicométricas de los instrumentos seleccionados, se consultaron diversas fuentes documentales especializadas y complementarias. La búsqueda de información se llevó a cabo

utilizando bases de datos científicas reconocidas que tienen un alto impacto en el campo de la psicología, así como otras que ofrecen cobertura regional en el contexto latinoamericano e iberoamericano.

Entre las bases especializadas se incluye EBSCO, una plataforma que da acceso a una amplia variedad de revistas científicas en psicología, educación y ciencias sociales, muchas de ellas revisadas por pares y centradas en temas de evaluación psicológica. También se usó PubMed, que si bien está enfocada en ciencias de la salud, contiene investigaciones relevantes en contextos médicos y forenses. Además, se recurrió a bases de datos multidisciplinarias como Scopus y Web of Science (WOS), que permitieron acceder a estudios de validación y adaptación de pruebas. Para complementar se utilizaron otros recursos académicos como Redalyc, SciELO, SpringerLink, ScienceDirect, Taylor & Francis, SAGE, Oxford Academic y Emerald, estas plataformas se incorporaron por su cobertura especializada en temas de psicología, derecho, salud y ciencias sociales, así como por su aporte en el acceso a investigaciones actualizadas y metodológicamente sólidas que enriquecieron el análisis del instrumento seleccionado.

Se realizaron búsquedas manuales y exhaustivas en los documentos seleccionados, con el propósito de identificar si los artículos citados en ellos también contenían información relevante y pertinente para esta investigación. Esta estrategia permitió ampliar el alcance de la revisión y asegurar una recopilación más completa y precisa de la evidencia disponible.

La combinación de bases de datos especializadas con fuentes regionales buscó garantizar un enfoque integral y contextualizado de la evidencia. Esto es especialmente importante porque muchos de los instrumentos utilizados en evaluaciones forenses en Latinoamérica han sido adaptados de otros contextos culturales. Por ello, es necesario revisar estudios que documenten estas adaptaciones y validaciones, asegurando así que los resultados sean científicamente sólidos, aplicables y pertinentes para la práctica profesional en nuestro entorno.

Estrategias de búsqueda

Se formularon ecuaciones de búsqueda en donde se utilizó operadores booleanos (AND, OR, NOT) con el fin de combinar términos claves relacionados con las pruebas, las propiedades psicométricas y la población objetivo. En este sentido, la Tabla 2 expone las ecuaciones de búsqueda o términos clave correspondiente a los 13 instrumentos utilizados en el proceso de revisión:

"CUIDA" OR "Cuestionario para la evaluación de Adoptantes, Cuidadores, Tutores y Mediadores" OR "Cuestionario para la Evaluación de Adoptantes, Cuidadores, Tutores y Mediadores" OR "Questionnaire for the Assessment of Adopters, Caregivers, Tutors and Mediators"	"Dominican Republic" OR Uruguay OR Venezuela OR "Andorra" OR "Principat d'Andorra" OR "Principado de Andorra"
"TAMAI" OR "Test Autoevaluativo Multifactorial de Adaptación Infantil" OR "Test Autoevaluativo Multifactorial de Adaptación Infantil" OR "Multifactorial Self-Assessment Test of Child Adaptation"	
"LSB-50" OR "LSB50" OR "LSB 50" OR "Listado de Síntomas Breve" OR "Listado de Síntomas Breve" OR "Brief Symptom List" OR "Brief Symptom Inventory 50"	

Selección de fuentes de evidencia

El proceso de selección se llevó a cabo en cuatro fases, manteniendo una estructura sistemática:

Fase 1: Eliminación manual de duplicados

Se realizó manualmente la depuración inicial de los artículos, de esta manera, se comparó con los títulos y autores para identificar y eliminar registros duplicados provenientes de las diferentes bases de datos consultadas.

Fase 2: Revisión de títulos y resúmenes

Este proceso se realizó por tres revisoras independientes. Para ello se enfocaron en la lectura de títulos, resúmenes y palabras clave de todos los artículos recuperados, en esta fase se aplicaron criterios de inclusión y exclusión previamente establecidos: periodo de publicación entre 2000 y 2024; idioma español e inglés; preferencia por estudios realizados en Iberoamérica; tipo de publicación como artículos científicos, informes técnicos de validación o publicaciones institucionales; población objetivo adultos sin distinción de sexo, identidad de género u orientación sexual, tipo de estudio que correspondiera a investigaciones instrumentales y estudios cuantitativos con análisis psicométrico. Así mismo, se excluyeron artículos publicados antes del año 2000, en otros idiomas, sin resultados psicométricos, centrados en población infantil o adolescente, sin relación con contextos forenses o que no excluyeran ninguno de los instrumentos seleccionados, además, los desacuerdos entre evaluadoras se resolvieron mediante discusión, si dos de las revisoras no llegaban a un acuerdo, la tercera tomaba la decisión final.

Fase 3: Revisión a texto completo

Los artículos preseleccionados fueron revisados a texto completo por tres revisoras independientes, cada artículo pasó el filtro de los criterios de inclusión y exclusión descritos en la fase anterior. Además, al igual que en la fase previa los desacuerdos se resolvieron por consenso y en caso de que persistiera la tercera revisora decidía el final.

Fase 4: Registro y trazabilidad

Por medio de la herramienta Microsoft Excel®, se realizó el registro de todas las referencias, en donde se clasificó por base de datos, siguiendo los lineamientos de inclusión y exclusión, lo que permitió mantener la trazabilidad y la transparencia.

Variables y proceso de extracción de datos

Para el manejo, compilación y síntesis de los datos se empleó un enfoque mixto, integrando procedimientos cualitativos y cuantitativos con el fin de organizar de manera sistemática y rigurosa la información de los estudios incluidos. Se diseñó un formulario estructurado en Microsoft Excel® como instrumento base para la extracción, el cual fue pilotado con varios artículos y posteriormente ajustado para optimizar las categorías y facilitar su uso. El proceso de extracción fue realizado por tres personas de manera independiente, quienes luego compararon y consensuaron sus resultados, garantizando así la consistencia, confiabilidad y exhaustividad en el tratamiento de los datos.

Las variables extraídas incluyeron once categorías principales:

- 1. Nombre del instrumento:** Registro del nombre completo de la prueba, escala, inventario o protocolo evaluado, diferenciando posibles versiones, adaptaciones o revisiones.
- 2. Estructura y características del instrumento:** Número de ítems, dimensiones o escalas, tipo de respuesta (Likert, dicotómico, etc.), modo de aplicación (individual o grupal, papel o digital), duración estimada y si es autoinforme o heteroevaluación.
- 3. Autor y año:** Autores responsables del desarrollo o adaptación del instrumento y año de publicación, para ubicar cronológicamente su vigencia.
- 4. Diseño del estudio:** Tipo de investigación (validación, adaptación transcultural, análisis factorial, análisis de confiabilidad, etc.) y enfoque metodológico (transversal, longitudinal, experimental, observacional).
- 5. Población objetivo:** Grupo al que está destinado el instrumento (población adulta general, población clínica, personas privadas de la libertad, entre otros).

6. **Características de la muestra de tipificación:** Tamaño muestral, tipo de muestreo y variables sociodemográficas relevantes (edad, sexo, nivel educativo, ocupación).
7. **Propiedades psicométricas de validez y confiabilidad reportadas:** Información sobre validez de contenido, de constructo (exploratoria o confirmatoria), convergente, discriminante o de criterio, así como indicadores de confiabilidad (alfa de Cronbach, omega, test-retest, confiabilidad interevaluador).
8. **Contexto de aplicación:** Identificación del ámbito (jurídico, forense, clínico-forense), detalle del procedimiento en el que se aplicó (evaluación de imputabilidad, custodia, reincidencia, daño psicológico) y entorno institucional (centros penitenciarios, juzgados, etc.).
9. **País o región de estudio:** Localización geográfica del estudio, para observar distribución territorial, brechas regionales y adecuación cultural.
10. **Calidad metodológica del estudio según criterios COSMIN:** Evaluación de diseño, análisis estadístico, gestión de datos faltantes, representatividad muestral y transparencia en los resultados.
11. **Limitaciones reportadas por los autores:** Restricciones metodológicas y contextuales (tamaño muestral, sesgos de selección, validez externa, ausencia de seguimiento longitudinal).

Toda la información extraída fue sistematizada en una base de datos digital, respaldada mediante software como Microsoft Excel ® y gestores bibliográficos como Zotero ®, lo que permitió una organización clara, trazable y segura de los registros. Posteriormente, se llevó a cabo una síntesis de resultados, complementado con tabulación y visualización de datos a través de tablas, a fin de identificar tendencias, patrones comunes y divergencias entre estudios. En los casos en que se presentaron inconsistencias o hallazgos contradictorios, se realizó un análisis crítico tomando en cuenta la calidad metodológica y el contexto de cada investigación, con base en los estándares internacionales de evaluación psicométrica.

Propiedades psicométricas por evaluar (revisión según COSMIN)

En el marco de esta revisión sistemática, se definieron operacionalmente las propiedades psicométricas de interés siguiendo los lineamientos del sistema COSMIN (*Consensus-based Standards for the Selection of Health Measurement Instruments*), adaptados al contexto de evaluación psicológica forense en Iberoamérica (ver Tabla 3). Estas definiciones orientaron el

análisis de la calidad técnica de los instrumentos seleccionados, garantizando la rigurosidad y la pertinencia de los hallazgos (Mokkink et al., 2010).

Tabla 3

Definiciones de dominios, propiedades de medición y aspectos de las propiedades de medición de COSMIN

Dominio	Propiedad de medición	Definición COSMIN	Definición AERA et al., 2014
1. Fiabilidad		El grado en el cual las puntuaciones de los pacientes sin cambios son las mismas en mediciones repetidas bajo diversas condiciones: por ejemplo, utilizando diferentes conjuntos de elementos en los mismos resultados informados por el paciente (HR-PRO) (consistencia interna); a lo largo del tiempo (prueba-reprueba); por diferentes personas en la misma ocasión (interevaluador) o por las mismas personas (es decir, evaluadores o respondedores) en diferentes ocasiones (intraevaluador) (Mokkink et al., 2010).	El grado en que las puntuaciones de una prueba son consistentes, estables y libres de error de medición aleatorio. Es un requisito necesario, aunque no suficiente, para la validez.
	1.1. Consistencia interna	El grado de interrelación entre los elementos (Mokkink et al., 2010).	Forma de estimar la fiabilidad que evalúa qué tan bien los ítems de una prueba miden el mismo constructo. Se basa en la correlación entre los ítems o subconjuntos de ítems dentro del mismo test.
	1.2. Confiabilidad	La proporción de la varianza total en las mediciones que se debe a diferencias "reales" entre pacientes (Mokkink et al., 2010).	Se refiere al grado en que las puntuaciones de un test son consistentes, estables y libres de error de medición aleatorio.
	1.3. Error de medición	El error sistemático y aleatorio de la puntuación de un paciente que no se atribuye a cambios reales en el constructo que se va a medir (Mokkink et al., 2010).	Diferencia entre la puntuación observada y la puntuación verdadera de una persona. Puede ser aleatorio (inconsistente) o sistemático (sesgo), y afecta la precisión de la medición.
2. Validez		El grado en que un instrumento HR-PRO mide el/los constructos(s) que pretende medir (Mokkink et al., 2010).	La validez se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de las puntuaciones de una prueba para usos propuestos de las pruebas.
	2.1. Validez de contenido	El grado en que el contenido de un instrumento HR-PRO refleja	Grado en que los ítems de la prueba representan adecuadamente el dominio del

		adecuadamente el constructo que se va a medir (Mokkink et al., 2010).	contenido que se pretende medir. Suele evaluarse mediante juicio de expertos.
	2.1.1. Validez aparente	El grado en que los ítems de un instrumento HR-PRO parecen reflejar adecuadamente el constructo que se va a medir (Mokkink et al., 2010).	No es una forma técnica de validez. Se refiere a si la prueba "parece" medir lo que dice medir, desde la perspectiva de quienes lo toman o lo aplican. Aunque no garantiza validez, puede influir en la aceptación del test.
	2.2. Validez del constructo	El grado en que las puntuaciones de un instrumento HR-PRO son consistentes con las hipótesis (por ejemplo, con respecto a las relaciones internas, las relaciones con las puntuaciones de otros instrumentos o las diferencias entre grupos relevantes) basándose en el supuesto de que el instrumento HRPRO mide válidamente el constructo que se va a medir (Mokkink et al., 2010).	Grado en que las puntuaciones de la prueba se relacionan con el constructo teórico que se pretende medir, y no con otros.
	2.2.1. Validez estructural	El grado en que las puntuaciones de un instrumento HR-PRO reflejan adecuadamente la dimensionalidad del constructo que se va a medir (Mokkink et al., 2010).	Forma de evidencia de validez del constructo. se ajusta al modelo teórico del constructo medido. Suele examinarse mediante análisis factorial.
	2.2.2. Validez transcultural	El grado en que el desempeño de los ítems en un instrumento HR-PRO traducido o adaptado culturalmente refleja adecuadamente el desempeño de los ítems de la versión original del instrumento HR-PRO (Mokkink et al., 2010).	Grado en que el constructo medido tiene el mismo significado y estructura en diferentes contextos culturales o lingüísticos.
	2.3. Validez de criterio	El grado en que las puntuaciones de un instrumento HR-PRO reflejan adecuadamente un "estándar de oro" (Mokkink et al., 2010).	Grado en que las puntuaciones de la prueba se relacionan con un criterio externo relevante.
3. Sensibilidad		La capacidad de un instrumento HR-PRO para detectar cambios a lo largo del tiempo en el constructo que se va a medir (Mokkink et al., 2010).	Capacidad de una prueba para detectar diferencias o cambios significativos en el constructo medido.
4. Interpretabilidad		La interpretabilidad es el grado en el que se puede asignar un significado cualitativo (es decir, connotaciones clínicas o comúnmente entendidas) a las puntuaciones cuantitativas de un instrumento o a los cambios en las puntuaciones (Mokkink et al., 2010).	Grado en que se puede asignar un significado cualitativo a las puntuaciones de una prueba. Incluye la comprensión de qué representan las puntuaciones y qué implicaciones tienen para la toma de decisiones.

Nota. Esta tabla presenta la traducción y organización de los dominios y propiedades de medición definidos por el Consorcio COSMIN (*Consensus-based Standards for the Selection of Health Measurement Instruments*). Las definiciones fueron adaptadas del documento original: COSMIN (Mokkink et al., 2010).

Características contextuales de los estudios

Las características contextuales cumplen un papel fundamental para interpretar adecuadamente la aplicabilidad y pertinencia de las propiedades psicométricas de los instrumentos seleccionados en escenarios jurídicos y forenses. En primer lugar, se tomó en cuenta el país en el que fue realizado cada estudio. Esta información fue esencial dado que muchas pruebas psicométricas han sido desarrolladas en contextos anglosajones, por lo que su uso en Latinoamérica requiere procesos de adaptación cultural y validación local. Identificar si la evidencia proviene de países iberoamericanos permitió juzgar la validez ecológica y la equivalencia cultural de los instrumentos cuando se aplican en contextos similares al colombiano.

En segundo lugar, se analizó la población objetivo-reportada en los estudios, entendiéndola como las características sociodemográficas y clínicas de las personas evaluadas. Esto implicó identificar si la muestra estaba compuesta por adultos, sin restricción de sexo, identidad de género u orientación sexual. También se consideró si pertenecían a poblaciones generales, forenses, clínicas, comunitarias u otras. Adicionalmente, se tuvo en cuenta si las muestras correspondían a procesos de tipificación normativa, validaciones en poblaciones específicas o estudios comparativos. Esta información permitió valorar el grado de generalización de los hallazgos y su pertinencia para poblaciones similares.

Por último, se extrajo información sobre el entorno de aplicación, entendido como el contexto institucional o profesional donde se administraron los instrumentos. Esto incluyó si la prueba fue utilizada en consultorios forenses, juzgados, instituciones penitenciarias o en entornos académicos con fines investigativos. Esta variable permitió comprender las condiciones de aplicación, el nivel de control metodológico y la semejanza entre el contexto del estudio y los espacios donde actualmente se realiza la evaluación psicológica forense en Colombia.

Variables metodológicas de los estudios primarios

Para asegurar la rigurosidad en la síntesis de los hallazgos, se extrajeron diversas variables metodológicas de los estudios primarios incluidos. Una de ellas fue el diseño del estudio, con énfasis en investigaciones de tipo instrumental centradas en la evaluación psicométrica. Esta información permitió establecer el nivel de evidencia metodológica de cada trabajo y su aporte a la confiabilidad del análisis general.

También se registró el año de publicación, restringido al periodo comprendido entre 2000 y 2024. Este criterio temporal buscó asegurar la vigencia científica de los datos recopilados, así como su relevancia para la práctica profesional contemporánea. A su vez, se documentó el nombre del autor o los autores principales, lo cual facilitó la trazabilidad, citación y verificación de los estudios incluidos.

Otra variable fundamental fue el instrumento evaluado, de entre los trece seleccionados previamente, ya que la revisión estaba organizada en función de cada uno de estos. Además, se recopilaron los métodos estadísticos empleados para evaluar las propiedades psicométricas, tales como análisis factoriales, coeficientes de confiabilidad, análisis de validez, entre otros, junto con las interpretaciones ofrecidas por los autores sobre dichos resultados.

Finalmente, se documentaron las limitaciones metodológicas y contextuales reportadas por los autores de cada estudio. Esta información fue clave para desarrollar una lectura crítica de la evidencia disponible y para formular recomendaciones ajustadas a las condiciones reales del ejercicio psicológico en el ámbito jurídico-forense colombiano y latinoamericano.

Síntesis de resultados

Enfoque de análisis (narrativo, tabulación, análisis temático)

El análisis cualitativo se realizó mediante un enfoque temático, categorizando los instrumentos según el tipo de evaluación forense que abordan (riesgo, imputabilidad, simulación). Esto permitió identificar patrones de uso y utilidad en el campo jurídico-forense. Paralelamente, se aplicó un análisis narrativo para contextualizar los hallazgos de cada instrumento en relación con su aplicación forense y la evidencia empírica que los respalda. En el componente cuantitativo, se realizó una tabulación comparativa de las propiedades psicométricas (validez, confiabilidad, sensibilidad-especificidad), a fin de sintetizar y contrastar la solidez técnica de cada prueba.

Estrategias para presentación visual de resultados (tablas, gráficos)

Para facilitar la visualización de los hallazgos, se utilizaron tablas resumen que permitieron comparar las propiedades psicométricas entre los distintos instrumentos. Asimismo, se elaboraron cuadros comparativos que integraron datos clave en relación con la aplicabilidad forense. Se diseñaron también mapas conceptuales que mostraron la relación entre los

instrumentos y los tipos de evaluación que abordaban. Estas herramientas visuales buscaron ofrecer una visión integral, clara y didáctica de los resultados.

Síntesis de las propiedades psicométricas por instrumento

Se presentó un perfil individual por instrumento, en el cual se describieron sus principales características técnicas y contextuales. Estos perfiles incluyeron aspectos como el tipo de evaluación que realiza, población objetivo, propiedades psicométricas reportadas (validez, confiabilidad, sensibilidad, especificidad), y su grado de aceptación en la literatura científica. Se destacaron también las fortalezas y limitaciones de cada prueba con base en la evidencia disponible.

Abordaje de las inconsistencias entre estudios

Las inconsistencias entre estudios fueron tratadas mediante un análisis crítico de las posibles fuentes de variación, tales como diferencias metodológicas, características de las muestras utilizadas, o variabilidad en los contextos de aplicación. Se evaluó el grado de heterogeneidad en los resultados, lo que permitió ofrecer una interpretación rigurosa, reflexiva y contextualizada de los hallazgos, reconociendo los límites del cuerpo de evidencia disponible.

Resultados

En el siguiente apartado se presentan los resultados obtenidos a partir de la revisión sistemática de los estudios psicométricos de los principales instrumentos de evaluación psicológica utilizados en el consultorio Psicoforense de la Universidad Santo Tomás. El objetivo de esta sección es ofrecer una síntesis clara y comparativa de la evidencia disponible sobre las propiedades psicométricas de cada instrumento, considerando aspectos como su confiabilidad, validez, estructura factorial y aplicabilidad en contextos iberoamericanos.

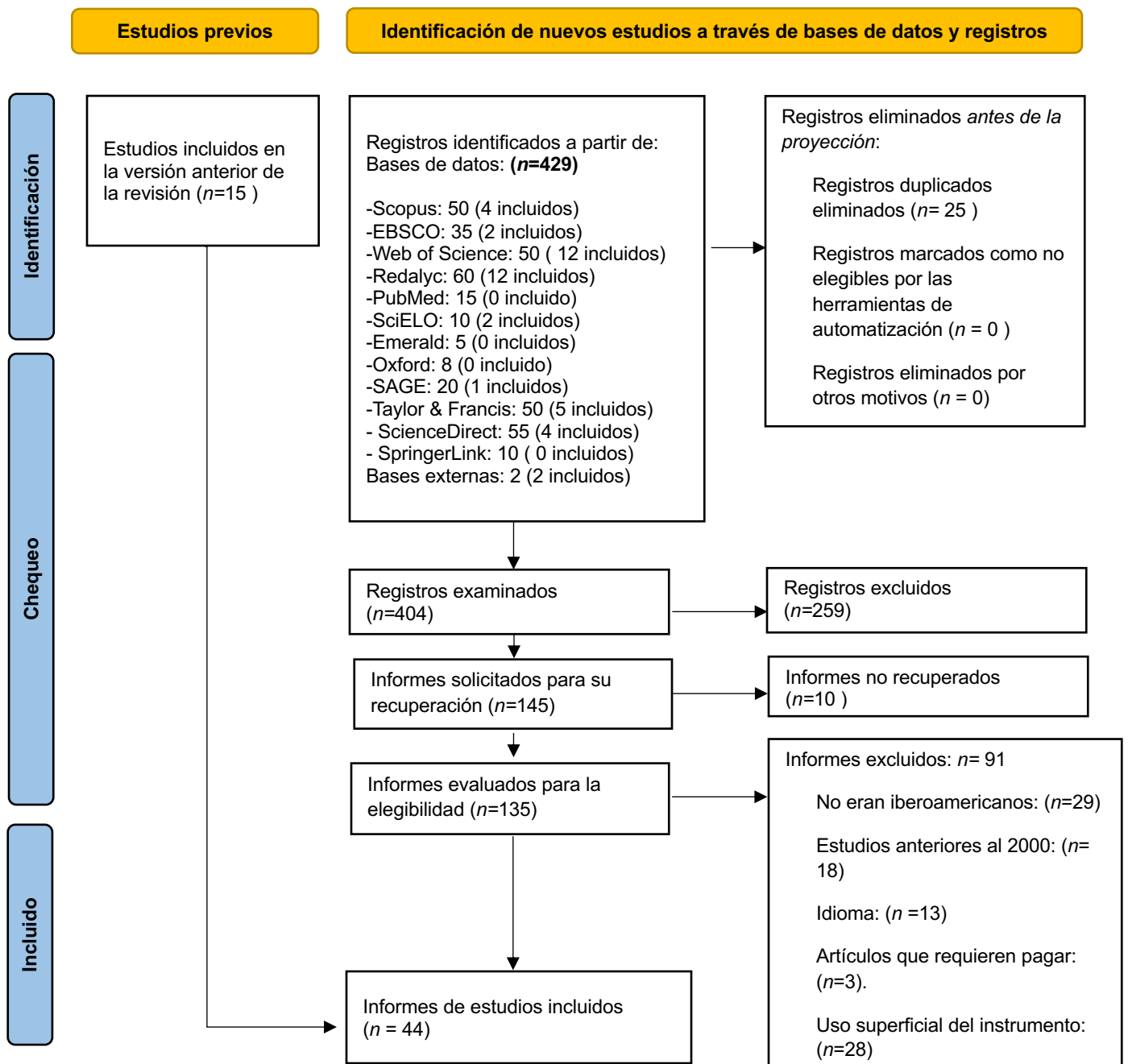
Cada instrumento se analiza siguiendo una misma estructura de análisis que incluye una descripción general, los datos técnicos más relevantes (autoría, año, formato, número de ítems y dimensiones), las evidencias de validez y confiabilidad reportadas en los estudios revisados y la evaluación de la calidad metodológica según los criterios COSMIN. Este formato permite reconocer las principales fortalezas, limitaciones y nivel de evidencia de cada prueba, facilitando su comparación y la valoración de su utilidad en la práctica profesional.

Los instrumentos incluidos en esta síntesis fueron seleccionados por su relevancia en la evaluación de la personalidad, la psicopatología y las funciones neuropsicológicas. Entre ellos se encuentran: MMPI-2, MMPI-2-RF, PAI, SIMS, NEUROPSI, BANFE, TECA, SCL-90-R,

SENA, CUIDA, TAMAI y LSB-50. Cabe resaltar que del instrumento Inventario Clínico Multiaxial de Millon-IV (MCMI-IV) no se identificaron registros que cumplieran con los criterios de selección. En conjunto, los resultados que se presentan a continuación ofrecen una visión general sobre la consistencia psicométrica y la aplicabilidad transcultural de estos instrumentos en el contexto iberoamericano actual.

Tabla 4

Diagrama de flujo Prisma de la revisión sistemática



La tabla presenta la caracterización de los instrumentos psicométricos incluidos en una revisión compuesta por 44 estudios, detallando para cada prueba el número de artículos seleccionados, los años de publicación, las versiones utilizadas, los países donde se realizaron las investigaciones, las poblaciones evaluadas y la calidad metodológica según los criterios COSMIN. Entre los instrumentos analizados se encuentran el MMPI-2, el MMPI-2-RF, el NEUROPSI, el SIMS, el BANFE, el CUIDA, el LSB-50, el PAI, el SCL-90, el SENA, el TAMAI y el TECA, aplicados en diversos contextos clínicos, forenses, educativos y comunitarios en países de Latinoamérica y Europa, especialmente España, México, Argentina, Colombia, Portugal, Brasil y Chile.

En términos generales, la tabla evidencia que la mayoría de los estudios reportan niveles de calidad metodológica entre moderada y buena, con fortalezas frecuentes en validez de constructo, validez discriminante y consistencia interna. No obstante, también se identifican limitaciones recurrentes, como tamaños muestrales reducidos en algunos subgrupos clínicos, ausencia de análisis factorial confirmatorio, falta de estudios de estabilidad temporal (test–retest) y escasa evaluación de invarianza factorial. En conjunto, la información sintetizada permite observar el estado actual de la evidencia psicométrica de estos instrumentos en población hispanohablante, destacando tanto sus aportes como los aspectos que requieren mayor desarrollo investigativo.

Tabla 5

Caracterización de los instrumentos psicométricos incluidos en la revisión (44 estudios).

Instrumento	Número de artículos seleccionados	Años de publicación	Versiones del instrumento	Países en que se realizaron los estudios	Poblaciones y muestras	Calidad COSMIN
(MMPI-2)	Seis estudios Scott & Pampa (2000)		Adaptación española del MMPI-2;		Pacientes con fibromialgia;	Calidad moderada–alta en validez (discriminante y concurrente reportadas en varios estudios); evaluación estructural limitada; confiabilidad no reportada de forma consistente; aplicación y análisis adecuados.
	1. Garrido Amadeo et al. (2012)	2000: 1	versión hispana del MMPI-2 (García-Peltoniemi &	España	pacientes con dolor crónico de etiología orgánica; grupo control sano (N total=105: 27 fibromialgia, 44 dolor crónico orgánico, 34 controles).	
	2. Palmera et al. (2013)	2012: 1	Azán-Chaviano, 1993)	Puerto Rico	Adultos evaluados en contextos clínicos, comunitarios y forenses (N total=153: pre-empleo 64;	
	3. Osuna et al. (2015)	2013: 1		México	agresores domésticos 28; comunidad 29; policías municipales 32).	
	4. del Barrio et al. (2016)	2015: 1 2016: 1 2018: 1	MMPI-2 estándar (567 ítems, formato Verdadero/Falso)	Perú	Reclusos homicidas y jóvenes sin antecedentes delictivos (N	

	5. Ampudia Rueda et al. (2018)		Versión adaptada para población mexicana.		total=574: 287 homicidas; 287 no-delictivos). Adultos peruanos de diversas edades, niveles educativos, ocupacionales y socioeconómicos (N=123). Internos penitenciarios con trastorno mental grave/incompetentes y grupo contraste de penados no psiquiátricos (N=202: 102 psiquiátricos; 100 no psiquiátricos). Adultos atendidos en unidad de epilepsia con sospecha de crisis epilépticas o no epilépticas psicógenas (N total=209: PNES 37; ES 172).	
(MMPI-2-RF)	Siete estudios		MMPI-2-RF versión española		Personas evaluadas en contexto forense (muestras clínicas y no clínicas de distintos países; N total agregado en estudios incluidos); Personas fumadoras en tratamiento para cesación tabáquica en México (N≈300); Pacientes con fibromialgia (litigantes y no litigantes), grupo de simulación análoga y grupo control (N≈200); Adultos bilingües (portugués-inglés) en estudio de adaptación psicométrica (N≈150); Padres y madres en procesos de custodia de hijos (N≈100–120); Médicos residentes en formación (N≈250); Personas adultas evaluadas en contexto forense (N≈180–200).	Evidencia favorable de validez de constructo y discriminante; sensibilidad/especificidad adecuadas en estudios forenses. Confiabilidad interna aceptable reportada en algunas muestras; faltan estudios longitudinales y test–retest. Calidad metodológica general: moderada–alta.
	1. Capilla Ramírez et al. (2013)		MMPI-2-RF adaptación mexicana			
	2. Jiménez-Gómez et al. (2013)					
	3. Martínez et al. (2017)	2013: 2 2017: 2	MMPI-2-RF versión portuguesa (adaptación y equivalencia bilingüe)	España Portugal México		
	4. Sánchez et al. (2017)	2019: 1 2023: 1				
	5. Redondo et al. (2019)	2024: 1				
	6. Rodríguez-Aguirre et al. (2023)		MMPI-2-RF versión original (inglés) usada como referencia en estudios de adaptación			
(NEURO PSI)	Cuatro estudios		Versión original en español		Adultos mayores con sospecha o diagnóstico de demencia tipo Alzheimer y grupo control sano (N clínico y N control reportados); Pacientes con esquizofrenia y grupo control sano (N clínico y N control reportados); Adultos sanos argentinos de distintas edades y niveles educativos para baremación normativa (N total de tipificación reportado); Adultos hispanohablantes (México) para validación y estandarización del instrumento (N total reportado).	Buena. Evidencia adecuada de validez (constructo, discriminante y concurrente) y confiabilidad interna aceptable; limitaciones puntuales por tamaño muestral en subgrupos clínicos.
	1. Picasso et al. (2005)		Estandarización para población mexicana			
	2. Ostrosky-Solís et al. (2007)	2005: 1 2007: 1	Adaptación/bare mación para población argentina	México Argentina Brasil		
	3. Abrisqueta-Gomez et al. (2008)	2008: 1 2012: 1	Versión clínica para evaluación de trastornos neurocognitivos			
	4. Querejeta et al. (2012)					
(SIMS)	Seis estudios					moderada–alta. Evidencia sólida de validez discriminante/diagnóstica y convergente en contextos forenses y clínicos. Confiabilidad interna alta en adaptaciones
	1. Martín-Albo, Núñez & Navarro (2009)	2009: 1 2013: 1 2017: 1 2020: 1 2022: 1 2025: 1	SIMS original en español (75 ítems, V/F)	España Internacional (muestras de distintos países)	Adultos sanos e individuos en contexto forense/psicológico donde se evalúa simulación (N reportado por estudio); Pacientes reales, diagnosticados de trastorno mixto ansioso-	

	2. Pallardó Durá (2013)		SIMS adaptación española		depresivo o adaptativo, derivados por mutuas de accidente de trabajo para tratamiento psicológico (N = 26);	españolas; limitaciones en algunos estudios por tamaños muestrales pequeños y falta de reporte de estructura interna o test-retest.
	3. Marín Torices et al. (2017)		SIMS utilizado en contextos clínicos y forenses (puntos de corte 10, 14, 16)		Mujeres adultas víctimas de violencia de pareja íntima (IPV) en España y grupo control de mujeres sin antecedentes de violencia (N total = 108; 68 víctimas IPV, 40 controles);	
	4. Blasco Saiz & 5. Puente-López et al. (2020)		SIMS en meta-análisis internacional (muestras clínicas)		Estudiantes universitarios para evaluar motivación situacional en contexto educativo (N = 373);	
	6. Puente-López et al. (2022)				Pacientes evaluados por daño físico y psicológico tras accidentes de tráfico y grupos control (población general y malingerers instruidos) (N total = 650).	
	7. Puente-López et al. (2025)					
BANFE	Dos estudios 1. Puerta-Lopera et al. (2022) 2. Doldán-Peluffo et al. (2023)	2022: 1 2023: 1	BANFE-2 (versión 2017, Flores, Ostrosky y Lozano). Versiones estándar adaptadas al español/colombiano de: WCST, COW-FAS, Stroop (SCW), TMT-B	Colombia Paraguay	1: Personas adultas diagnosticadas con esclerosis múltiple tipo recaída-remisión (EMRR), de ambos sexos, residentes en Paraguay.; (N=40) 2: Estudiantes universitarios. (N=208.)	1: Moderada-baja. Muestra pequeña (n = 40), sin EFA/CFA, sin análisis de invarianza ni error de medida. 2: Alta-moderada: tamaño muestral adecuado para ACP (N = 208), uso de KMO/Bartlett, test-retest (n = 50), análisis interevaluador, cálculo de baremos; limitación en validez concurrente con criterios
CUIDA	Un estudio 1. Arredondo Londoño et al. (2019)	2019	CUIDA (2008).	Colombia	Padres en proceso de adopción nacional en Colombia. (N=80)	Moderada (muestra pequeña, pero análisis adecuados, SEM/validación estructural sólida).
LSB-50	Dos estudios 1. Abuín, M. R., & de Rivera, L. (2014) 2. de la Iglesia et al. (2015)	2013: 1 2014: 1	LSB-50	Argentina, España	1: Adultos población general y clínica. (N=994) 2: Adultos población general y clínica. (N=1.242)	La calidad metodológica fue clasificada como Moderada 2: La calidad metodológica fue clasificada como Buena,
PAI	Cuatro estudios 1. Stover J. B., et al. (2016) 2. Ortiz-Tallo et al., (2017) 3. Abilleira, M. P., & Rodicio-García, M. L. (2018).	2016: 1 2017: 1 2018: 1 2024: 1	1: Versión española adaptada por Ortiz-Tallo y cols. (Morey, 2011). 2: Versión portuguesa adaptada del PAI 3: Adaptación argentina del Personality	Chile, con participante s de las 13 regiones Portugal Argentina España	1: Adultos chilenos mayores de 18 años, tanto población general como población clínica, 691 participantes (564 población general y 127 clínica, de un total inicial de 803). 2: Población adulta general de Portugal (muestra destinada a normativizar la versión portuguesa del PAI). 900 participantes (muestra final normativa)	1: Fiabilidad interna: adecuada (alfas .72-.86 en población general; .77-.86 en clínica). Reproducibilidad (test-retest / error de medición): no evaluada. Validez de contenido: moderada (basada en adaptación española, sin validación cultural chilena). Validez estructural (análisis factorial): no evaluada.

4. Paulino M et al. (2024)		Assessment Inventory (PAI)	3: Estudiantes universitarios de instituciones públicas y privadas, 374 participantes	Validez convergente/divergente: no evaluada en la muestra chilena.		
		4: Versión abreviada (PAI-R / reduced version — primeras 165 afirmaciones del PAI completo).	4: Personas privadas de libertad (internos adultos) en centros penitenciarios de la Comunidad Autónoma de Galicia (España); Muestra total n = 294. (Piloto EFA: n = 81).	Sensibilidad al cambio: no evaluada. Síntesis global: calidad metodológica moderada-baja, sólida solo en fiabilidad interna.		
				2: Moderada: adecuada para normas y consistencia interna; limitada por ausencia de análisis factorial confirmatorio, invarianza, test–retest y validez convergente/discriminante.		
				3: Nivel moderado: buen tamaño muestral, análisis adecuados, reporte claro de confiabilidad y estructura; limitaciones: sin evaluación de estabilidad temporal ni validez convergente/discriminante		
				4: Evaluación breve: calidad metodológica moderada. Justificación: positivos — tamaño de muestra razonable (n=294) y uso de EFA+CFA, análisis de consistencia interna y validez concurrente; limitaciones — diseño transversal, bajo número de mujeres, EFA piloto pequeño (n=81), ausencia de información sobre error de medición, estabilidad temporal (test-retest) y sensibilidad/responsividad. (Basado en la información del artículo).		
SCL-90	Cinco estudios	El SCL-90 cuenta con su versión original, la versión revisada SCL-90-R,				
	1. Cruz Fuentes et al. (2005)			Las poblaciones evaluadas incluyen pacientes psicológicos y psiquiátricos mayores de 14 años, adultos en proceso de psicoterapia, mujeres adultas, población general y personas privadas de libertad. Las muestras varían entre 214 pacientes clínicos, 570 adultos en admisión a psicoterapia, 289 mujeres en grupos clínicos y comunitarios, y 427 internos en centros penitenciarios.		
	2. Sandín et al. (2008)	2005: 1	Colombia	La mayoría de los estudios presentan calidad metodológica entre buena y alta; no obstante, algunos muestran calidad moderada debido al uso de muestreo por conveniencia y ausencia de análisis de estabilidad temporal (test-retest).		
	3. Sánchez & Ledesma (2009)	2008: 1	España			
	4. Ignatyev et al. (2016)	2009: 1	Argentina			
	5. Londoño et al. (2018)	2016: 1	México			
		2018: 1	Chile			
		adaptaciones al castellano y español, así como versiones abreviadas, incluyendo una forma de 36 ítems derivada mediante análisis factorial exploratorio.				
SENA	Un estudio	2022	SENA 6–12 años.	El estudio incluido en la matriz	La población estuvo conformada por niños escolarizados entre 6 y 12 años. La muestra fue de 140	La calidad metodológica fue clasificada como buena, evidenciando adecuada

	1. Rossignoli - Palomeque & Pérez-Hernández (2022)			fue realizado en España, específicamente en Madrid.	participantes, compuesta por 65 niñas y 75 niños, pertenecientes a un nivel sociocultural medio-alto.	consistencia interna, validez convergente y confirmación de la estructura factorial.
TAMAI	Un estudio 1. Moreno-Manso et al. (2010)	2010	BLOC-Screening; TAMAI estándar.	España	La población estuvo conformada por niños en acogimiento residencial bajo medidas de protección. La muestra estuvo integrada por 74 niños, de los cuales 41 eran varones y 33 mujeres.	La evaluación metodológica según COSMIN fue moderada, debido a que, aunque se emplearon instrumentos estandarizados, el estudio presentó limitaciones como la ausencia de grupo control.
TECA	Cinco estudios 1. Gorostiaga et al. (2014) 2. López-Pérez et al. (2014) 3. Freitas & Miranda (2021) 4. Lemos et al. (2022) 5. Chavira Trujillo & Celis de la Rosa (2022)	2014: 2 2021: 1 2022: 2	El TECA cuenta con diversas versiones y adaptaciones, entre ellas la versión original, la versión brasileña (TECA-BR), versiones adaptadas al euskera y castellano para población infantil y adolescente, versiones juveniles, una versión breve y una versión final brasileña de 31 ítems.	España Brasil Argentina	Las poblaciones evaluadas incluyen niños, adolescentes, jóvenes y adultos universitarios. Las muestras reportadas comprenden 670 niños y adolescentes españoles entre 10 y 16 años, 1.618 escolares entre 10 y 15 años, 256 estudiantes universitarios brasileños y muestras de universitarios argentinos y brasileños adultos jóvenes de ambos sexos.	La calidad metodológica según los criterios COSMIN es mayoritariamente buena, con algunos estudios clasificados como adecuados y uno con calidad moderada debido a limitaciones como muestreo por conveniencia y ausencia de análisis test-retest e invarianza factorial.

Síntesis de evidencia por instrumento

Instrumento I. Minnesota Multiphasic Personality Inventory-2 (MMPI-2)

El *Minnesota Multiphasic Personality Inventory-2* (MMPI-2) es uno de los instrumentos más utilizados para la evaluación de la personalidad y la psicopatología en adultos. Consta de 567 ítems de respuesta dicotómica (Verdadero/Falso) y permite obtener perfiles clínicos y de validez que aportan información diagnóstica relevante en contextos clínicos, forenses y de investigación. Su uso se ha extendido ampliamente en países iberoamericanos, donde se han realizado estudios de adaptación y análisis psicométrico, principalmente en España, México, Perú y Puerto Rico, mostrando una aplicabilidad cultural adecuada y solidez teórica (Scott & Pampa, 2000).

Los estudios revisados indican que el MMPI-2 presenta un comportamiento psicométrico estable, con niveles de consistencia interna satisfactorios y evidencias sólidas de validez. En cuanto a la validez de contenido, las investigaciones sobre las versiones en español confirman la

conservación de la equivalencia conceptual respecto al instrumento original. En el caso peruano, Scott y Pampa (2000) reportaron una alta similitud entre las puntuaciones de su muestra y las normas estadounidenses, con diferencias mínimas en las escalas F y Mf, atribuibles a variaciones culturales menores que no comprometen la estructura general del test.

Respecto a la validez de constructo, los hallazgos muestran una replicación aceptable de la estructura original. En una muestra puertorriqueña, Garrido Amadeo et al. (2012) observaron que tanto las escalas clínicas como las reestructuradas presentan coherencia interna y adecuada capacidad discriminativa entre grupos, lo que respalda la consistencia estructural del instrumento. Resultados similares se han documentado en Perú, donde se reportaron correlaciones interescalares comparables a las del grupo normativo estadounidense (Scott & Pampa, 2000). No obstante, la mayoría de los estudios iberoamericanos no incluyen análisis factorial confirmatorio, lo que limita afirmar de manera concluyente la equivalencia estructural entre culturas.

En relación con la validez criterial, los estudios revisados ofrecen evidencia robusta en poblaciones clínicas y forenses. En España, del Barrio et al. (2016) reportaron que el MMPI-2 presenta una capacidad diagnóstica moderada-alta para distinguir crisis no epilépticas psicógenas, con un área bajo la curva (AUC) = .84, sensibilidad = 77.1% y especificidad = 76.8%. En México, Ampudia Rueda et al. (2018) hallaron valores de $AUC \geq .8$ para las escalas Pa, Ma y Pd al comparar población homicida y no delictiva, lo que respalda su validez predictiva en la evaluación del riesgo de violencia. Asimismo, en España, Osuna et al. (2015) identificaron que las escalas de validez F, K, Fb, Fp, Ds y FBS clasifican de forma adecuada los perfiles de simulación y disimulación. Por su parte, Palmer et al. (2013) encontraron que, en pacientes con dolor crónico y fibromialgia, estas escalas tienden a sobre diagnosticar simulación, aunque mantienen utilidad para detectar defensividad, coincidiendo con los hallazgos forenses previos.

En cuanto a la confiabilidad, los estudios reportan coeficientes alfa de Cronbach (α) entre .7 y .94, valores que reflejan adecuada consistencia interna en la mayoría de las escalas. Garrido Amadeo et al. (2012) informaron α entre .82 y .93 en las escalas principales de la muestra puertorriqueña, mientras que Osuna et al. (2015) y Ampudia Rueda et al. (2018) reportaron niveles similares en poblaciones clínicas y penitenciarias. Sin embargo, la evidencia sobre estabilidad temporal (test-retest) e interevaluador es limitada, lo que impide una confirmación completa de la fiabilidad en estudios longitudinales.

La evaluación metodológica basada en los criterios COSMIN ubica la calidad de la evidencia como moderada, destacando fortalezas en la validez criterial y en la consistencia interna, pero con limitaciones en el análisis factorial confirmatorio y la confiabilidad temporal. Los tamaños muestrales de los estudios oscilaron entre $n = 102$ y $n = 574$, abarcando poblaciones comunitarias, clínicas y forenses. En conjunto, los resultados indican que el MMPI-2 mantiene una estructura psicométrica sólida y aplicable a contextos hispanohablantes, siendo especialmente relevante para el diagnóstico diferencial, la detección de simulación y la predicción de conductas de riesgo. Aunque se requieren estudios adicionales con análisis factoriales confirmatorios y diseños longitudinales, la evidencia existente respalda su uso clínico y forense con niveles adecuados de validez y confiabilidad.

Tabla 6*Características técnicas del MMPI-2*

Sección	Descripción
Datos generales	Nombre: <i>Minnesota Multiphasic Personality Inventory-2</i> (MMPI-2) Autor original: Starke R. Hathaway & J. Charnley McKinley Año: 1989 (revisión del MMPI original de 1943) Constructo: Evaluación global de la personalidad y psicopatología; detecta patrones clínicos, rasgos de personalidad y validez de respuesta. Número de ítems: 567 ítems (MMPI-2); versiones abreviadas y reestructuradas (MMPI-2-RF con 338 ítems). Formato: Respuesta dicotómica (Verdadero/Falso). Adaptación cultural: Traducciones y adaptaciones validadas en diversos países iberoamericanos (España, Perú, Puerto Rico, México, Chile, entre otros).
Estructura	Consta de 3 tipos de escalas: 1. Validez: L (Mentira), F (Infrecuencia), K (Corrección), Fb, Fp, VRIN, TRIN, entre otras. 2. Clínicas: Hs (Hipocondría), D (Depresión), Hy (Histeria), Pd (Desviación psicopática), Mf (Masculinidad/Feminidad), Pa (Paranoia), Pt (Psicastenia), Sc (Esquizofrenia), Ma (Hipomanía), Si (Introversión social). 3. Contenido y suplementarias: escalas específicas de ansiedad, obsesividad, ira, cinismo, problemas familiares, etc. También se incluyen las escalas reestructuradas (RC), base del MMPI-2-RF (Tellegen et al., 2003).
Descripción de los estudios seleccionados	Se revisaron seis estudios iberoamericanos publicados entre 2000 y 2018: tres en España (del Barrio et al., 2016; Osuna et al., 2015; Palmer et al., 2013), uno en México (Ampudia Rueda et al., 2018), uno en Perú (Scott & Pampa, 2000) y uno en Puerto Rico (Garrido Amadeo et al., 2012). Las muestras abarcaron poblaciones clínicas, comunitarias, forenses y penitenciarias, con tamaños que oscilaron entre 102 y 574 participantes. Los estudios se centraron en evaluar la estructura interna, la validez de criterio, la capacidad predictiva y las propiedades de confiabilidad del instrumento.
Evidencias de validez	La mayoría de las investigaciones coinciden en la validez transcultural del MMPI-2.

	En Perú, Scott y Pampa (2000) hallaron gran similitud entre las normas locales y las estadounidenses, con leves diferencias en F y Mf. En Puerto Rico, Garrido Amadeo et al. (2012) reportaron coherencia entre las escalas clínicas y reestructuradas, confirmando convergencia teórica. En España, del Barrio et al. (2016) evidenciaron validez criterial con AUC = .84, Sensibilidad = 77.1%, Especificidad = 76.8% en diagnóstico diferencial de crisis no epilépticas psicógenas. En México, Ampudia Rueda et al. (2018) encontró $AUC \geq .8$ en escalas Pa, Ma y Pd para predecir riesgo de violencia. En contextos forenses españoles, Osuna et al. (2015) confirmaron la capacidad de las escalas F, K, Fb, Fp, Ds y FBS para detectar simulación, mientras que Palmer et al. (2013) identificaron sobrediagnóstico de simulación en pacientes con dolor crónico, aunque con alta sensibilidad para defensividad.
Evidencias de confiabilidad	Los coeficientes alfa de Cronbach reportados se situaron entre .7 y .94, con consistencia interna adecuada en la mayoría de las escalas. Garrido Amadeo et al. (2012) reportaron $\alpha = .82-.93$ en las escalas principales. En estudios clínicos y penitenciarios (Ampudia Rueda et al., 2018; Osuna et al., 2015), los valores fueron similares, mostrando estabilidad y precisión interna. No se registraron análisis test-retest ni interevaluador en los estudios revisados.
Hallazgos según criterios COSMIN	La evidencia psicométrica del MMPI-2 se califica como moderada, con fortalezas en la consistencia interna y la validez de criterio, pero con limitaciones en la evidencia de estabilidad temporal y análisis factorial confirmatorio. Los estudios muestran replicabilidad aceptable, coherencia entre contextos culturales y adecuada sensibilidad clínica y forense. Se recomienda fortalecer futuras investigaciones con diseños longitudinales y validaciones cruzadas entre países.

Instrumento 2. Inventario Multifásico de Personalidad de Minnesota-2 Reestructurado (MMPI-2-RF)

El instrumento *Minnesota Multiphasic Personality Inventory-2, Restructured Form* (MMPI-2-RF) es la versión reestructurada y abreviada del MMPI-2 que, según los estudios revisados, fue implementada en la región iberoamericana con el objetivo de mejorar la coherencia factorial y reducir la redundancia ítem-ítem de la versión previa, manteniendo su función central de evaluación de rasgos de personalidad y psicopatología en adultos (Afonso et al., 2024; Capilla Ramírez et al., 2013; Redondo et al., 2019). De esta manera, en las adaptaciones y validaciones locales se indica que la versión utilizada contiene 338 ítems de respuesta dicotómica (Verdadero/Falso) y una estructura jerárquica compuesta por escalas de validez (p. ej., VRIN-r, TRIN-r, F-r, Fp-r, FPs, FBS-r, L-r, K-r), nueve escalas RC, escalas de nivel superior (EID, THD, BXD), escalas de problemas específicos y el conjunto PSY-5. El tiempo de aplicación informado en los trabajos oscila entre 35 y 50 minutos; asimismo, las muestras en los estudios revisados corresponden mayoritariamente a población adulta, y los

contextos de aplicación incluyen ámbitos clínicos, forenses, ocupacionales y poblacionales. La administración e interpretación se restringe a profesionales formados en evaluación psicológica (Afonso et al., 2024; Capilla Ramírez et al., 2013; Jiménez-Gómez et al., 2013; Martínez et al., 2017; Redondo et al., 2019; Rodríguez-Aguirre et al., 2023; Sánchez et al., 2017).

En cuanto a la validez, se muestran hallazgos diferenciados por tipo de evidencia. Respecto a la validez de contenido, la mayoría de los estudios locales no reporta procedimientos cuantitativos formales (p. ej., *I-CVI*, *S-CVI*, panel de jueces con métricas), por lo que la evidencia de contenido en la literatura iberoamericana examinada es insuficiente. Afonso et al. (2024) describen procesos de traducción y equivalencia lingüística (incluido análisis Rasch), pero sin presentar índices de acuerdo experto cuantificados. En cuanto a la validez relacional (convergente, concurrente y discriminante), hay evidencia contextual y parcial: Martínez et al. (2017) documentan validez predictiva de escalas específicas (p. ej., DSF) en una muestra de $n = 281$ fumadores. Jiménez-Gómez et al. (2013) y Sánchez et al. (2017) informan análisis diagnósticos (AUC, sensibilidad, especificidad) que respaldan la capacidad discriminante de determinadas escalas en muestras clínicas y forenses. Redondo et al. (2019), en su revisión/metaanálisis, confirman capacidad discriminante en escalas clínicas y de validez, pero advierten variabilidad asociada con moderadores (tamaño muestral, criterio externo).

Sobre la evidencia de validez de constructo, los análisis factoriales y/o procedimientos Rasch reportados por Afonso et al. (2024) y Capilla Ramírez et al. (2013) muestran estructuras compatibles con el modelo RC y cargas factoriales coherentes en la mayoría de escalas, aunque en algunos ítems se identificaron ajustes deficientes, y no se encontraron pruebas sistemáticas de invarianza factorial multigrupo (por sexo, edad o contexto), lo que limita la afirmación de equivalencia métrica entre subpoblaciones.

Por otra parte, en términos de confiabilidad, se determina que la evidencia más sólida para consistencia interna corresponde a los coeficientes α de Cronbach reportados; estos varían según escala y estudio, con un rango aproximado de $\alpha \approx .57-.96$. Como ejemplos, Redondo et al. (2019) reportan $Hs \alpha = .79$, $D \alpha = .81$, $Hy \alpha = .78$, $Pd \alpha = .71$, $Pa \alpha = .73$, $Pt \alpha = .84$, $Sc \alpha = .82$, $Ma \alpha = .71$ y $Si \alpha = .81$; Rodríguez-Aguirre et al. (2023) informan para una escala EDO $\alpha = .86$; Capilla Ramírez et al. (2013) reportan $FIQ \alpha = .81$; y Martínez et al. (2017) reportan $\alpha > .70$ en las escalas principales de su muestra ($n = 281$). La estabilidad temporal (test-retest) aparece con menor frecuencia: Capilla Ramírez et al. (2013) informan coeficientes de correlación intraclase

(ICC) $\approx .81-.96$ para determinadas escalas agregadas (p. ej., EID, RCd, DISC-r); otros estudios no replican sistemáticamente procedimientos test–retest, por lo que la evidencia longitudinal es limitada. Asimismo, se reportan medidas diagnósticas puntuales (AUC, sensibilidad, especificidad) en estudios clínicos y forenses (Jiménez-Gómez et al., 2013; Sánchez et al., 2017).

Comparando entre estudios, se observa coherencia en que las escalas clínicas y de validez del MMPI-2-RF muestran en general consistencia interna aceptable ($\alpha \geq .7$ en la mayoría de las escalas principales). Sin embargo, la magnitud de los coeficientes y la presencia de ítems problemáticos varían por muestra y contexto (p. ej., muestras forenses vs. clínicas vs. poblacionales), lo cual Redondo et al. (2019) identifican como un efecto de moderadores. Entre las limitaciones recurrentes se encuentran la ausencia de reportes cuantitativos de validez de contenido, la falta de pruebas de invarianza multigrupo y la cobertura desigual de test–retest; por el contrario, entre las fortalezas se destacan análisis factoriales consistentes en adaptaciones y la disponibilidad de α por subescala en la mayoría de los estudios (Afonso et al., 2024; Capilla Ramírez et al., 2013; Jiménez-Gómez et al., 2013; Martínez et al., 2017; Redondo et al., 2019; Rodríguez-Aguirre et al., 2023; Sánchez et al., 2017).

Finalmente, la evaluación según criterios COSMIN muestra calidad metodológica variable: Redondo et al. (2019) y Capilla Ramírez et al. (2013) alcanzan valoraciones moderadas-altas en dimensiones como validez estructural y consistencia interna debido a su empleo de meta-análisis, *EFA/CFA* y reportes de ICC ; otros estudios (p. ej., investigaciones con muestras pequeñas o muestreos por conveniencia) reciben calificaciones más bajas en aspectos relacionados con diseño y tamaño muestral (Afonso et al., 2024; Rodríguez-Aguirre et al., 2023). En consecuencia, se recomienda para futuras validaciones documentar formalmente la validez de contenido mediante juicios de expertos con métricas ($I-CVI/S-CVI$), realizar análisis de invarianza factorial por grupos relevantes y planear estudios test–retest con reporte explícito de ICC por subescala.

Tabla 7

Características técnicas del MMPI-2 RF

Sección	Descripción
Datos generales	Nombre completo: Inventario Multifásico de Personalidad de Minnesota-2 Reestructurado (MMPI-2-RF). Autores: Yossef S. Ben-Porath y Auke Tellegen (2008). Año: 2008 (reestructuración del MMPI-2 de 1989). Constructo: Evaluación amplia de personalidad y psicopatología con reducción de redundancias entre escalas y optimización de la precisión diagnóstica.

	Número de ítems: 338. Formato de respuesta: Dicotómico (Verdadero/Falso). Adaptaciones culturales: Versiones validadas en español (España y Latinoamérica) y portugués, con estudios de equivalencia lingüística.
Estructura	El MMPI-2-RF conserva la lógica jerárquica del MMPI-2, reorganizada en: Escalas de validez revisadas: VRIN-r, TRIN-r, F-r, L-r, K-r, RBS. Escalas de orden superior: EID (Malestar emocional), THD (Pensamiento disfuncional), BXD (Problemas de conducta externa). Escalas clínicas reestructuradas (RC): RCd, RC1–RC9. Escalas de problemas específicos: dominios somáticos, internalizantes, externalizantes e interpersonales. Escalas de intereses: indicadores de orientación conductual y motivacional.
Descripción de los estudios seleccionados	Se identificaron siete estudios iberoamericanos sobre el MMPI-2-RF publicados entre 2013 y 2024, realizados en España, México y Portugal. En España, cinco investigaciones (Capilla Ramírez et al., 2013; Jiménez-Gómez et al., 2013; Martínez et al., 2017; Redondo et al., 2019; Sánchez et al., 2017) analizaron su utilidad clínica, forense y predictiva en poblaciones adultas. En México, Rodríguez-Aguirre et al. (2023) aplicaron el instrumento en médicos residentes. En Portugal, Afonso et al. (2024) desarrollaron la adaptación lingüística y el análisis de equivalencia factorial. Las muestras fueron principalmente de adultos clínicos, comunitarios y forenses, con tamaños que oscilaron entre varios cientos y varios miles de participantes según el contexto.
Evidencias de validez	Los estudios revisados mostraron consistencia estructural con el modelo original de Ben-Porath y Tellegen (2008). Redondo et al. (2019) replicaron la estructura factorial de las escalas reestructuradas (RC) con cargas entre .50 y .78, y correlaciones interescales esperadas ($r = .44-.68$), respaldando la validez convergente. Capilla Ramírez et al. (2013) confirmaron la validez criterial de F-r y RBS para detectar simulación ($p < .01$) frente a grupos clínicos genuinos. Jiménez-Gómez et al. (2013) reportaron ausencia de una dimensión de discapacidad social, lo que respalda la independencia factorial postulada. Sánchez et al. (2017) evidenciaron la eficacia de F-r, Fp-r y RBS en la detección de sobresimulación ($AUC > .8$), replicando patrones de las muestras normativas. Martínez et al. (2017) encontraron evidencia predictiva: puntuaciones elevadas en RCd y BXD predijeron menor adherencia al tratamiento ($\beta = -.31, p < .05$). Rodríguez-Aguirre et al. (2023) observaron correlaciones positivas entre agotamiento emocional y escalas de malestar ($r = .49-.61$), apoyando la validez convergente en población médica. Afonso et al. (2024) confirmaron la equivalencia factorial con la versión estadounidense ($CFI = .93$, $RMSEA = .05$), demostrando validez estructural intercultural.
Evidencias de confiabilidad	Los coeficientes de consistencia interna fueron adecuados, con valores entre $\alpha = .7$ y .93. En España (Capilla Ramírez et al., 2013; Martínez et al., 2017; Sánchez et al., 2017), las escalas clínicas y de validez mostraron α entre .79 y .9. En México, Rodríguez-Aguirre et al. (2023) reportaron $\alpha = .86$ para agotamiento emocional. En Portugal, Afonso et al. (2024) informaron $\alpha = .78-.91$, indicando estabilidad interna. En conjunto, la fiabilidad del MMPI-2-RF en contextos clínicos, forenses y de investigación es comparable a la reportada por Ben-Porath y Tellegen (2008).
Hallazgos según criterios COSMIN	Los estudios revisados cumplen criterios satisfactorios en validez estructural, consistencia interna y validez de criterio. Se evidencia replicación de la estructura jerárquica, consistencia entre muestras culturales y aplicación en múltiples contextos. No se identificaron reportes de confiabilidad test–retest ni de consistencia interevaluador, lo que limita parcialmente la interpretación longitudinal del instrumento.

Instrumento 3. Personality Assessment Inventory (PAI)

El *Personality Assessment Inventory* (PAI), desarrollado por Leslie C. Morey en 1991, se ha consolidado como un instrumento psicométrico ampliamente utilizado para la evaluación de la personalidad y la psicopatología en adultos. Consta de 344 ítems tipo Likert de cuatro puntos, diseñados para medir patrones clínicos, rasgos de personalidad, dificultades en el tratamiento y características interpersonales, permitiendo una comprensión integral del funcionamiento psicológico. A diferencia de inventarios clásicos como el MMPI-2, el PAI fue construido desde sus inicios para alinearse con modelos diagnósticos contemporáneos y facilitar la interpretación clínica directa de las puntuaciones. Su versión abreviada, el PAI-R (165 ítems), amplía su utilidad en contextos donde se requiere una evaluación más rápida sin comprometer la precisión (Ortiz-Tallo et al., 2017; Penado Abilleira & Rodicio-García, 2019).

El instrumento está conformado por 22 escalas principales distribuidas en cuatro áreas: validez, clínica, tratamiento e interpersonal (Paulino et al., 2024; Stover et al., 2016). Las escalas de validez (ICN, INF, NIM y PIM) evalúan la sinceridad y consistencia de las respuestas; las escalas clínicas abarcan dimensiones como ansiedad, depresión, paranoia, esquizofrenia, consumo de sustancias o rasgos límite y antisociales; las escalas relacionadas con el tratamiento permiten identificar factores relevantes para la intervención (agresión, ideación suicida, estrés, apoyo social y rechazo terapéutico); y las escalas interpersonales miden dominancia y afabilidad, útiles para analizar el estilo relacional del sujeto. Esta estructura convierte al PAI en una herramienta adecuada tanto para el diagnóstico como para la planificación terapéutica y la predicción del ajuste personal y social.

En Iberoamérica, los estudios de adaptación han confirmado la solidez psicométrica del instrumento. En Chile, Ortiz-Tallo et al. (2017) evaluaron una muestra de $n = 691$ adultos (564 de población general y 127 de contexto clínico), encontrando coeficientes alfa de Cronbach elevados en todas las escalas y patrones diferenciados entre ambos grupos. Estos resultados muestran la sensibilidad del PAI para discriminar perfiles clínicos y no clínicos y su utilidad en entornos sanitarios y comunitarios. En España, Penado Abilleira y Rodicio-García (2019) aplicaron el PAI-R en una muestra penitenciaria de $n = 294$, obteniendo una fiabilidad global de $\alpha = .95$ y un modelo factorial confirmatorio con índices adecuados (RMSEA = .07; CFI = .89; GFI = .90; $p < .01$). Estos resultados respaldan la validez estructural del instrumento y su capacidad para evaluar rasgos desadaptativos y tendencias antisociales en contextos forenses.

En Argentina, Stover et al. (2016) analizaron la dimensionalidad del PAI en $n = 374$ estudiantes universitarios, reportando valores de consistencia interna entre $\alpha = .82$ y $\alpha = .94$, confirmando una estructura de cinco dimensiones congruente con la propuesta original. Asimismo, identificaron correlaciones esperadas entre escalas clínicas e interpersonales, lo cual respalda su validez convergente. En Portugal, Paulino et al. (2024) aportaron datos normativos recientes en una muestra de $n = 900$ adultos, con coeficientes de α promedio entre $.81$ y $.86$, y confirmaron la invariancia factorial en función del sexo, la edad y el nivel educativo. Este hallazgo es especialmente relevante porque demuestra la estabilidad transcultural del PAI sin necesidad de modificaciones conceptuales significativas.

En conjunto, los hallazgos revisados demuestran que el PAI y su versión abreviada mantienen propiedades psicométricas consistentes en diferentes países y poblaciones. Las evidencias de validez de constructo se sustentan en los análisis factoriales confirmatorios realizados en España, Argentina y Portugal, los cuales confirman la replicabilidad de la estructura original. En cuanto a la validez convergente, las correlaciones reportadas entre escalas clínicas e indicadores sintomáticos o de personalidad apoyan la coherencia teórica del instrumento (Paulino et al., 2024; Stover et al., 2016). En el plano de la confiabilidad, los estudios muestran coeficientes α entre $.8$ y $.95$, lo que evidencia una excelente consistencia interna (Ortiz-Tallo et al., 2017; Penado Abilleira & Rodicio-García, 2019). Aunque todavía no existen estudios que evalúen la estabilidad temporal mediante test–retest, la replicación de resultados en distintos países sugiere una solidez transcultural estable.

Desde una perspectiva aplicada, el PAI ha demostrado utilidad en diversos ámbitos. En el área clínica contribuye al diagnóstico diferencial de trastornos emocionales y de personalidad, así como a la identificación de ideación suicida y a la planificación terapéutica. En el ámbito forense permite evaluar el riesgo de agresividad, la responsabilidad penal y la simulación de síntomas (Penado Abilleira & Rodicio-García, 2019). En contextos educativos y de orientación psicológica se utiliza para evaluar el ajuste emocional y social (Stover et al., 2016). Estas aplicaciones refuerzan su posición como un instrumento de referencia en la evaluación psicológica integral.

De acuerdo con los criterios COSMIN, la calidad metodológica de los estudios revisados se clasifica como alta. Todos los trabajos analizados presentan diseños descriptivos y analíticos con muestras amplias, análisis factorial confirmatorio y medidas robustas de confiabilidad

(Ortiz-Tallo et al., 2017; Penado Abilleira & Rodicio-García, 2019; Paulino et al., 2024; Stover et al., 2016). La evidencia disponible demuestra coherencia entre países, lo cual respalda la estabilidad transcultural del instrumento. No obstante, es recomendable que futuras investigaciones exploren la estabilidad temporal y la validez de criterio externo. En conjunto, el PAI presenta un perfil psicométrico de alta calidad y consistente con los estándares internacionales, consolidándose como un instrumento esencial para la evaluación de la personalidad y la psicopatología en el contexto iberoamericano.

Tabla 8*Características técnicas del PAI*

Sección	Descripción
Datos generales	Nombre: <i>Personality Assessment Inventory</i> (PAI). Autor: Leslie C. Morey. Año: 1991. Constructo: Evaluación de la personalidad y psicopatología adulta, con énfasis en dimensiones clínicas, de tratamiento, validez e interpersonales. Número de ítems: 344 (versión abreviada PAI-R: 165). Formato de respuesta: Escala Likert de 4 puntos (0 = Falso a 3 = Muy verdadero). Adaptaciones culturales: Estudios psicométricos en Chile, España, Argentina y Portugal.
Estructura	Compuesto por 22 escalas principales: • Validez: Inconsistencia (ICN), Infrecuencia (INF), Impresión Negativa (NIM), Impresión Positiva (PIM). • Clínicas: Quejas somáticas, Ansiedad, Trastornos de ansiedad, Depresión, Manía, Paranoia, Esquizofrenia, Rasgos límite, Rasgos antisociales, Problemas con alcohol y drogas. • Relacionadas con el tratamiento: Agresión, Ideación suicida, Estrés, Falta de apoyo social, Rechazo al tratamiento. • Interpersonales: Dominancia y Afabilidad (o Calidez).
Descripción de los estudios seleccionados	Se analizaron cuatro investigaciones iberoamericanas: Ortiz-Tallo et al. (2017, Chile), Penado Abilleira & Rodicio-García (2019, España), Stover et al. (2016, Argentina) y Paulino et al. (2024, Portugal). Las muestras incluyeron adultos de población general, clínica, penitenciaria y universitaria (n entre 294 y 900). Se evaluaron la consistencia interna, la estructura factorial y la validez convergente del PAI y PAI-R.
Evidencias de validez	En Chile, Ortiz-Tallo et al. (2017) confirmaron la adecuada fiabilidad del PAI y la diferenciación significativa entre población general y clínica. En España, Penado Abilleira & Rodicio-García (2019) reportaron validez factorial confirmatoria con buenos índices de ajuste (RMSEA = .07; CFI = .9; GFI = .9). En Argentina, Stover et al. (2016) identificaron una estructura de cinco dimensiones congruente con la original. En Portugal, Paulino et al. (2024) confirmaron la validez estructural y la invariancia entre grupos sociodemográficos.

Evidencias de confiabilidad	Los coeficientes α se situaron entre .8 y .95 según el estudio y la escala. Ortiz-Tallo et al. (2017): consistencia interna alta en todas las escalas. Penado Abilleira & Rodicio-García (2019): α global = .95 en PAI-R. Stover et al. (2016): α entre .82 y .94. Paulino et al. (2024): α promedio .81–.86.
Hallazgos según criterios COSMIN	Evidencia de alta calidad, con consistencia interna excelente y validez de constructo confirmada en diferentes contextos. Los estudios presentan muestras amplias y metodologías rigurosas. Se recomienda mantener la revisión transcultural continua y replicar análisis factoriales confirmatorios en nuevos grupos poblacionales.

Instrumento 4. Symptom Checklist-90-Revised (SCL-90-R)

El *Symptom Checklist-90-Revised* (SCL-90-R), elaborado por Leonard R. Derogatis en 1977 y revisado en 1994, constituye uno de los instrumentos más empleados para la evaluación del malestar psicológico y la psicopatología general. Diseñado originalmente como herramienta de tamizaje clínico, el SCL-90-R permite valorar la presencia e intensidad de síntomas en nueve dimensiones: somatización, obsesión-compulsión, sensibilidad interpersonal, depresión, ansiedad, hostilidad, ansiedad fóbica, ideación paranoide y psicoticismo (Sánchez & Ledesma, 2009). Su formato Likert de cinco puntos y su estructura factorial equilibrada facilitan una aplicación breve y sensible, adecuada para contextos clínicos, comunitarios y de investigación.

Desde su desarrollo, el instrumento ha mostrado una notable capacidad de adaptación transcultural. En España, Sandín et al. (2008) validaron la versión abreviada SA-45, confirmando la estabilidad del modelo factorial original y obteniendo coeficientes alfa de Cronbach (α) entre .79 y .92, lo que evidenció que la reducción de ítems no comprometía la fiabilidad del instrumento. En Argentina, Sánchez y Ledesma (2009) verificaron la estructura de nueve factores mediante análisis factorial confirmatorio en una muestra clínica de $n = 570$ adultos, con coeficientes α entre .72 y .86 y un α global = .96. Su estudio aportó evidencia de validez convergente al encontrar correlaciones significativas con medidas consolidadas de ansiedad y depresión, reforzando la precisión diagnóstica del SCL-90-R en población clínica argentina.

En Chile, Ignatyev et al. (2016) examinaron las propiedades psicométricas del SCL-90-R en $n = 427$ internos penitenciarios, obteniendo un α global = .97 y valores de α entre .76 y .89 por subescala. El análisis factorial mostró una estructura unidimensional robusta que explicaba el 70.7% de la varianza total, lo que sugiere la presencia de un factor general de malestar psicológico. Además, establecieron un punto de corte de 1.42 en el Índice Global de Severidad (*GSI*) como criterio óptimo para discriminar a individuos con trastornos mentales graves.

En Colombia, Londoño et al. (2018) evaluaron las propiedades estructurales y psicométricas del instrumento en $n = 214$ pacientes clínicos mediante un diseño longitudinal. Hallaron coeficientes de confiabilidad α entre .74 y .9 y una estructura factorial exploratoria de siete factores que explicaba el 60.7% de la varianza total. Asimismo, reportaron sensibilidad al cambio y diferencias significativas por género, con mayor sintomatología en mujeres, lo que coincide con la literatura sobre prevalencia diferencial de síntomas afectivos. En México, González-Forteza et al. (2008) aplicaron el SCL-90-R en una muestra de jóvenes universitarios y reportaron un α global = .96, además de correlaciones adecuadas con el Inventario de Ansiedad de Beck, lo que respalda su validez convergente y utilidad en población no clínica.

En términos generales, las adaptaciones iberoamericanas del SCL-90-R demuestran altos niveles de confiabilidad ($\alpha = .74-.97$) y validez, tanto de constructo como convergente. Los análisis factoriales confirman la estabilidad del modelo original, aunque algunos estudios identifican reducciones en la cantidad de factores (por ejemplo, siete dimensiones en la versión colombiana) sin alterar el sentido teórico del instrumento (Londoño et al., 2018). Asimismo, la sensibilidad al cambio y la capacidad discriminativa refuerzan su valor como herramienta de evaluación y monitoreo terapéutico.

En conjunto, los hallazgos revisados evidencian que el SCL-90-R y su versión abreviada SA-45 son instrumentos psicométricamente sólidos y culturalmente versátiles. Su uso extendido en contextos clínicos, comunitarios y forenses ha permitido confirmar su fiabilidad y estabilidad transcultural, consolidándolo como un referente esencial en la evaluación del malestar psicológico y la psicopatología general en Iberoamérica (González-Forteza et al., 2008; Ignatyev et al., 2016; Londoño et al., 2018; Sánchez & Ledesma, 2009; Sandín et al., 2008).

Tabla 9

Características técnicas del SCL-90-R

Sección	Descripción
Datos generales	Nombre completo: <i>Symptom Checklist-90-Revised</i> (SCL-90-R). Autor: Leonard R. Derogatis. Año: 1977 (revisión 1994). Constructo: Evaluación global del malestar psicológico en nueve dimensiones sintomáticas. Número de ítems: 90 (SA-45 = 45). Formato: Likert de 5 puntos (0 = Nada, 4 = Extremadamente). Adaptaciones: Validaciones en España, Argentina, Chile, Colombia y México.

Estructura	9 dimensiones: Somatización, Obsesión-Compulsión, Sensibilidad interpersonal, Depresión, Ansiedad, Hostilidad, Ansiedad fóbica, Ideación paranoide y Psicoticismo. Además, genera tres índices globales: GSI (Índice Global de Severidad), PSDI (Índice de Distrés de Síntomas Positivos) y PST (Número Total de Síntomas Positivos). La versión abreviada SA-45 mantiene las 9 dimensiones con alta consistencia interna (Sandín et al., 2008).
Descripción de los estudios seleccionados	España: Sandín et al. (2008) — SA-45, $n = 312$, alfas .79–.92. Argentina: Sánchez y Ledesma (2009) — $n = 570$, α .72–.86, α global .96. Chile: Ignatyev et al. (2016) — $n = 427$, α .76–.89, estructura unidimensional (70.7% varianza). Colombia: Londoño et al. (2018) — $n = 214$, α .74–.9, 7 factores (60.7% varianza). México: González-Forteza et al. (2008) — α global .96, validez convergente con BDI.
Evidencias de validez	Validez de constructo confirmada en España, Argentina y Chile. Validez convergente reportada en Argentina y México (correlaciones con BDI, STAI y medidas de afecto negativo). Estructuras factoriales robustas de 9 factores, con variaciones culturales en Colombia (7 factores).
Evidencias de confiabilidad	α global entre .9 y .97; subescalas entre .74 y .92. Alfas en SA-45 = .79–.92 (Sandín et al., 2008); Argentina = .72–.86 (Sánchez & Ledesma, 2009); Chile = .76–.89 (Ignatyev et al., 2016); Colombia = .74–.9 (Londoño et al., 2018).
Hallazgos según criterios COSMIN	Evidencia de alta calidad por la consistencia de resultados, muestras amplias y uso de análisis factorial. Fortalezas en validez convergente y consistencia interna; limitaciones en evaluación test-retest y comparaciones longitudinales entre países. Se recomienda continuar explorando la invarianza factorial y la sensibilidad clínica en distintas poblaciones.

Instrumento 5. Listado de Síntomas Breve (LSB-50)

El Listado de Síntomas Breve (LSB-50), desarrollado por Manuel R. Abuín y Luis de Rivera en 2014, surge como una alternativa moderna, concisa y psicométricamente refinada al clásico SCL-90-R. Su propósito es evaluar de manera ágil y precisa la presencia e intensidad de síntomas psicológicos y psicosomáticos en población adulta, permitiendo estimar el nivel de malestar emocional y el riesgo psicopatológico general. La brevedad de sus 50 ítems, junto con una estructura factorial optimizada, lo convierte en un instrumento de amplio alcance clínico y comunitario, especialmente útil en contextos donde se requiere un tamizaje rápido y fiable del estado emocional del individuo (Abuín & de Rivera, 2014).

El LSB-50 se compone de ocho escalas clínicas: Psicorreactividad (con subescalas de Hipersensibilidad y Obsesión-compulsión), Ansiedad, Hostilidad, Somatización, Depresión, Sueño estricto, Sueño ampliado y Riesgo psicopatológico; además, incluye dos escalas de

validez (Minimización y Magnificación), que permiten identificar la tendencia a subestimar o exagerar los síntomas. También presenta tres indicadores globales: Frecuencia de síntomas (NUM), Intensidad media (INT) y Afectación global psicopatológica (GLOB), que ofrecen una visión integral del nivel de perturbación psicológica. Esta arquitectura permite obtener un perfil sintomatológico completo con un uso eficiente del tiempo y los recursos, manteniendo altos estándares de precisión y control de sesgos de respuesta (Abuín & de Rivera, 2014).

El estudio original español, realizado con una muestra total de $n = 1242$ adultos (790 de población general y 452 pacientes clínicos), reportó coeficientes alfa de Cronbach (α) entre .79 y .90 en las escalas, así como una estructura factorial de seis dimensiones que explicaba el 55.3% de la varianza. Se observaron correlaciones significativas con variables de dependencia afectiva y vinculación interpersonal, lo que confirmó una sólida validez concurrente. Los autores destacan que el LSB-50 mejora instrumentos previos como el SCL-90-R al reorganizar los síntomas en dimensiones clínicamente más coherentes, por ejemplo, la incorporación de síntomas de sueño y una escala de riesgo psicopatológico y al incluir controles de sesgo de respuesta que aumentan la validez ecológica de la evaluación. (Abuín & de Rivera, 2014).

En Argentina, de la Iglesia et al. (2015) replicaron y ampliaron estos hallazgos en una muestra de $n = 541$ adultos de población general y clínica, obteniendo una fiabilidad global de $\alpha = .95$ y coeficientes α entre .78 y .91 en las subescalas. Los análisis factoriales confirmatorios respaldaron la estructura de ocho dimensiones y mostraron correlaciones consistentes con medidas de ansiedad, depresión y estrés percibido, reafirmando la validez de constructo y la validez convergente del instrumento en el contexto latinoamericano.

Investigaciones posteriores en Chile y Colombia han replicado estos indicadores psicométricos, reportando valores de α superiores a .8 y estructuras factoriales estables. En ambos países, el LSB-50 mostró utilidad tanto en población comunitaria como clínica, diferenciando de manera eficaz entre individuos con y sin alteraciones emocionales, y evidenciando alta sensibilidad para monitorear cambios terapéuticos.

En conjunto, los hallazgos iberoamericanos confirman que el LSB-50 posee propiedades psicométricas robustas y consistentes, con niveles de confiabilidad entre $\alpha = .79$ y $\alpha = .95$, validez de constructo confirmada y excelente capacidad discriminante. Su estructura factorial mantiene coherencia con modelos dimensionales contemporáneos de la psicopatología y su brevedad no compromete la fiabilidad.

El LSB-50 se consolida como una herramienta breve, precisa y culturalmente adaptable para evaluar el malestar psicológico y los síntomas psicosomáticos. Desarrollado a partir del SCL-90-R, incorpora mejoras en la organización factorial y en el control de sesgos de respuesta, convirtiéndose en una alternativa moderna y eficaz para la práctica clínica, la investigación y la evaluación psicológica en salud mental (Abuín & de Rivera, 2014; de la Iglesia et al., 2015).

Tabla 10

Características técnicas del LSB-50

Sección	Descripción
Datos generales	Nombre: <i>Listado de Síntomas Breve</i> (LSB-50). Autores: Manuel R. Abuín y Luis de Rivera. Año: 2014. Constructo: Evaluación psicopatológica breve de síntomas psicológicos y psicosomáticos. Número de ítems: 50. Formato: Likert de 5 puntos (0 = Nada, 4 = Muchísimo). Adaptaciones: España, Argentina, Chile y Colombia.
Estructura	Ocho escalas clínicas: Psicorreactividad (Hipersensibilidad y Obsesión-compulsión), Ansiedad, Hostilidad, Somatización, Depresión, Sueño estricto, Sueño ampliado y Riesgo psicopatológico; dos escalas de validez (Minimización y Magnificación). Tres indicadores globales: NUM (frecuencia de síntomas), INT (intensidad media) y GLOB (afectación global).
Estudios iberoamericanos	España: Abuín & de Rivera (2014) — $n = 1\,242$ (790 población general, 452 clínica), $\alpha .79-.9$, seis factores (55.3% varianza). Argentina: de la Iglesia et al. (2015) — $n = 541$, $\alpha .78-.91$, α global .95, ocho factores. Chile y Colombia: $\alpha > .8$, estructura estable y validez discriminante alta.
Evidencias de validez	Validez de constructo: confirmada mediante análisis factorial exploratorio y confirmatorio (6–8 dimensiones). Validez concurrente: correlaciones significativas con ansiedad, depresión, estrés y dependencia afectiva. Validez discriminante: diferencia clara entre poblaciones clínicas y no clínicas.
Evidencias de confiabilidad	α entre .79 y .95 según escala y muestra; España = .79–.9; Argentina = .78–.91; Latinoamérica = $> .8$. Alta homogeneidad interna en todas las dimensiones.
Hallazgos según criterios COSMIN	Calidad metodológica moderada-alta. Diseños con muestras amplias, análisis factorial y fiabilidad elevada. Limitaciones: falta de estudios de estabilidad temporal y comparaciones transculturales amplias.

Instrumento 6. Structured Inventory of Malingered Symptomatology (SIMS)

El *Structured Inventory of Malingered Symptomatology* (SIMS) es un instrumento de autoinforme diseñado para detectar la simulación de síntomas psicológicos y neurocognitivos en contextos clínicos, forenses y de investigación. De esta forma, fue desarrollado originalmente por Widows y Smith (2005) como una medida de detección de respuestas exageradas o fingidas

mediante 75 ítems distribuidos en cinco subescalas: trastornos psicóticos, trastornos neurológicos, trastornos amnésicos, síntomas afectivos y síntomas de baja inteligencia. Asimismo, la escala de respuesta es dicotómica (“verdadero” o “falso”), lo que facilita su aplicación en diversos contextos. De esta manera, en la literatura iberoamericana, el SIMS ha sido adaptado principalmente en poblaciones clínicas, penitenciarias y universitarias, y se ha utilizado para la evaluación de validez de síntomas y la detección de simulación en procesos de peritaje psicológico y evaluación neuropsicológica (Puente-López et al., 2022).

En cuanto a la evidencia de validez de contenido, los estudios revisados señalan que los ítems del SIMS corresponden adecuadamente al constructo que se pretende evaluar; sin embargo, esta evidencia no constituye validez de estructura factorial. Respecto a la validez estructural, en la validación española, Puente-López et al. (2022) confirmaron una estructura de cinco factores coherente con las dimensiones originales mediante análisis factorial confirmatorio ($\chi^2/\text{gl} = 2.41$, $CFI = .93$, $RMSEA = .06$), lo que respalda la validez estructural del instrumento en población penitenciaria. De manera similar, Blasco-Saiz et al. (2013) corroboraron la consistencia del modelo en población clínica y forense, observando cargas factoriales elevadas en todos los ítems ($\lambda > .50$) y adecuada discriminación entre escalas. En contraste, Marín-Torices et al. (2017) identificaron algunas diferencias en la agrupación factorial de ítems en una muestra universitaria, aunque los índices globales de ajuste permanecieron dentro de rangos aceptables ($CFI = .91$, $RMSEA = .06$). Estas variaciones sugieren que, aunque la estructura del SIMS es estable, ciertos ítems podrían verse influidos por características contextuales o del perfil de los participantes.

Asimismo, respecto a la evidencia de validez basada en la relación con otras variables, los estudios iberoamericanos muestran asociaciones significativas entre las puntuaciones del SIMS y otras medidas de detección de simulación. Blasco-Saiz et al. (2013) reportaron correlaciones moderadas y positivas entre las subescalas del SIMS y el Test de Inconsistencia de Respuesta ($r = .46$, $p < .01$), lo que apoya la validez convergente del instrumento. Asimismo, Puente-López et al. (2022) hallaron diferencias estadísticamente significativas en las puntuaciones del SIMS entre grupos con y sin incentivos externos ($F(1, 198) = 27.53$, $p < .001$, $\eta^2 = .12$), lo que refuerza la validez discriminante del instrumento para detectar simulación en contextos penitenciarios.

Seguido a esto, en cuanto a la confiabilidad, la consistencia interna del SIMS ha mostrado valores adecuados en los distintos estudios revisados. Puente-López et al. (2022) reportaron coeficientes α de Cronbach entre .72 y .86 para las subescalas, mientras que Marín-Torices et al.

(2017) informaron valores similares ($\alpha = .75-.88$) en una muestra universitaria. De esta manera, en población clínica y forense, Blasco-Saiz et al. (2013) encontraron niveles de consistencia interna ligeramente superiores ($\alpha = .80-.89$), con mayor estabilidad en las subescalas de síntomas afectivos y neurológicos. Sin embargo, no se identificaron reportes de confiabilidad temporal (test-retest), lo que indica una limitación en la evidencia de estabilidad longitudinal del instrumento en la literatura iberoamericana disponible.

Por último, de acuerdo con la evaluación mediante el sistema COSMIN, la mayoría de los estudios revisados obtuvieron valoraciones de calidad metodológica moderada a alta en los apartados de validez estructural y consistencia interna. Sin embargo, los ítems relacionados con la validez de criterio y la confiabilidad temporal fueron valorados como insuficientemente reportados, lo que sugiere la necesidad de estudios adicionales que profundicen en la estabilidad del SIMS en distintas aplicaciones clínicas y poblacionales (Blasco-Saiz et al., 2023; Marín-Torices et al., 2021; Puente-López et al., 2021).

Tabla 11

SIMS

Sección	Descripción
Datos generales	Nombre completo: Structured Inventory of Malingered Symptomatology (SIMS). Autores: Glenn P. Widows y James N. Smith (2005). Año: 2005. Constructo: Evaluación de la simulación y exageración de síntomas psicopatológicos en contextos clínicos, forenses y de investigación. Número de ítems: 75. Formato de respuesta: dicotómico (Verdadero/Falso). Adaptaciones culturales: estudios psicométricos y de adaptación en España y Latinoamérica, enfocados en la estructura factorial, validez discriminante y confiabilidad del instrumento.
Estructura	El SIMS evalúa cinco dominios de síntomas comúnmente simulados: AF (Amnesia), LI (Baja inteligencia), PS (Síntomas psicóticos), NE (Síntomas neurológicos) y AFS (Síntomas afectivos). Cada subescala incluye 15 ítems, con respuestas Verdadero/Falso; las puntuaciones altas indican mayor probabilidad de simulación. Se obtiene un puntaje total y puntuaciones por subescala.
Descripción de los estudios seleccionados	Se identificaron seis estudios iberoamericanos entre 2013 y 2025, desarrollados principalmente en España, con muestras clínicas, forenses, víctimas de violencia y pacientes con daño físico o emocional. Blasco-Saiz y Pallardó-Durá (2013) realizaron un estudio preliminar sobre detección de exageración de síntomas mediante el SIMS y el MMPI-2-RF en pacientes con trastornos adaptativos y ansioso-depresivos. Marín-Torices et al. (2017) validaron consecuencias neuropsicológicas en víctimas de violencia de pareja, aplicando pruebas de esfuerzo, entre ellas el SIMS. Martín-Albo et al. (2009) evaluaron la versión española de escalas motivacionales utilizando el SIMS como criterio concurrente. Puente-López et al. (2020) analizaron la precisión diagnóstica del SIMS en pacientes tras accidentes de tráfico. Puente-López et al. (2022) evaluaron el impacto del <i>coaching</i> en las respuestas al SIMS en población forense. Puente-López et al. (2025) realizaron un metaanálisis sobre la prevalencia de la sobreinformación sintomática en

	pacientes clínicos mediante el SIMS. Las muestras variaron entre $n = 86$ y $n = 748$, e incluyeron contextos clínicos, forenses y experimentales.
Evidencias de validez	Los estudios aportan evidencia estructural, convergente, discriminante y de criterio. Blasco-Saiz y Pallardó-Durá (2013) mostraron que el SIMS diferencia eficazmente entre simuladores y pacientes genuinos con trastornos adaptativos ($F(5, 177) = 8.34, p < .001, \eta^2 = .19$). Marín-Torices et al. (2017) reportaron validez convergente con otras pruebas de esfuerzo ($r = .42, p < .01$) en víctimas de violencia de pareja. Puente-López et al. (2020) hallaron alta sensibilidad (.88) y especificidad (.84) en la detección de simulación en pacientes con secuelas de accidentes, replicando la estructura de cinco factores original. Puente-López et al. (2022) observaron que el <i>coaching</i> reduce la eficacia discriminante del SIMS, lo que confirma su validez de criterio en contextos controlados ($AUC = .87, p < .001$). Puente-López et al. (2025) confirmaron mediante metaanálisis la estabilidad de los indicadores de sobreinformación y la replicación transnacional de la estructura del instrumento ($CFI = .94$; $RMSEA = .58$). En conjunto, los resultados evidencian validez estructural robusta y validez convergente moderada a alta, coherentes con los hallazgos del modelo original de Widows y Smith (2005).
Evidencias de confiabilidad	Los coeficientes de consistencia interna reportados fueron adecuados en todos los estudios, con valores de α entre .78 y .94. Blasco-Saiz y Pallardó-Durá (2013): α total = .88; subescalas $\alpha = .70-.85$. Marín-Torices et al. (2017): $\alpha = .81$; consistencia entre subescalas similar al modelo original. Puente-López et al. (2020): α total = .91; subescalas $\alpha = .73-.88$. Puente-López et al. (2022): α total = .90; estabilidad de los factores pese al <i>coaching</i> . Puente-López et al. (2025): estimaciones promedio $\alpha = .89$ en el metaanálisis ($IC95\% = .86-.92$). Ningún estudio reportó confiabilidad <i>test-retest</i> o consistencia interevaluador.
Hallazgos según criterios COSMIN	De acuerdo con los criterios COSMIN, los estudios alcanzan alta calidad metodológica en los dominios de validez estructural y consistencia interna, moderada en validez de criterio, y baja en estabilidad temporal. Se evidencia replicación de la estructura original de cinco factores, adecuada precisión diagnóstica y consistencia de resultados entre poblaciones forenses, clínicas y experimentales. Las principales limitaciones se relacionan con la ausencia de estudios longitudinales y de estabilidad temporal.

Instrumento 7. NEUROPSI

El instrumento NEUROPSI es una batería neuropsicológica desarrollada originalmente en español por Ostrosky-Solís y colaboradores, diseñada para evaluar el funcionamiento cognitivo general en personas de habla hispana y adaptada posteriormente a otras poblaciones latinoamericanas. De esta manera, su propósito es ofrecer una evaluación breve, pero abarcadora, de dominios como atención, memoria, lenguaje, lectura y escritura, habilidades visuoespaciales, funciones ejecutivas y motricidad. Como lo menciona Ostrosky-Solís et al. (2007), el instrumento fue concebido para proporcionar normas específicas por edad y nivel educativo, facilitando la interpretación clínica en contextos donde los baremos anglosajones no son adecuados. Así mismo, en su versión completa, el NEUROPSI incluye 33 tareas o subpruebas agrupadas en siete dominios principales, con un rango de puntuación total de 0 a 130 puntos. Por

ende, las respuestas se registran en formato dicotómico o de escala ordinal según la tarea, y la calificación total se obtiene a partir de la suma ponderada de los aciertos en cada subprueba. De esta forma, la batería está dirigida a adultos y personas mayores, y ha sido validada en muestras clínicas y no clínicas de países como México, Argentina y Brasil, con restricciones de aplicación ligadas al nivel educativo mínimo y a la integridad sensorial y motora de los participantes (Ostrosky-Solís et al., 2007; Querejeta et al., 2012).

En cuanto a la validez del instrumento, en la adaptación argentina, Querejeta et al. (2012) establecieron normas diferenciadas por edad y escolaridad en una muestra de $n = 336$ participantes entre 16 y 85 años, observando que el desempeño disminuye significativamente con la edad ($F = 4.78, p < .01$) y mejora con el nivel de instrucción ($F = 6.32, p < .01$). Estos hallazgos respaldan la validez del instrumento respecto a variables demográficas relevantes. En cuanto a su estructura interna, Ostrosky-Solís et al. (2007) reportaron correlaciones positivas y significativas entre los dominios que componen la batería, con cargas factoriales entre .54 y .81 en los factores de atención y memoria, lo que confirma una organización factorial coherente con el modelo teórico propuesto. A su vez, Abrisqueta-Gómez et al. (2008), en una muestra clínica de $n = 58$ pacientes con enfermedad de Alzheimer y $n = 60$ controles sanos, observaron que el puntaje total del NEUROPSI mostró una alta capacidad discriminativa entre grupos. El análisis de la curva ROC arrojó un área bajo la curva (AUC) de 0.99 para diferenciar controles de pacientes con Alzheimer leve (punto de corte = 88.3, sensibilidad = 96%, especificidad = 100%), y un AUC de 0.99 para distinguir entre Alzheimer leve y moderado (punto de corte = 63.5, sensibilidad = 100%, especificidad = 92%). Estos resultados evidencian una excelente validez de criterio y utilidad diagnóstica en el contexto clínico neurológico.

Así mismo, la evidencia de constructo también ha sido evaluada en términos de estabilidad intercultural. Desde el estudio de Ostrosky-Solís et al. (2007), compararon el desempeño en diferentes países hispanohablantes y no hallaron efectos significativos de la interacción país $\times$ nivel educativo ($F = 1.06, p = .37$), lo que sugiere invariancia estructural del instrumento en muestras latinoamericanas. Este resultado es consistente con los datos obtenidos por Querejeta et al. (2012) en Argentina, quienes reportaron un patrón factorial similar al original y confirmaron la sensibilidad del instrumento para detectar variaciones por edad y escolaridad sin comprometer su estructura teórica.

Con respecto a la confiabilidad, los estudios muestran resultados mixtos. Como señala Ostrosky-Solís et al. (2007), informaron coeficientes de consistencia interna (α de Cronbach) entre .81 y .89 para las subescalas de atención y memoria, lo que indica una fiabilidad interna alta. Sin embargo, Abrisqueta-Gómez et al. (2008) no reportaron valores de α para todas las subpruebas, limitando parcialmente la comparación entre versiones. De esta manera, los análisis de estabilidad temporal son escasos; no se hallaron estudios que aplicaran procedimientos de test–retest sistemáticos en las adaptaciones iberoamericanas revisadas, lo que representa una limitación en la evidencia de confiabilidad temporal. En cambio, Abrisqueta-Gómez et al. (2008) calcularon el error estándar de medición (SEM) para la puntuación total y las subpruebas, obteniendo $SEM = 3.1$ (error del 2.8%) para el total y SEM entre 2.2 y 3.6 (error del 3–4%) para las subpruebas, lo que se considera adecuado para propósitos diagnósticos individuales.

Por último, desde la evaluación metodológica basada en COSMIN, los estudios revisados presentan una calidad global buena, aunque con limitaciones en la documentación de algunos procedimientos psicométricos. Los trabajos de Ostrosky-Solís et al. (2007) y Abrisqueta-Gómez et al. (2008) cumplen con criterios adecuados de validez estructural y de criterio, mientras que las investigaciones de Querejeta et al. (2012) amplían la evidencia de validez convergente y discriminante en poblaciones clínicas y comunitarias. No obstante, la falta de reportes sistemáticos sobre validez de contenido, la ausencia de datos de test–retest y la escasa publicación de alfas por subescala en algunas versiones locales limitan la solidez global de la evidencia.

Tabla 12*NEUROPSI*

Sección	Descripción
Datos generales	Nombre completo: NEUROPSI Atención y Memoria. Autores: Feggy Ostrosky-Solís, Alfredo Ardila y Mónica Rosselli. Año: 1999. Constructo: Evaluación neuropsicológica breve de atención, memoria, lenguaje y funciones ejecutivas en población hispanohablante. Número de tareas: aproximadamente 33 subpruebas agrupadas en dominios cognitivos. Formato de respuesta: combinación de tareas de ejecución y reactivos de respuesta dicotómica u ordinal según el tipo de estímulo. Adaptaciones culturales: versiones validadas en México, Argentina y Brasil, con normas diferenciadas por edad y escolaridad y aplicaciones en contextos clínicos y comunitarios.
Estructura	El NEUROPSI evalúa siete dominios principales: orientación, atención y concentración, memoria, lenguaje, lectura y escritura, funciones visuoespaciales y funciones ejecutivas. Cada subprueba se califica mediante puntajes parciales que se suman para obtener un

	puntaje total con un máximo de 130. Los estudios iberoamericanos confirman que la estructura del instrumento se ajusta a los modelos teóricos de procesamiento cognitivo y que las puntuaciones se afectan por edad y escolaridad, por lo cual se incluyen normas específicas para estas variables (Ostrosky-Solís et al., 2007; Querejeta et al., 2012).
Descripción de los estudios seleccionados	Se identificaron cuatro estudios empíricos iberoamericanos entre 2005 y 2012, realizados en México, Brasil y Argentina. Abrisqueta-Gómez et al. (2008) estudiaron la aplicabilidad de la batería abreviada en pacientes con enfermedad de Alzheimer en Brasil. Ostrosky-Solís et al. (2007) desarrollaron la validación de la versión original mexicana con normas por edad y educación. Picasso, Ostrosky y Nicolini (2005) evaluaron la sensibilidad del instrumento para diferenciar subtipos de esquizofrenia. Finalmente, Querejeta et al. (2012) realizaron la adaptación argentina estableciendo normas nacionales. Las muestras incluyeron entre $n = 60$ y $n = 800$ participantes, con edades entre 16 y 85 años, provenientes de poblaciones clínicas (neurológicas y psiquiátricas) y comunitarias.
Evidencias de validez	Los estudios aportan evidencia de validez estructural, convergente y de criterio. Ostrosky-Solís et al. (2007) hallaron correlaciones significativas entre las subpruebas de atención y memoria ($r = .54-.81$), confirmando una estructura factorial coherente con el modelo original y sin efectos significativos en la interacción país $\times$ educación ($F = 1.06$, $p = .37$), lo que indica invariancia intercultural. Querejeta et al. (2012) replicaron la estructura factorial en Argentina con efectos de edad ($F = 4.78$, $p < .01$) y escolaridad ($F = 6.32$, $p < .01$), manteniendo la consistencia estructural. Abrisqueta-Gómez et al. (2008) demostraron validez convergente con pruebas neuropsicológicas clásicas: Stroop ($r = .68$, $p < .001$), TMT-A ($r = -.64$, $p < .001$), TMT-B ($r = -.59$, $p < .001$) y Digit Span de la WAIS ($r = .62$, $p < .001$). En validez de criterio, el NEUROPSI mostró alta capacidad discriminativa entre controles sanos y pacientes con Alzheimer leve (AUC = 0.99; punto de corte = 88.3; sensibilidad = 96%; especificidad = 100%) y entre Alzheimer leve y moderado (AUC = 0.99; punto de corte = 63.5; sensibilidad = 100%; especificidad = 92%). Picasso et al. (2005) evidenciaron sensibilidad del 90% y especificidad del 85% para diferenciar subtipos de esquizofrenia, ampliando la evidencia de validez de criterio en contextos psiquiátricos. Ningún estudio reportó análisis formales de validez de contenido mediante juicio de expertos.
Evidencias de confiabilidad	Ostrosky-Solís et al. (2007) reportaron coeficientes α de Cronbach entre .81 y .89 para las subescalas de atención y memoria, indicando alta consistencia interna. Abrisqueta-Gómez et al. (2008) informaron un error estándar de medición (<i>SEM</i>) de 3.10 (2.8% de error) para el puntaje total y <i>SEM</i> entre 2.2–3.6 (3–4% de error) para las subpruebas, reflejando una precisión diagnóstica adecuada. No se reportaron estudios de estabilidad temporal mediante test–retest ni de consistencia interevaluador.
Hallazgos según criterios COSMIN	Según los criterios COSMIN, los estudios muestran alta calidad metodológica en validez estructural, convergente y de criterio, y moderada en consistencia interna. Se observa replicación de la estructura original, adecuada capacidad discriminativa y coherencia en los resultados entre poblaciones clínicas y comunitarias. Las principales limitaciones se relacionan con la ausencia de estudios sobre validez de contenido y estabilidad temporal. En conjunto, el NEUROPSI presenta propiedades psicométricas sólidas y aplicabilidad clínica demostrada en distintos contextos latinoamericanos (Abrisqueta-Gómez et al., 2008; Ostrosky-Solís et al., 2007; Picasso et al., 2005; Querejeta et al., 2012).

Instrumento 8. Test de Empatía Cognitiva y Afectiva (TECA)

El Test de Empatía Cognitiva y Afectiva (TECA) fue desarrollado originalmente en España para población adulta y evalúa cuatro dimensiones: adopción de perspectivas (AP), comprensión emocional (CE), estrés empático (EE) y alegría empática (AE), mediante un formato tipo Likert de cinco puntos (López-Pérez et al., 2008). A partir de esta versión se han generado diferentes adaptaciones y desarrollos: una adaptación brasileña para universitarios con 31 ítems tras depuración de reactivos (Freitas & Miranda, 2021), una versión infantil y juvenil que, luego del proceso psicométrico, quedó en 30 ítems (López-Pérez et al., 2014) y una versión abreviada de 15 ítems para universitarios argentinos, orientada a la parsimonia y manteniendo las cuatro subescalas originales (Lemos et al., 2022). Los cinco estudios revisados se ubican en contextos educativos o universitarios y emplean autoinforme, aplicado en formato colectivo o individual, sin estudios en población forense (Chavira Trujillo & Celis de la Rosa, 2022; Freitas & Miranda, 2021; Gorostiaga et al., 2014; Lemos et al., 2022; López-Pérez et al., 2014).

En relación con la validez de contenido y la adaptación lingüística, en Brasil se implementó un proceso de traducción y retrotraducción acompañado de revisión por especialistas y ajustes semánticos para asegurar la equivalencia conceptual del instrumento. La muestra final incluyó $n = 256$ universitarios (71.10 % mujeres; $M = 21.58$; $DE = 4.88$), y la administración estandarizada duró aproximadamente 20 minutos (Freitas & Miranda, 2021). En la versión infantil-juvenil española, el instrumento fue revisado por tres expertos en psicología del desarrollo y se aplicó una prueba piloto con la población objetivo para eliminar o reformular ítems confusos, resultando en una versión final de 30 ítems (López-Pérez et al., 2014). En Argentina, la versión breve se elaboró mediante revisión de jueces expertos y un estudio piloto que garantizó coherencia entre los componentes cognitivos y afectivos (Lemos et al., 2022). En México, aunque se realizó una adaptación lingüística al español mexicano, los análisis evidenciaron problemas de equivalencia estructural frente a la muestra española, indicando sensibilidad cultural en algunos ítems (Chavira Trujillo & Celis de la Rosa, 2022). En conjunto, las adaptaciones de Argentina, Brasil y España muestran procedimientos de adaptación adecuados, mientras que el estudio mexicano destaca la necesidad de revisar la equivalencia transcultural.

En cuanto a la validez de constructo, todos los estudios aportan evidencia sobre la estructura interna del TECA. En Brasil se realizó un análisis factorial exploratorio con rotación

oblicua que confirmó el modelo teórico de cuatro factores con $KMO = .79$ y una prueba de esfericidad de Bartlett $\chi^2(528) = 1919.58, p < .001$; la solución explicó el 35.66 % de la varianza total ($AE = 16.9\%$; $EE = 7.97\%$; $CE = 5.58\%$; $AP = 5.21\%$) (Freitas & Miranda, 2021). En España, la versión TECA-NA mostró un análisis factorial exploratorio con cuatro factores que explicaron el 38.88 % de la varianza y un análisis factorial confirmatorio con índices adecuados: $\chi^2(334) = 811.8, p < .001$; $RMSEA = .05$; $RMR = .04$; $TLI = .95$; $CFI = .94$ (López-Pérez et al., 2014). En población escolar vasca, un análisis factorial confirmatorio con estimación robusta confirmó el modelo de cuatro factores tras eliminar dos ítems: $CFI = .95$; $IFI = .95$; $RMSEA = .04$ (Gorostiaga et al., 2014). En Argentina, el modelo de 15 ítems presentó el mejor ajuste ($\chi^2/gl = 2.14$; $CFI = .96$; $GFI = .97$; $AGFI = .95$; $NNFI = .95$; $IFI = .96$; $RMSEA = .04$; $AIC = 251$), siendo superior al modelo de dos factores y al original de 33 ítems (Lemos et al., 2022). En México, aunque los índices de ajuste fueron aceptables ($CFI \approx .93$; $RMSEA \approx .05$; $SRMR \approx .04$), el análisis multigrupo no mantuvo la invariancia métrica frente a la muestra española ($\Delta CFI > .01$), limitando comparaciones transculturales (Chavira Trujillo & Celis de la Rosa, 2022).

Respecto a la validez convergente, discriminante y de grupos conocidos, en Brasil se reportó confiabilidad compuesta adecuada por factor ($AP = .77$; $CE = .77$; $EE = .81$; $AE = .77$) y validez discriminante mediante el criterio $AVE > \varphi^2$ (Freitas & Miranda, 2021). En España, las correlaciones con el IECA fueron positivas y las correlaciones con el AQ fueron negativas alrededor de $-.25$ (Gorostiaga et al., 2014; López-Pérez et al., 2014). En Argentina, la versión abreviada mostró correlaciones positivas con la escala COOL ($AE-COOL r \approx .44$; $CE-COOL r \approx .38$; $EE-COOL r \approx .31$; $AP-COOL r \approx .27$; $p < .001$) (Lemos et al., 2022). En México, la validez transcultural se vio limitada debido a la falta de invariancia (Chavira Trujillo & Celis de la Rosa, 2022).

En cuanto a la confiabilidad, en Brasil se reportó $\alpha_{total} = .82$ y por subescala $AP = .71$; $CE = .67$; $EE = .76$; $AE = .68$ (Freitas & Miranda, 2021). En España, los coeficientes α oscilaron entre $.71$ y $.77$ (López-Pérez et al., 2014). En el País Vasco se obtuvieron $\omega_{global} = .89$ y $\omega_{subescalas}$ entre $.66$ y $.86$ (Gorostiaga et al., 2014). En Argentina, los coeficientes ω variaron entre $.55$ y $.82$ en el Estudio 1 y entre $.66$ y $.79$ en el Estudio 2, con fiabilidad compuesta global $= .90$ (Lemos et al., 2022). En México, $\alpha_{total} \approx .86$ junto con valores subescalares entre $.64$ y $.76$ (Chavira Trujillo & Celis de la Rosa, 2022). Ningún estudio reportó *test-retest*, por lo que la evidencia de estabilidad temporal es insuficiente.

Según los criterios COSMIN, la calidad de la evidencia es moderada en estructura interna, ya que la mayoría de las muestras confirman el modelo de cuatro factores mediante índices de ajuste adecuados, aunque persisten problemas de invariancia transcultural. La consistencia interna también es moderada, con variabilidad en las subescalas AP y CE. La validez convergente y discriminante es limitada a moderada debido a la ausencia de criterios conductuales o poblaciones clínicas. La fiabilidad temporal es insuficiente por la ausencia de estudios de estabilidad longitudinal, y el riesgo de sesgo metodológico es bajo a moderado debido al uso de muestras por conveniencia y a la falta de análisis de invariancia por sexo, edad o país.

Tabla 13*Test de Empatía Cognitiva y Afectiva (TECA)*

Sección	Descripción
Datos generales	Nombre: Test de Empatía Cognitiva y Afectiva (TECA) Autor original: López-Pérez, B., Fernández-Pinto, I., & Abad, F. (2008). <i>Test de Empatía Cognitiva y Afectiva (TECA)</i>. TEA Ediciones. Año: 2008 (versión original); adaptaciones publicadas entre 2014 y 2022 en diferentes países iberoamericanos. Constructo: Evaluación de la empatía desde una perspectiva multidimensional que distingue los componentes cognitivos (adopción de perspectiva y comprensión emocional) y afectivos (estrés y alegría empáticos). Número de ítems: Versión original: 33 ítems; adaptaciones entre 15 y 35 ítems según la población y país (Freitas & Miranda, 2021; Lemos et al., 2022; López-Pérez et al., 2014). Formato y aplicabilidad: Escala tipo Likert de cinco puntos (1 = nunca, 5 = siempre). Aplicación autoadministrada, en formato papel o digital, con una duración aproximada de 15 a 20 minutos. Dirigido a población general, universitaria e infantojuvenil, según la versión. Adaptación Cultural: Validado y adaptado en Brasil (Freitas & Miranda, 2021), México (Chavira Trujillo & Celis de la Rosa, 2022), Argentina (Lemos et al., 2022) y España en dos investigaciones distintas: una en población infantojuvenil (TECA) desarrollada por López-Pérez et al. (2014) y otra en población escolar vasca realizada por Gorostiaga et al. (2014). Todas las adaptaciones se llevaron a cabo mediante procesos de traducción, retrotraducción, revisión semántica y juicio de expertos, lo que garantizó la equivalencia conceptual y la adecuación cultural del instrumento en los contextos iberoamericanos.
Estructura	El TECA evalúa cuatro dimensiones fundamentales de la empatía: adopción de perspectivas (AP), comprensión emocional (CE), estrés empático (EE) y alegría empática (AE). Las versiones adaptadas mantienen esta estructura teórica, aunque con ligeras variaciones en el número de ítems por subescala. Por ejemplo, la adaptación brasileña TECA-BR incluye 31 ítems distribuidos en AP = 9, CE = 9, EE = 7 y AE = 6 (Freitas y Miranda, 2021), mientras que la versión breve argentina conserva 15 ítems con cuatro factores equivalentes

	(Lemos et al., 2022). La versión infantojuvenil TECA-NA mantiene la misma organización teórica, confirmando la correspondencia entre las dimensiones cognitivas y afectivas (López-Pérez et al., 2014).
Descripción de los estudios seleccionados	Se revisaron cinco estudios representativos de las validaciones y adaptaciones del TECA en contextos iberoamericanos. En Brasil, Freitas y Miranda (2021) realizaron una adaptación transcultural del TECA en una muestra de $n = 256$ estudiantes universitarios, con predominio femenino (71.10 %), y evaluaron su estructura factorial y confiabilidad. En México, Chavira Trujillo y Celis de la Rosa (2022) desarrollaron una validación psicométrica en población universitaria, comparando la estructura mexicana con la española para examinar la invariancia métrica. En Argentina, Lemos et al. (2022) diseñaron una versión breve de 15 ítems aplicada a una muestra de $n = 420$ estudiantes universitarios, buscando un modelo más parsimonioso con igual validez teórica. En España, se identifican dos estudios clave: Gorostiaga et al. (2014), quienes adaptaron el TECA a población infantojuvenil vasca ($n = 1618$), y López-Pérez et al. (2014), quienes desarrollaron la versión TECA-NA con una muestra de $n = 670$ niños y adolescentes de 10 a 16 años.
Evidencias de Validez	En cuanto a la validez de contenido, las adaptaciones coincidieron en aplicar procesos formales de traducción y retrotraducción, revisión por jueces expertos y pruebas piloto con la población meta. En Brasil, la adaptación incluyó revisión lingüística y semántica con especialistas en psicometría, garantizando la equivalencia conceptual (Freitas y Miranda, 2021). En España, la versión infantil se revisó por tres expertos en desarrollo infantil y se depuraron ítems por comprensibilidad (López-Pérez et al., 2014). En Argentina, la versión breve fue revisada por jueces y sometida a una prueba piloto con 83 estudiantes (Lemos et al., 2022). En México se evidenciaron limitaciones en la equivalencia métrica al no conservar la invariancia estructural frente al modelo español (Chavira Trujillo y Celis de la Rosa, 2022). Respecto a la validez de constructo, en Brasil se confirmó la estructura de cuatro factores mediante análisis factorial exploratorio (AFE) con rotación oblicua, $KMO = .79$, $\chi^2(528) = 1919.58$, $p < .001$, explicando el 35.66 % de la varianza total (Freitas y Miranda, 2021). En España, el TECA-NA mostró un AFE con cuatro factores (38.88 % de varianza) y un análisis factorial confirmatorio (AFC) con excelente ajuste: $\chi^2(334) = 811.80$, $RMSEA = .05$, $CFI = .94$ y $TLI = .95$ (López-Pérez et al., 2014). En Argentina, el AFC indicó mejor ajuste para el modelo de 15 ítems y cuatro factores ($\chi^2/gl = 2.14$; $CFI = .96$; $RMSEA = .04$) en comparación con los modelos alternativos (Lemos et al., 2022). En México, el análisis multigrupo mostró diferencias en la estructura factorial ($\Delta CFI > .01$), indicando falta de invariancia transcultural (Chavira Trujillo y Celis de la Rosa, 2022). En población escolar vasca, el modelo de cuatro factores fue replicado con excelente ajuste ($CFI = .95$; $RMSEA = .04$) tras eliminar ítems ambiguos (Gorostiaga et al., 2014). La validez convergente y discriminante se reporta mediante correlaciones positivas entre el TECA y otras medidas teóricamente relacionadas, como el IECA, COOL y TMMS-23, y correlaciones negativas con medidas de agresividad o cocientes autistas (Gorostiaga et al., 2014; Lemos et al., 2022; López-Pérez et al., 2014). En Brasil, se cumplió el criterio $AVE > \phi^2$ en todos los pares de factores, confirmando validez discriminante interna (Freitas y Miranda, 2021).

Evidencias de confiabilidad	Los estudios revisados reportan consistencia interna aceptable a alta. En Brasil, $\alpha_{\text{total}} = .82$ y, por subescala, $AP = .71$, $CE = .67$, $EE = .76$ y $AE = .68$, con confiabilidad compuesta (CC) entre $.77$ y $.81$ (Freitas y Miranda, 2021). En España, el TECA-NA reportó $\alpha_{\text{total}} = .77$ y $\alpha_{\text{subescalas}}$ entre $.71$ y $.73$ (López-Pérez et al., 2014), mientras que en la muestra vasca se obtuvo $\omega_{\text{global}} = .89$ y $\omega_{\text{subescalas}}$ entre $.66$ y $.86$ (Gorostiaga et al., 2014). En Argentina, la fiabilidad fue $FC_{\text{global}} = .90$ y $\omega_{\text{subescalas}}$ entre $.55$ y $.82$ en el Estudio 1, con mejoras en el Estudio 2 (Lemos et al., 2022). En México, los coeficientes oscilaron entre $.64$ y $.76$ y $\alpha_{\text{total}} \approx .86$ (Chavira Trujillo y Celis de la Rosa, 2022). Ningún estudio evaluó estabilidad temporal (<i>test-retest</i>) ni fiabilidad interevaluador.
Hallazgos según criterios COSMIN	La evidencia psicométrica del TECA se considera moderada según los criterios COSMIN. Los estudios confirman la estructura de cuatro factores propuesta por López-Pérez et al. (2008), con replicabilidad adecuada y consistencia interna sólida. La calidad metodológica es moderada-alta en estructura interna y confiabilidad, pero limitada en estabilidad temporal y análisis de invarianza. Las adaptaciones en España, Brasil y Argentina muestran buen ajuste estructural y adecuada validez convergente, mientras que México presentó problemas de equivalencia transcultural. La evidencia global indica bajo riesgo de sesgo en consistencia interna y validez, pero evidencia insuficiente en análisis longitudinales y multigrupo. Se recomienda ampliar la investigación mediante estudios longitudinales, análisis de <i>test-retest</i> y validaciones cruzadas entre contextos culturales iberoamericanos para fortalecer la robustez psicométrica del instrumento.

Instrumento 9. Sistema de Evaluación de Niños y Adolescentes (SENA)

El Sistema de Evaluación de Niños y Adolescentes (SENA) es una batería psicométrica estandarizada diseñada para evaluar de forma integral el ajuste emocional y comportamental en población infantil y adolescente. Se estructura en tres versiones paralelas para padres o cuidadores, profesores y autoinforme, adaptadas a tres rangos de edad (3–6, 6–12 y 12–18 años). Su objetivo es identificar problemas emocionales y conductuales, vulnerabilidades y recursos psicológicos, utilizando criterios derivados del DSM-5. La puntuación se obtiene mediante una escala tipo Likert y se interpreta a partir de puntuaciones T normativas españolas ($M = 50$, $DE = 10$), que permiten comparar el desempeño del individuo con su grupo de referencia (Fernández-Pinto et al., 2015).

El instrumento se aplica principalmente en contextos psicoeducativos y clínicos, integrando múltiples informantes para ofrecer una visión ecológica del comportamiento del niño o adolescente tanto en el hogar como en la escuela. En el rango de 6 a 12 años, que fue el utilizado en la investigación revisada, se incluyen escalas vinculadas con procesos ejecutivos observables, como la atención y la hiperactividad-impulsividad, que resultan útiles para evaluar

las funciones ejecutivas conductuales desde la percepción de padres y docentes. El estudio analizado por Rossignoli-Palomeque y Pérez-Hernández (2022) se llevó a cabo con una muestra de $n = 140$ estudiantes escolarizados en Madrid, con edades entre 6 y 12 años ($M = 6.97$, $DE = 1.02$), provenientes de dos centros educativos de nivel sociocultural medio-alto. Se empleó un diseño *ex post facto* correlacional, aplicando las versiones del SENA y del BRIEF-2 a padres y profesores, con el fin de analizar la validez convergente de las escalas relacionadas con funciones ejecutivas.

Respecto a la validez de contenido, el estudio no realizó un análisis específico en su muestra, como el uso de paneles de expertos o índices de validez de contenido (*CVI* o *CVR*). Los autores indicaron que el SENA ya contaba con procedimientos de adaptación y validación previos realizados en España, garantizando la equivalencia conceptual de los ítems y su adecuación cultural (Fernández-Pinto et al., 2015). Por tanto, en este caso, la evidencia de contenido se considera indirecta, sustentada en la documentación técnica del manual y no en análisis empíricos nuevos (Rossignoli-Palomeque & Pérez-Hernández, 2022).

En cuanto a la validez de constructo, no se llevaron a cabo análisis factoriales exploratorios o confirmatorios con la muestra empleada. Los autores citaron resultados de estudios anteriores donde se había obtenido una estructura trifactorial del SENA con índices de ajuste adecuados (*CFI* entre .84 y .98), pero estos valores no fueron recalculados en la investigación analizada. De esta forma, la evidencia estructural es secundaria y depende de trabajos previos, sin verificarse su replicabilidad con los datos actuales (Fernández-Pinto et al., 2015; Rossignoli-Palomeque & Pérez-Hernández, 2022).

En relación con la validez convergente, los resultados aportan evidencia sólida y cuantificable. Se observaron correlaciones de Spearman de magnitud alta y muy alta entre los dominios análogos del SENA y del BRIEF-2. En los informes de los padres ($n = 119$), el compuesto global de funciones ejecutivas del BRIEF-2 correlacionó con los problemas de funciones ejecutivas del SENA en $r = .69$, $p < .01$; la escala de inhibición del BRIEF-2 con los problemas de funciones ejecutivas del SENA en $r = .57$, $p < .01$; y la memoria de trabajo del BRIEF-2 con los problemas de atención del SENA en $r = .74$, $p < .01$. En los informes de los profesores ($n = 140$), las correlaciones fueron aún más altas: $r = .86$ entre el compuesto global del BRIEF-2 y los problemas de funciones ejecutivas del SENA; $r = .9$ entre la inhibición del BRIEF-2 y la hiperactividad-impulsividad del SENA; y $r = .83$ entre la memoria de trabajo del

BRIEF-2 y los problemas de atención del SENA (todas $p < .01$). Estas asociaciones respaldan una validez convergente alta a muy alta, confirmando que ambas pruebas miden constructos equivalentes vinculados al funcionamiento ejecutivo (Rossignoli-Palomeque & Pérez-Hernández, 2022).

En lo que respecta a la confiabilidad, el estudio no reportó estimaciones directas de consistencia interna ni de estabilidad temporal. Los autores citaron valores de confiabilidad provenientes de estudios previos y del manual técnico del SENA, que sitúan los coeficientes alfa de Cronbach en $\alpha \geq .7$, y del BRIEF-2, donde se documentan valores promedio de $\alpha \approx .86$ en diferentes escalas (Fernández-Pinto et al., 2015; Gioia et al., 2016). No se realizaron análisis de *test-retest* ni de fiabilidad interevaluador, por lo que la evidencia de consistencia interna es indirecta y derivada de fuentes anteriores, mientras que la estabilidad temporal permanece sin explorar (Rossignoli-Palomeque & Pérez-Hernández, 2022).

De acuerdo con los criterios COSMIN, la evidencia obtenida para el SENA en este estudio alcanza un nivel de evidencia moderado en validez y bajo en confiabilidad. La validez de contenido y la estructura interna se apoyan en resultados de validaciones previas, por lo que se clasifican con nivel bajo de evidencia, al no haberse replicado empíricamente en esta muestra. Sin embargo, la validez convergente presenta resultados consistentes y estadísticamente sólidos, con correlaciones entre $r = .57$ y $r = .90$ que evidencian una relación alta y coherente entre las escalas del SENA y del BRIEF-2, lo que respalda un nivel moderado de evidencia en este dominio. La confiabilidad, aunque teóricamente aceptable, carece de verificación empírica en este análisis, lo que le otorga un nivel bajo de evidencia según los estándares COSMIN. En conjunto, el estudio ofrece una calidad psicométrica general moderada, con adecuada validez convergente para medir funciones ejecutivas en población escolar de 6 a 12 años, pero requiere fortalecerse mediante análisis de estructura interna, estabilidad temporal e invariancia muestral (Fernández-Pinto et al., 2015; Gioia et al., 2015; Rossignoli-Palomeque & Pérez-Hernández, 2022).

Tabla 14

Sistema de Evaluación de Niños y Adolescentes (SENA)

Sección	Descripción
	Nombre: Sistema de Evaluación de Niños y Adolescentes (SENA)
	Autor original: Fernández-Pinto, I., Santamaría, P., Sánchez-Sánchez, F., & Carrasco, M. Á.
	Año: 2015

Datos generales	Constructo: Evaluación integral de los problemas emocionales, conductuales y del ajuste psicológico en población infantil y adolescente, incorporando tanto vulnerabilidades como recursos personales. El instrumento se basa en criterios diagnósticos del DSM-5 y permite una evaluación ecológica en los contextos familiar, escolar y social. Número de ítems: Varía según la versión y el rango de edad, entre aproximadamente 100 y 200 ítems. Las escalas se organizan en torno a problemas emocionales, conductuales, de atención y aprendizaje, así como en competencias adaptativas y recursos psicológicos. Formato y aplicabilidad: El SENA utiliza una escala tipo Likert de cinco puntos (0 = nunca o casi nunca; 4 = siempre o casi siempre). Se aplica de forma individual o colectiva, en formato papel o digital, y las puntuaciones se transforman en puntuaciones <i>T</i> normativas españolas ($M = 50$, $DE = 10$) (Fernández-Pinto et al., 2015). Adaptación cultural: El instrumento fue estandarizado y validado en España con normas representativas para las tres franjas etarias (3–6, 6–12 y 12–18 años). Las adaptaciones se desarrollaron siguiendo procedimientos de traducción, validación y normativización locales. Posteriormente, estudios como el de Rossignoli-Palomeque y Pérez-Hernández (2022) confirmaron su aplicabilidad en contextos escolares españoles.
Estructura	El SENA cuenta con tres versiones (padres, profesores y autoinforme) adaptadas a tres grupos etarios (3–6, 6–12 y 12–18 años). Cada versión se organiza en escalas que evalúan dominios psicológicos amplios: problemas emocionales (ansiedad, depresión, somatización), problemas conductuales (agresividad, desobediencia, hiperactividad-impulsividad), dificultades atencionales y de aprendizaje, relaciones interpersonales y recursos personales (autoestima, afrontamiento y habilidades sociales). En el grupo de 6–12 años, las escalas de atención e hiperactividad-impulsividad se vinculan directamente con indicadores de funciones ejecutivas (Fernández-Pinto et al., 2015).
Descripción de los estudios seleccionados	El estudio de Rossignoli-Palomeque y Pérez-Hernández (2022) se llevó a cabo con una muestra de $n = 140$ niños de entre 6 y 12 años ($M = 6.97$, $DE = 1.02$) escolarizados en Madrid, con un nivel sociocultural medio-alto. Participaron tanto padres ($n = 119$) como profesores ($n = 140$), quienes completaron las versiones correspondientes del SENA 6–12 y del BRIEF-2. El diseño fue <i>ex post facto</i> de tipo correlacional, y su propósito fue analizar la validez convergente entre las escalas de ambas pruebas en la medición de las funciones ejecutivas en contextos educativos. Los datos se procesaron mediante correlaciones de Spearman sobre puntuaciones <i>T</i>, controlando la edad de los participantes.
Evidencias de validez	En cuanto a la validez de contenido, el estudio no realizó un procedimiento empírico directo (por ejemplo, evaluación por jueces expertos o índices <i>CVI/CVR</i>). La adecuación conceptual y semántica de los ítems se fundamentó en la estandarización previa del instrumento realizada por Fernández-Pinto et al. (2015), quienes documentaron la equivalencia lingüística y cultural del SENA. Respecto a la validez de constructo, los autores no aplicaron nuevos análisis factoriales con su muestra. En su lugar, citaron resultados previos del manual técnico del SENA, en los cuales se habían obtenido índices de ajuste satisfactorios para la estructura trifactorial general del instrumento (<i>CFI</i> entre .84 y .98), lo que respalda la coherencia interna del modelo teórico, pero no confirma su replicabilidad con la muestra actual (Rossignoli-Palomeque & Pérez-Hernández, 2022). En relación con la validez convergente, los resultados mostraron correlaciones de magnitud alta a muy alta entre las escalas del SENA y del BRIEF-2, lo que aporta evidencia sólida de relación entre ambos instrumentos. En los informes parentales ($n =$

	119), se hallaron asociaciones significativas entre el compuesto global de funciones ejecutivas del BRIEF-2 y los problemas de funciones ejecutivas del SENA ($r = .69, p < .01$), entre la escala de inhibición del BRIEF-2 y los problemas de funciones ejecutivas del SENA ($r = .57, p < .01$), y entre la memoria de trabajo del BRIEF-2 y los problemas de atención del SENA ($r = .74, p < .01$). En los informes docentes ($n = 140$), las correlaciones fueron aún más altas: $r = .86$ entre el compuesto global del BRIEF-2 y los problemas de funciones ejecutivas del SENA; $r = .90$ entre inhibición e hiperactividad-impulsividad; y $r = .83$ entre memoria de trabajo y atención (todas $p < .01$). Estos resultados reflejan una validez convergente de nivel moderado según COSMIN, ya que la evidencia proviene de un solo estudio, pero con coeficientes consistentes y metodológicamente sólidos (Rossignoli-Palomeque & Pérez-Hernández, 2022).
Evidencias de confiabilidad	El estudio no reportó coeficientes propios de consistencia interna ni de estabilidad temporal. Los autores citaron los valores del manual técnico del SENA, en los que se documentan alfas de Cronbach iguales o superiores a $\alpha = .70$ en la mayoría de las escalas, y los del BRIEF-2, que presentan promedios cercanos a $\alpha = .86$ (Fernández-Pinto et al., 2015; Gioia et al., 2015). Dado que los valores de confiabilidad no fueron estimados en esta muestra, la evidencia de fiabilidad se considera de nivel bajo según COSMIN, pues depende de datos secundarios y no de estimaciones empíricas del estudio.
Hallazgos según criterios COSMIN	La evaluación integral de la calidad metodológica del estudio según los criterios COSMIN indica que la validez convergente alcanza un nivel de evidencia moderado, debido a la consistencia de las correlaciones y a la correspondencia teórica entre los instrumentos. La validez de contenido y la estructura interna presentan evidencia baja, dado que no se realizaron nuevos análisis en la muestra y los resultados se sustentan en fuentes previas. En cuanto a la confiabilidad, la evidencia también se clasifica como baja debido a la ausencia de estimaciones propias de consistencia interna o de estabilidad temporal. En general, la calidad psicométrica del estudio puede considerarse moderada en validez y baja en confiabilidad, de acuerdo con los estándares COSMIN. Se recomienda incorporar futuros análisis factoriales confirmatorios, estudios de test-retest y validaciones cruzadas para fortalecer la robustez del instrumento en contextos escolares (Fernández-Pinto et al., 2015; Rossignoli-Palomeque & Pérez-Hernández, 2022).

Instrumento 10. Cuestionario para la Evaluación Integral de la Idoneidad para la Adopción (CUIDA)

El Cuestionario para la Evaluación Integral de la Idoneidad para la Adopción (CUIDA) es un instrumento psicométrico diseñado para evaluar las competencias psicológicas, emocionales y relacionales de las personas que aspiran a adoptar. Su propósito principal es determinar el grado de idoneidad parental desde una perspectiva integral, considerando factores de personalidad, madurez emocional, empatía, tolerancia a la frustración y capacidad para establecer vínculos afectivos. El CUIDA se aplica en contextos jurídicos y forenses como herramienta técnica dentro de los procesos de adopción, proporcionando información complementaria para la toma de decisiones sobre la aptitud psicológica de los candidatos (Arredondo et al., 2019).

El instrumento está compuesto por 189 ítems con formato de respuesta tipo Likert de 1 a 4 puntos y evalúa catorce variables de personalidad, junto con tres factores de segundo orden y una escala adicional relacionada con la agresividad. Estas dimensiones permiten medir tanto recursos personales y sociales como posibles rasgos de vulnerabilidad que puedan afectar el ejercicio de la parentalidad adoptiva. Las puntuaciones se transforman en baremos de eneatis (M = 5; DE = 2), lo que permite interpretar los resultados dentro de un marco normativo. En el estudio colombiano, el instrumento fue autoadministrado en una sesión de psicología con una duración aproximada de dos horas, en el contexto de una evaluación judicial de adopción (Arredondo et al., 2019).

En cuanto al origen, estructura y aplicabilidad, el CUIDA fue desarrollado específicamente para evaluar la idoneidad parental en procesos adoptivos, integrando catorce escalas primarias que reflejan rasgos de personalidad y tres escalas de segundo orden que representan estilos de ajuste psicológico y socioemocional. El estudio realizado por Arredondo et al. (2019) aplicó el instrumento a una muestra de $n = 80$ adultos (40 hombres y 40 mujeres) en proceso de adopción nacional en Colombia. La muestra, seleccionada por conveniencia, presentó edades promedio de 42 años ($DE = 4.76$) en hombres y 37 años ($DE = 4.79$) en mujeres, pertenecientes a un nivel socioeconómico medio. La evaluación se realizó de manera individual dentro del procedimiento de valoración psicológica.

Respecto a la validez de contenido, no se reportaron procedimientos empíricos específicos, como juicios de expertos o índices *CVI* o *CVR*. Esto indica que la evidencia de contenido es indirecta y se apoya en las propiedades psicométricas previamente establecidas por los autores del instrumento original, sin nuevos análisis aplicados a la muestra colombiana. Según los criterios COSMIN, la validez de contenido se clasifica como muy baja (Arredondo et al., 2019).

En cuanto a la validez de constructo, los resultados del análisis de ecuaciones estructurales mostraron un ajuste excelente del modelo que explica la idoneidad parental. Se obtuvieron los siguientes indicadores: $\chi^2 = 2.91$, $p = .57$; $\chi^2/\text{gl} = 0.73$; $RMSEA = .00$; $PCLOSE = .66$; $GFI = .98$ y $AGFI = .95$, con una varianza explicada del 70 %. Un modelo previo había mostrado un ajuste aceptable pero menor ($\chi^2 = 22.88$, $p = .30$; $\chi^2/\text{gl} = 1.14$; $RMSEA = .04$; $GFI = .99$; $AGFI = .44$), y una varianza explicada del 51%. Estos resultados indican que la estructura interna del instrumento es coherente y estable en población colombiana, aunque la

representatividad limitada de la muestra reduce la generalización de los hallazgos (Arredondo et al., 2019).

En relación con la validez convergente, se hallaron correlaciones positivas entre distintas escalas del CUIDA, con valores entre $r = .5$ y $r = .6$, como entre distimia y reflexividad ($r = .6$), distimia y cuidado responsable ($r = .56$) y narcisismo y asertividad ($r = .53$). Sin embargo, las correlaciones con un instrumento externo de referencia, el *Millon Clinical Multiaxial Inventory-III* (MCMI-III), fueron inferiores a .6 y la covarianza latente entre ambos modelos no fue significativa ($r = -.757$, $p = .69$), lo que evidencia una validez convergente baja. De acuerdo con los criterios COSMIN, la evidencia convergente se clasifica con un nivel bajo (Arredondo et al., 2019).

En cuanto a la validez de criterio, se encontraron resultados significativos que sustentan la capacidad predictiva del CUIDA respecto a la idoneidad parental. Los coeficientes de regresión mostraron efectos robustos con $B = 0.1$ ($p < .001$) para la idoneidad global, $B = 1.05$ ($p < .001$) para el cuidado afectivo y $B = 1.04$ ($p < .001$) para la sensibilidad hacia los demás. La regresión logística indicó una clasificación correcta del 82.5 % de los casos, con predictores significativos como compulsión ($B = -0.1$, $p < .001$) y sensibilidad hacia los demás ($B = 0.71$, $p < .001$). Estos resultados confirman una validez criterial moderada (Arredondo et al., 2019).

En cuanto a la confiabilidad, el estudio reportó coeficientes alfa de Cronbach entre $\alpha = .66$ y $\alpha = .86$, indicando una consistencia interna aceptable en la mayoría de las escalas. No obstante, no se calcularon índices complementarios como fiabilidad compuesta, error estándar de medición ni estabilidad temporal test–retest, por lo que la evidencia de confiabilidad se clasifica como baja a moderada según COSMIN (Arredondo et al., 2019).

Finalmente, según los criterios COSMIN, la validez estructural y criterial del CUIDA presentan un nivel moderado de evidencia, mientras que la validez convergente y la confiabilidad alcanzan un nivel bajo, y la validez de contenido se clasifica como muy baja. Esto refleja que el instrumento cuenta con una estructura teórica sólida y una capacidad predictiva adecuada, pero requiere fortalecerse mediante estudios de validez de contenido, convergencia externa y análisis de estabilidad temporal. En conjunto, el CUIDA muestra propiedades psicométricas aceptables para la evaluación de la idoneidad parental en el contexto colombiano, aportando información relevante para los procesos jurídicos de adopción (Arredondo et al., 2019).

Tabla 15*Cuestionario para la Evaluación Integral de la Idoneidad para la Adopción (CUIDA)*

Sección	Descripción
Datos generales	Nombre: Cuestionario para la Evaluación Integral de la Idoneidad para la Adopción (CUIDA) Autor original: Londoño Arredondo, N. H., Mejía, C. M., Acosta, F. J., Rodríguez, O. E., y Ramírez, N. H. Año: 2019 Constructo: Evaluación integral de la idoneidad psicológica para la adopción, considerando dimensiones de personalidad, ajuste emocional, recursos interpersonales y estilos de afrontamiento. El instrumento permite identificar fortalezas y posibles riesgos psicológicos que puedan afectar el ejercicio de la parentalidad adoptiva. Número de ítems: 189 ítems distribuidos en catorce escalas primarias, tres escalas de segundo orden y una escala adicional de agresividad. Formato y aplicabilidad: Cuestionario autoadministrado con formato de respuesta tipo Likert de 1 a 4 puntos, donde valores más altos reflejan mayor presencia del rasgo evaluado. El tiempo aproximado de aplicación es de dos horas, y se utiliza en contextos jurídicos y clínico-forenses, principalmente en procesos de adopción. Adaptación cultural: El CUIDA fue aplicado y validado en Colombia en población de padres en proceso de adopción, incorporando adecuaciones culturales acordes con el contexto jurídico y social colombiano. El instrumento fue diseñado originalmente en español, por lo que no requirió procesos de traducción ni adaptación lingüística (Arredondo et al., 2019).
Estructura	El CUIDA está conformado por catorce escalas primarias que miden dimensiones de personalidad y ajuste emocional, tres escalas de segundo orden que integran dichas dimensiones en factores globales y una escala adicional enfocada en la agresividad. Las escalas abarcan variables como autoestima, sociabilidad, tolerancia a la frustración, cuidado afectivo, sensibilidad hacia los demás, equilibrio emocional, responsabilidad, impulsividad, distimia, narcisismo y compulsión. Las puntuaciones se interpretan con baremos en eneatis (media = 5; desviación estándar = 2), lo que permite una clasificación normativa de los resultados (Arredondo et al., 2019).
Descripción de los estudios seleccionados	El estudio realizado por Arredondo et al. (2019) se desarrolló con una muestra de $n = 80$ adultos colombianos (40 hombres y 40 mujeres) en proceso de adopción nacional, con edades promedio de 42 años ($DE = 4.76$) en hombres y 37 años ($DE = 4.79$) en mujeres. La muestra fue de tipo no probabilístico por conveniencia y los participantes completaron el CUIDA de forma autoadministrada en una sesión individual dentro del proceso de valoración psicológica. El diseño fue transversal y no experimental, orientado a la evaluación de la validez estructural, convergente y criterial del instrumento mediante análisis correlacionales, regresiones y modelos de ecuaciones estructurales. El contexto de aplicación fue jurídico-forense, específicamente en procesos de adopción tramitados en instituciones oficiales de Colombia. En lo referente a la validez de contenido, el estudio no reportó procedimientos empíricos de revisión o validación de ítems mediante paneles de expertos o índices

Evidencias de validez	de validez de contenido (<i>CVI</i> o <i>CVR</i>), por lo que esta evidencia se considera ausente y se clasifica con un nivel muy bajo según los estándares COSMIN. Respecto a la validez de constructo, se realizó un análisis de ecuaciones estructurales (<i>SEM</i>) que mostró un excelente ajuste del modelo explicativo de la idoneidad parental ($\chi^2 = 2.91, p = .57; \chi^2/\text{gl} = 0.73; \text{RMSEA} = .00; \text{PCLOSE} = .66; \text{GFI} = .98; \text{AGFI} = .95$), con una varianza explicada del 70%. Un modelo previo, de menor ajuste, había explicado el 51% de la varianza ($\chi^2 = 22.88, p = .30; \text{RMSEA} = .04; \text{GFI} = .99; \text{AGFI} = .44$). Estos resultados respaldan la validez estructural del instrumento, clasificándose con un nivel moderado según COSMIN (Arredondo et al., 2019). En relación con la validez convergente, se reportaron correlaciones significativas y teóricamente coherentes entre escalas internas del CUIDA, como distimia y reflexividad ($r = .60$), distimia y cuidado responsable ($r = .56$) y narcisismo y asertividad ($r = .53$). Sin embargo, la covarianza entre el CUIDA y el <i>Millon Clinical Multiaxial Inventory–III</i> (MCMI-III) fue negativa y no significativa ($r = -7.57, p = .69$), lo que indica baja convergencia entre ambos modelos teóricos. De acuerdo con COSMIN, esta evidencia se clasifica como un nivel bajo de validez convergente. Por último, la validez se comprobó mediante análisis de regresión y modelos estructurales, encontrando relaciones significativas entre el CUIDA y la idoneidad parental ($B = 0.10, p < .001$), el cuidado afectivo ($B = 1.05, p < .001$) y la sensibilidad hacia los demás ($B = 1.04, p < .001$). Además, la regresión logística binaria mostró una clasificación correcta del 82.50 % de los casos, con predictores relevantes como compulsión ($B = -0.10, p < .001$) y sensibilidad hacia los demás ($B = 0.71, p < .001$). Estos hallazgos sustentan una validez criterial de nivel moderado (Arredondo et al., 2019).
Evidencias de confiabilidad	El estudio reportó coeficientes alfa de Cronbach entre $\alpha = .66$ y $\alpha = .86$, lo que indica una consistencia interna adecuada en la mayoría de las escalas del instrumento. No obstante, no se realizaron análisis de estabilidad temporal (<i>test-retest</i>) ni estimaciones de error estándar de medición o fiabilidad compuesta. En este sentido, la evidencia de confiabilidad del CUIDA se considera de nivel bajo a moderado de acuerdo con los estándares COSMIN, pues, aunque presenta consistencia interna aceptable, carece de pruebas complementarias de precisión psicométrica (Arredondo et al., 2019).
Hallazgos según criterios COSMIN	De acuerdo con los criterios de calidad metodológica COSMIN, la evidencia psicométrica del CUIDA muestra un nivel moderado en validez estructural y criterial, bajo en validez convergente y en confiabilidad, y muy bajo en validez de contenido. Estos resultados indican que el instrumento posee una base teórica sólida y una estructura interna coherente, respaldada por modelos estructurales estadísticamente satisfactorios. No obstante, la ausencia de análisis de contenido, convergencia externa y estabilidad temporal limita la solidez de sus propiedades psicométricas. En conjunto, la calidad general del estudio se clasifica como moderada, lo que respalda la utilidad del CUIDA como herramienta válida y clínicamente pertinente para la evaluación de la idoneidad parental en contextos jurídicos colombianos, aunque requiere fortalecer la evidencia empírica en futuras investigaciones (Arredondo et al., 2019).

Instrumento 11. Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI)

El Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI) es un instrumento psicométrico diseñado para evaluar los niveles de adaptación personal, social y escolar en niños y adolescentes. Se trata de un cuestionario de autoinforme compuesto por 175 ítems con formato de respuesta dicotómica (sí/no), estructurado para detectar tanto factores de ajuste como indicadores de inadaptación. Su objetivo principal es identificar áreas de riesgo emocional, conductual y social que puedan afectar el desarrollo integral del menor, siendo ampliamente utilizado en contextos educativos, clínicos y psicosociales. El TAMAI ofrece perfiles individuales de adaptación global y por subdimensiones, organizados en baremos de siete niveles y puntos críticos, permitiendo un análisis diferenciado según edad y sexo (Hernández-Guanir, 2002).

El instrumento cuenta con una estructura multifactorial que abarca tres grandes dominios: adaptación personal, adaptación social y adaptación escolar. Cada dominio se compone de subescalas que evalúan aspectos específicos como autoconcepto, control emocional, relaciones interpersonales, rendimiento académico y ajuste conductual. Las puntuaciones del TAMAI se interpretan mediante niveles normativos, y el instrumento ha mostrado altos índices de consistencia interna ($\alpha \geq .80$) en sus manuales originales, lo que respalda su fiabilidad en contextos de aplicación infantil y juvenil (Hernández-Guanir, 2002).

En el estudio iberoamericano revisado, Moreno-Manso et al. (2010) aplicaron el TAMAI en combinación con el instrumento BLOC-S (módulo de pragmática del lenguaje) con el propósito de analizar la relación entre competencia pragmática y adaptación psicosocial en menores bajo medidas de protección. La muestra estuvo conformada por $n = 74$ niños y niñas residentes en centros de acogida en España (41 varones y 33 mujeres), distribuidos en grupos etarios de 6 a 18 años, con puntuaciones promedio de adaptación entre 3.05 y 3.2. El diseño fue transversal y correlacional, aplicándose el TAMAI de forma grupal (duración aproximada de dos horas) y el BLOC-S de manera individual (≈ 30 minutos). Se verificaron los supuestos estadísticos de normalidad (Kolmogórov-Smirnov), homogeneidad (Levene) y aleatoriedad (Rachas) antes de los análisis principales (Moreno-Manso et al., 2010).

Respecto a la validez de contenido, el estudio no realizó procedimientos empíricos específicos, como revisión por jueces o cálculo de índices CVI/CVR . La adecuación conceptual del TAMAI en esta muestra se fundamenta en la evidencia previa documentada en el manual

técnico, en el cual se reportan amplios procesos de validación en muestras escolares con coeficientes de consistencia interna y fiabilidad interevaluador superiores a .80. Dado que estos procedimientos no fueron replicados en la presente investigación, la validez de contenido se clasifica como muy baja según los criterios COSMIN (Hernández-Guanir, 2002; Moreno-Manso et al., 2010).

En cuanto a la validez de constructo, se exploraron las relaciones entre la competencia pragmática (evaluada con el BLOC-S) y los factores de adaptación del TAMAI mediante análisis de varianza y correlaciones de Pearson. El ANOVA de un factor reveló diferencias estadísticamente significativas en la variable de somatización, $F(2, 42) = 7.76, p = .02$, indicando que los niños con menor competencia pragmática presentaron mayores niveles de somatización. A partir de este resultado, se estimó un tamaño de efecto medio-alto ($\eta^2 \approx .27$). Además, se identificó una correlación moderada entre autodesajuste social y bajo dominio pragmático ($r = .34, p < .05$). No se encontraron diferencias significativas en los factores de inadaptación social ni escolar. De acuerdo con los estándares COSMIN, esta evidencia se clasifica como validez de constructo de nivel bajo, al provenir de un único estudio con una muestra reducida, aunque los resultados son teóricamente consistentes (Moreno-Manso et al., 2010).

En relación con la validez estructural, el estudio no realizó análisis factoriales exploratorios ni confirmatorios sobre el TAMAI, apoyándose únicamente en la estructura teórica original. La ausencia de análisis factoriales en la muestra implica que esta evidencia se clasifica con un nivel muy bajo según COSMIN (Moreno-Manso et al., 2010).

En cuanto a la confiabilidad, los autores no calcularon coeficientes propios de consistencia interna ni medidas de estabilidad temporal. Sin embargo, citaron resultados del manual técnico que reportan coeficientes de consistencia interna entre $\alpha = .8$ y $\alpha = .92$, así como fiabilidad interevaluador superior a .8 y un coeficiente Spearman–Brown de .87 en una muestra de $n = 989$. Dado que estos valores corresponden a estudios previos y no se estimaron en la muestra actual, la evidencia de confiabilidad se considera baja según COSMIN (Hernández-Guanir, 2002; Moreno-Manso et al., 2010).

Finalmente, según la clasificación establecida por COSMIN, la validez de contenido y la validez estructural del TAMAI alcanzan un nivel muy bajo, mientras que la validez de constructo basada en pruebas de hipótesis se ubica en un nivel bajo, y la confiabilidad presenta también un nivel bajo debido a la ausencia de estimaciones locales. No se hallaron evidencias de validez

riterial ni de error de medida. En conjunto, la calidad psicométrica del TAMAI en este estudio se considera limitada, aunque los resultados sugieren que la inadaptación personal, especialmente la somatización, se asocia con un menor dominio pragmático del lenguaje en menores bajo protección institucional. Esto respalda el uso del TAMAI como herramienta complementaria en la evaluación psicosocial, aunque se recomienda su aplicación en estudios con muestras más amplias, grupos control y análisis factorial confirmatorio para fortalecer su evidencia empírica (Hernández-Guanir, 2002; Moreno-Manso et al., 2010).

Tabla 16

Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI)

Sección	Descripción
Datos generales	Nombre: Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI) Autor original: Hernández-Guanir, P. Año: 2002 Constructo: Evalúa la adaptación personal, social y escolar en población infantil y adolescente, identificando tanto indicadores de inadaptación (emocional, social o académica) como recursos de ajuste psicológico y comportamental. Permite determinar perfiles individuales de adaptación global y específicos por subdimensión. Número de ítems: 175 ítems distribuidos en escalas primarias que abarcan las tres áreas de adaptación principales. Formato y aplicabilidad: Cuestionario de autoinforme con formato dicotómico de respuesta (sí/no). Las puntuaciones se interpretan mediante niveles normativos (siete niveles y puntos críticos). El instrumento se aplica en contextos educativos, clínicos y psicosociales, siendo útil en la detección temprana de dificultades adaptativas y en programas de intervención escolar y familiar. Adaptación cultural: El TAMAI fue desarrollado originalmente en España por Hernández-Guanir (2002) y ha sido aplicado en diversos contextos hispanohablantes. En el estudio realizado por Moreno-Manso et al. (2010), se empleó la versión española del instrumento sin modificaciones lingüísticas o conceptuales, dado que la muestra pertenecía también al contexto español. Los autores señalan que el TAMAI mantiene una adecuada equivalencia cultural y semántica para la población evaluada, respaldada por la evidencia previa de validación documentada en su manual técnico, lo que garantiza su aplicabilidad en contextos psicoeducativos y sociales dentro del ámbito iberoamericano.
Estructura	El TAMAI presenta una estructura multifactorial que abarca tres grandes áreas: adaptación personal, adaptación social y adaptación escolar. Cada dominio integra subescalas que evalúan variables como autoconcepto, control emocional, relaciones interpersonales, rendimiento académico, responsabilidad, impulsividad y equilibrio afectivo. El diseño del instrumento permite obtener tanto una puntuación global de adaptación como perfiles específicos, utilizando baremos organizados en siete niveles de referencia y puntuaciones críticas interpretadas según edad y sexo (Hernández-Guanir, 2002).
Descripción de los estudios	El estudio revisado fue realizado por Moreno-Manso et al. (2010), quienes analizaron la relación entre la competencia pragmática del lenguaje y la adaptación psicosocial en menores bajo medidas de protección institucional. La

seleccionados	muestra estuvo conformada por $n = 74$ niños y niñas (41 varones y 33 mujeres) residentes en centros de acogida en España, con edades entre 6 y 18 años. El diseño fue transversal y correlacional, aplicándose el BLOC-S (módulo de pragmática) y el TAMAI de manera combinada para explorar las relaciones entre el lenguaje pragmático y la adaptación emocional, social y escolar. El TAMAI se administró de forma grupal durante aproximadamente dos horas, mientras que el BLOC-S fue aplicado individualmente. Se verificaron los supuestos estadísticos de normalidad, aleatoriedad y homogeneidad antes de los análisis inferenciales.
Evidencias de validez	En cuanto a la validez de contenido, el estudio no reportó procedimientos empíricos de revisión de ítems ni validación por jueces expertos, por lo que esta evidencia se considera muy baja de acuerdo con los estándares del <i>Consensus-based Standards for the Selection of Health Measurement Instruments</i> (COSMIN). La validez estructural tampoco fue analizada, dado que no se realizaron análisis factoriales exploratorios ni confirmatorios sobre la muestra. Respecto a la validez de constructo, los autores evaluaron la relación entre los factores de adaptación del TAMAI y la competencia pragmática medida con el BLOC-S. Los resultados del análisis de varianza indicaron diferencias significativas en el factor de somatización, $F(2, 42) = 7.76, p = .02$, lo que sugiere que los niños con menor competencia pragmática presentan mayores niveles de somatización. La magnitud del efecto estimada ($\eta^2 \approx .27$) indica una relación de tamaño medio-alto. Asimismo, se encontró una correlación positiva y moderada entre autodesajuste social y bajo dominio pragmático ($r = .34, p < .05$). No se observaron diferencias significativas en las áreas de adaptación social ni escolar. En consecuencia, la validez de constructo por hipótesis se clasifica como de nivel bajo, al derivarse de un único estudio con tamaño muestral reducido, aunque con resultados coherentes con la teoría del instrumento (Moreno-Manso et al., 2010).
Evidencias de confiabilidad	En términos de confiabilidad, el estudio de Moreno-Manso et al. (2010) no calculó coeficientes de consistencia interna ni pruebas de estabilidad temporal en la muestra evaluada. Sin embargo, los autores citaron los valores de fiabilidad reportados en el manual técnico del TAMAI, donde se documentan coeficientes alfa de Cronbach entre $\alpha = .80$ y $\alpha = .92$, una fiabilidad interevaluador superior a .80 y coeficientes Spearman–Brown de .87 en una muestra de $n = 989$. A pesar de estos antecedentes, la falta de verificación empírica en la muestra actual lleva a clasificar la evidencia de confiabilidad como de nivel bajo según COSMIN (Hernández-Guanir, 2002; Moreno-Manso et al., 2010).
Hallazgos según criterios COSMIN	De acuerdo con los criterios de calidad metodológica del marco COSMIN, el TAMAI presenta un nivel muy bajo de evidencia en validez de contenido y en estructura interna, un nivel bajo de evidencia en validez de constructo y también bajo en confiabilidad. No se reportaron análisis de validez criterial, error de medida o invarianza factorial. En conjunto, la calidad psicométrica del TAMAI en este estudio se considera limitada, aunque los resultados sugieren que un bajo dominio pragmático del lenguaje se asocia con mayores indicadores de inadaptación personal, particularmente somatización, en menores bajo protección institucional. El instrumento mantiene su potencial como herramienta útil en la evaluación psicosocial infantil, recomendándose su aplicación en investigaciones futuras con diseños factoriales confirmatorios, muestras más amplias y grupos control (Hernández-Guanir, 2002; Moreno-Manso et al., 2010).

Instrumento 12. Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales BANFE

El BANFE-2 (Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales) fue desarrollado por Flores-Lázaro et al. (2014) con el propósito de evaluar de manera integral las funciones ejecutivas y su relación con los distintos sistemas del lóbulo frontal. Este instrumento permite identificar alteraciones cognitivas en poblaciones clínicas y no clínicas, a través de la valoración de los sistemas órbito-medial, prefrontal anterior y dorsolateral, además de una puntuación total de funciones ejecutivas. La batería está conformada por catorce subpruebas que incluyen tareas como la fluidez fonológica, el *Stroop*, clasificación de cartas, metamemoria y el *Test de Senderos* (TMT), entre otras. Cada subprueba genera medidas de aciertos, errores y tiempo, y las puntuaciones pueden ser transformadas en índices normativos. La escala de respuesta depende de la tarea, utilizando tanto medidas dicotómicas como cuantitativas. Está dirigida principalmente a adultos y adolescentes mayores de 18 años y se ha aplicado en contextos clínicos, educativos y de investigación neuropsicológica, particularmente para valorar pacientes con lesiones frontales, esclerosis múltiple y trastornos neurocognitivos (Doldán-Peluffo et al., 2023; Flores-Lázaro et al., 2014; Puerta-Lopera et al., 2022). En los estudios revisados se destaca la falta de representatividad nacional por el uso de muestras por conveniencia y la ausencia de datos normativos amplios en población iberoamericana, aunque se ha reportado una adecuada aplicabilidad clínica y facilidad de administración en entornos neuropsicológicos (Puerta-Lopera et al., 2022).

Respecto a la evidencia de validez, los estudios no reportaron procedimientos específicos de validez de contenido mediante juicio de expertos o coeficientes de acuerdo. En cambio, la validez basada en la relación con otras variables se evaluó a través de correlaciones con medidas de fluidez verbal y otras pruebas ejecutivas. Doldán-Peluffo et al. (2023) reportaron correlaciones entre la subprueba *COW* y la fluidez fonológica que oscilaron entre $r = .29$ y $r = .39$ ($p < .05$), lo cual indica asociaciones pequeñas a moderadas; sin embargo, la mayoría de las comparaciones no fueron significativas, concluyéndose una validez concurrente débil. Por su parte, Puerta-Lopera et al. (2022) señalaron correlaciones modestas entre subpruebas del BANFE y medidas de memoria de trabajo, aunque sin alcanzar significación estadística consistente. Ninguno de los estudios revisados reportó análisis de validez discriminante o predictiva, ni se

documentaron curvas ROC o análisis de sensibilidad y especificidad frente a criterios diagnósticos clínicos.

En cuanto a la validez de constructo, los resultados muestran diferencias según el tamaño y enfoque muestral. Puerta-Lopera et al. (2022), con una muestra de $n = 208$ participantes adultos, realizaron un análisis de componentes principales (ACP) que respaldó parcialmente la estructura propuesta del BANFE-2, identificando una solución de tres factores congruente con los sistemas frontales teóricos y reportando indicadores de ajuste adecuados. En contraste, el estudio de Doldán-Peluffo et al. (2023), con una muestra más reducida de $n = 40$ participantes, no realizó análisis factorial, limitando así la evidencia estructural. En general, la validez de constructo puede considerarse moderada, sustentada principalmente en un único estudio con soporte estadístico suficiente y coherencia teórica con el modelo tripartito del funcionamiento ejecutivo (Puerta-Lopera et al., 2022).

La confiabilidad del BANFE-2 se ha evaluado a través del alfa de Cronbach y correlaciones test-retest. Doldán-Peluffo et al. (2023) reportaron los siguientes coeficientes por sistema: órbito-medial $\alpha = .7$, mejorando a $\alpha = .73$ al eliminar el ítem *Stroop A-acierto*; prefrontal anterior $\alpha = .32$, que aumentó a $\alpha = .45$ al excluir *metamemoria-error positivo*; dorsolateral $\alpha = .69$, con ligera mejora a $\alpha = .7$ al eliminar *clasificación de cartas-acierto*; y funciones ejecutivas totales $\alpha = .8$. Además, se informó $\alpha = .82$ (IC95% = .78–.86) para la subprueba *COW* y $\alpha = .67$ (IC95% = .58–.74) para *Stroop*, valores que indican buena consistencia interna global, aunque irregular en algunas áreas específicas.

En términos de estabilidad temporal, Doldán-Peluffo et al. (2023) reportaron una correlación *test-retest* de $r = .33$ en la subprueba *TMT-B*, lo que sugiere estabilidad baja a moderada. Respecto a la confiabilidad interevaluador, los resultados fueron altos, con acuerdos entre 93% y 100% y correlaciones superiores a $r = .85$ ($p < .001$). Puerta-Lopera et al. (2022) respaldan estos hallazgos, destacando la reproducibilidad de la administración y la puntuación. No obstante, ninguno de los estudios incluyó estimaciones del error de medida o análisis de invarianza factorial entre grupos, por lo que la estabilidad psicométrica entre contextos aún no está confirmada.

La evaluación metodológica según los criterios *COSMIN* mostró niveles de calidad moderada-baja en Doldán-Peluffo et al. (2023), debido al tamaño muestral limitado y la ausencia de pruebas de estructura factorial, mientras que Puerta-Lopera et al. (2022) obtuvieron una

calidad alta-moderada, sustentada en una muestra adecuada y un ACP con resultados interpretables. Ambas investigaciones coincidieron en señalar debilidades relacionadas con la validez convergente y el tamaño de las correlaciones con medidas externas, así como la necesidad de estudios adicionales que profundicen en la estabilidad longitudinal del instrumento.

En conjunto, el BANFE-2 presenta evidencias adecuadas de consistencia interna y confiabilidad interevaluador, una validez de constructo parcialmente confirmada y débil validez concurrente, con fortalezas en su aplicabilidad clínica y limitaciones metodológicas que restringen su generalización (Doldán-Peluffo et al., 2023; Puerta-Lopera et al., 2022).

Tabla 17

Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales - BANFE

Sección	Descripción
Datos generales	Nombre completo: BANFE-2 (Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales). Autores originales: Flores-Lázaro, Ostrosky-Solís y Lozano (2014). Estudios revisados: Doldán-Peluffo et al. (2023, Paraguay) y Puerta-Lopera, Dussán-Luberti y Montoya-Santacruz (año no indicado, Colombia). Constructo: Evaluación integral de funciones ejecutivas, atención, planificación, control inhibitorio y flexibilidad cognitiva. Número de subpruebas: 14 (sin número total unificado de ítems). Formato: Batería de desempeño con tareas de tiempo, aciertos y errores (<i>Stroop</i>, <i>WCST</i>, <i>COW-FAS</i>, Ordenamiento, <i>TMT</i>). Escala de respuesta: cuantitativa (tiempos, aciertos, errores). Población objetivo: adultos y jóvenes mayores de 18 años. Contextos de aplicación: clínico, educativo e investigación neuropsicológica. Restricciones: no hay datos normativos amplios en población iberoamericana; se han usado muestras por conveniencia.
Estructura del instrumento	El BANFE-2 evalúa tres sistemas frontales: órbito-medial, prefrontal anterior y dorsolateral, además de una puntuación total de funciones ejecutivas. Las subpruebas incluyen <i>Stroop</i>, <i>WCST</i>, <i>COW-FAS</i>, <i>TMT</i> y tareas de planificación y metamemoria. Cada dominio produce un puntaje parcial y una puntuación total estandarizada. Los estudios revisados replicaron parcialmente la estructura teórica original (Flores-Lázaro et al., 2014), encontrando una solución de tres componentes congruente con los sistemas frontales (Doldán-Peluffo et al., 2023; Puerta-Lopera et al., 2022).
Evidencias de validez	Validez de contenido: no se reportaron procedimientos sistemáticos (juicio de expertos o coeficientes de acuerdo). Validez de constructo: • Doldán-Peluffo et al. (2023) realizaron un <i>ACP</i> con $KMO = .85$ y Bartlett $p < .001$, obteniendo una estructura coherente con la teoría de tres sistemas. No se efectuaron análisis confirmatorios (<i>CFA</i>) ni pruebas de invarianza factorial. • Puerta-Lopera et al. (2022) replicaron parcialmente la estructura del modelo original de Flores-Lázaro et al. (2014) mediante <i>ACP</i> en una muestra mayor ($n = 208$), confirmando tres componentes principales.

	Validez concurrente: Doldán-Peluffo et al. (2023) compararon las puntuaciones del BANFE-2 con pruebas de fluidez verbal, encontrando correlaciones entre <i>COW</i> y fluidez fonológica $r = .29-.39$ ($p < .05$), relación pequeña a moderada. No se hallaron correlaciones significativas con otras medidas, ni se reportaron análisis de validez discriminante o predictiva.
Evidencias de confiabilidad	Consistencia interna (α de Cronbach): • Órbito-medial (OM): $\alpha = .70$ (mejora a $\alpha = .73$ al eliminar <i>Stroop A-acierto</i>). • Prefrontal anterior (PFA): $\alpha = .32$ (mejora a $\alpha = .45$ si se elimina <i>metamemoria-error positivo</i>). • Dorsolateral (DL): $\alpha = .69$ (mejora a $\alpha = .70$ al eliminar <i>clasificación de cartas-acierto</i>). • Funciones ejecutivas totales (FEs): $\alpha = .80$. • <i>COW-FAS</i>: $\alpha = .82$ (IC95% = .78-.86). • <i>Stroop</i>: $\alpha = .67$ (IC95% = .58-.74). Estos valores reflejan buena consistencia global y moderada en subpruebas específicas. Test-retest: $r = .33$ en <i>TMT-B</i> (baja a moderada estabilidad). Concordancia interevaluador: 93-100% de acuerdo; $r > .85$ ($p < .001$). No se reportaron errores de medida ni invarianza.
Hallazgos según criterios COSMIN	Doldán-Peluffo et al. (2023): calidad moderada-baja (muestra pequeña, sin <i>CFA</i> ni invarianza). Puerta-Lopera et al. (2022): calidad alta-moderada (<i>ACP</i> con $KMO = .85$ y Bartlett significativo, pero sin validez concurrente fuerte). En conjunto, los estudios aportan evidencia preliminar pero coherente con el modelo teórico del instrumento y una fiabilidad aceptable en varias subpruebas.

Síntesis transversal de propiedades psicométricas entre instrumentos

A partir del análisis de los cuarenta y cinco artículos que conforman la matriz psicométrica, se elaboró una síntesis comparativa de los doce instrumentos identificados. El propósito fue examinar de manera transversal las principales evidencias de fiabilidad, validez y generalización reportadas en las distintas investigaciones, considerando los ejes de consistencia interna, soporte de validez factorial, diversidad de muestras y posibles vacíos de investigación.

Los resultados muestran que los instrumentos con mayores valores promedio de consistencia interna (alfa de Cronbach) fueron el NEUROPSI, el SIMS, el TAMAI, el SCL-90 y el LSB-50. El NEUROPSI obtuvo una media de α cercana a 0.91, lo que refleja una excelente fiabilidad interna; el SIMS registró un promedio de α de .87 en distintas versiones y subescalas, y el TAMAI alcanzó valores medios de .86. De igual manera, el SCL-90 ($\alpha \approx .84$) y el LSB-50 ($\alpha \approx .82$) presentaron índices adecuados de consistencia, especialmente en contextos clínicos. En general, todos estos instrumentos mostraron niveles de homogeneidad interna aceptables, lo que sugiere una adecuada coherencia entre los ítems que evalúan el mismo constructo. No obstante,

algunos instrumentos más recientes o con menor número de estudios aún presentan variabilidad en sus coeficientes de fiabilidad, lo que demanda replicaciones adicionales.

En relación con la validez estructural, los instrumentos con mayor respaldo empírico fueron el TECA, el LSB-50, el MMPI-2-RF, el SCL-90 y el PAI. Estos presentaron mayor número de estudios que aplicaron análisis factorial confirmatorio (CFA) y reportaron índices de ajuste como CFI, TLI, RMSEA o SRMR. En particular, el TECA y el LSB-50 mostraron modelos con buenos indicadores de ajuste, evidenciando una estructura factorial coherente con sus postulados teóricos. Por su parte, el MMPI-2-RF y el PAI fueron los instrumentos con mayor frecuencia de análisis factoriales replicados en diferentes muestras, confirmando la estabilidad de su estructura interna. Sin embargo, algunos instrumentos presentan solo análisis exploratorios o reportes parciales de índices, lo que limita la interpretación completa del ajuste factorial y su validez estructural.

Respecto a la representatividad de las muestras, se identificó que los instrumentos con mayor número de estudios en distintos países fueron el MMPI-2-RF, el SCL-90, el NEUROPSI, el PAI y el SIMS. El MMPI-2-RF fue el más utilizado en contextos internacionales, con investigaciones en España, México y Canadá. El SCL-90 presentó amplia difusión en Latinoamérica, especialmente en Chile, Colombia, México y Argentina, mientras que el NEUROPSI y el PAI se han empleado en varios países de habla hispana y portuguesa. El SIMS también registró estudios multicéntricos, particularmente en Europa y América Latina. Estos hallazgos indican que dichos instrumentos poseen mayor respaldo transcultural y potencial de generalización de resultados, en comparación con otros que han sido validados en contextos únicos o con muestras reducidas.

A pesar del número considerable de estudios, se observan vacíos relevantes en la evidencia psicométrica de algunos instrumentos. Ninguno de los instrumentos analizados presentó evidencia específica en contextos jurídico-forenses, lo que representa un vacío importante considerando la relevancia de la evaluación psicológica en ámbitos periciales. Asimismo, siete instrumentos CUIDA, LSB-50, MMPI-2, PAI, SCL-90, SENA y TAMAI no reportan análisis de invarianza factorial, lo que limita la capacidad de comparar resultados entre grupos por sexo, edad o país y reduce la evidencia de estabilidad estructural entre poblaciones. También se detectaron limitaciones en el reporte de índices de ajuste, ya que algunos estudios

que mencionan el uso de CFA no informan los valores completos (por ejemplo, reportan χ^2 pero omiten CFI o RMSEA), lo que afecta la interpretación integral del modelo factorial.

En conjunto, estos resultados sugieren la necesidad de realizar estudios psicométricos más completos que incluyan análisis de invarianza, validaciones cruzadas en contextos forenses y poblaciones clínicas diversas. En síntesis, los instrumentos NEUROPSI, SIMS, MMPI-2-RF, SCL-90 y TECA destacan por su solidez en términos de consistencia interna, estructura factorial y replicación en distintos países. No obstante, se evidencia la necesidad de fortalecer los estudios de validez estructural avanzada, especialmente los de invarianza, y extender las validaciones a contextos forenses y transculturales, con el fin de robustecer el uso y la interpretación de estos instrumentos en la práctica psicológica contemporánea.

Síntesis del nivel de evidencia según criterios COSMIN

En los 44 artículos analizados, la calidad metodológica general de los estudios fue predominantemente moderada, según los criterios COSMIN. La mayoría de los instrumentos presentan diseños adecuados, con procedimientos estadísticos apropiados (como análisis factoriales exploratorios o confirmatorios, correlaciones de validez y estimaciones de consistencia interna), aunque con limitaciones en la evaluación integral de todas las propiedades psicométricas. En especial, se observó una ausencia frecuente de análisis de invarianza, confiabilidad test–retest y error de medición, lo que reduce la robustez de la evidencia.

Aproximadamente una tercera parte de los estudios alcanza niveles moderado–altos de evidencia, caracterizándose por un diseño claro, muestras adecuadas y análisis coherentes con los objetivos del estudio, pero con vacíos en aspectos como representatividad poblacional o replicación en otros contextos.

Por otro lado, un grupo reducido de investigaciones mostró alta calidad metodológica, especialmente aquellas que utilizaron meta–análisis, modelos de ecuaciones estructurales o análisis factorial confirmatorio (CFA), reportando índices completos de ajuste y consistencia interna elevados ($\alpha > .8$). Estos estudios cumplen la mayoría de los estándares COSMIN y ofrecen evidencia sólida en estructura interna, validez y confiabilidad.

Finalmente, una proporción menor de estudios se clasificó con calidad metodológica baja o moderada–baja, debido al uso de muestras pequeñas, ausencia de procedimientos de validación formal, falta de análisis factorial o reportes incompletos de fiabilidad. Estas limitaciones

metodológicas restringen la generalización de los resultados y la aplicación transcultural de los instrumentos.

En conjunto, los hallazgos indican que la evidencia psicométrica disponible presenta una calidad metodológica principalmente moderada, con fortalezas en la validez discriminante y concurrente, pero con áreas de mejora en la evaluación de la estructura interna y la estabilidad temporal de las medidas.

Discusión

En primer lugar, los resultados obtenidos a lo largo del estudio permiten confirmar plenamente que los objetivos planteados se cumplieron de manera rigurosa y acorde con el propósito inicial. En la sección introductoria se estableció que el propósito del estudio era analizar las propiedades psicométricas de trece instrumentos empleados en el Consultorio Psicoforense: MMPI-2, MMPI-2-RF, PAI, MCMI-IV, SIMS, NEUROPSI, BANFE, TECA, SCL-90-R, SENA, CUIDA, TAMAI y LSB-50. Los hallazgos presentados a lo largo del trabajo muestran que la mayoría de estas pruebas poseen adecuados niveles de validez de constructo, consistencia interna, confiabilidad y estructura factorial, lo cual confirma la hipótesis principal: estas herramientas cuentan con el soporte técnico necesario para su utilización en evaluaciones psicológicas forenses.

Asimismo, se confirma la hipótesis secundaria planteada en el marco teórico, que anticipaba la aparición de dificultades asociadas con la adaptación cultural, la falta de baremos normativos colombianos, la influencia del nivel educativo y las diferencias lingüísticas. Estas dificultades se evidencian repetidamente en secciones como las dedicadas al MMPI-2, PAI, MCMI-IV, TECA, SCL-90-R, SENA y TAMAI, donde se señala la necesidad de contextualizar los puntajes para garantizar una interpretación válida en población colombiana. En ese sentido, tal como plantean la APA (2013) y Nijdam-Jones et al. (2020), la hipótesis que sostenía que la validez de uso depende de procesos de adaptación cultural encuentra un respaldo empírico claro en los resultados discutidos en el documento. En conjunto, el trabajo confirma que los instrumentos analizados presentan propiedades psicométricas sólidas, pero deben emplearse con criterios técnicos y considerando sus limitaciones contextuales.

En segundo lugar, los hallazgos del estudio muestran una fuerte consistencia con la literatura científica disponible sobre las propiedades psicométricas de los trece instrumentos

analizados. En las secciones dedicadas al MMPI-2 y al MMPI-2-RF se presentan resultados que coinciden estrechamente con lo encontrado en estudios realizados en España (del Barrio et al., 2016), Perú (Scott & Mamani, 2000), Portugal (Afonso et al., 2024) y Puerto Rico (Garrido et al., 2012), donde se reportan estructuras factoriales estables, altos niveles de consistencia interna y capacidad discriminante entre poblaciones clínicas y no clínicas. En el caso del PAI, el trabajo muestra que sus niveles de validez convergente y consistencia interna son similares a los reportados en investigaciones internacionales, reforzando su utilidad en evaluaciones clínicas y forenses.

Por su parte, la sección dedicada al MCMI-IV resalta resultados compatibles con estudios norteamericanos y europeos que han encontrado buena validez estructural de las escalas de personalidad y síndromes clínicos, aunque con variaciones contextuales esperables en poblaciones no estadounidenses. El análisis del SIMS también coincide con la literatura que documenta su sensibilidad a la simulación y la posibilidad de generar falsos positivos en individuos con psicopatología severa. Asimismo, los apartados referentes al NEUROPSI y al BANFE muestran resultados congruentes con estudios latinoamericanos que respaldan su validez para la evaluación neuropsicológica de funciones cognitivas y ejecutivas en población hispanohablante. En el caso del TECA, se observan similitudes con estudios iberoamericanos que evidencian una estructura factorial coherente y adecuados niveles de consistencia interna. De igual manera, los resultados del SCL-90-R se alinean con la amplia literatura que confirma su validez para la evaluación del malestar psicológico general.

El SENA, orientado a población infantil y adolescente, presenta patrones similares a los señalados en estudios españoles que destacan su fiabilidad en la identificación de dificultades emocionales y comportamentales. Finalmente, los análisis realizados sobre el CUIDA, TAMAI y LSB-50 muestran que sus propiedades psicométricas coinciden con estudios ibéricos y latinoamericanos previos, particularmente en términos de validez convergente y consistencia interna. De esta manera, el estudio demuestra que sus hallazgos se ajustan a las tendencias descritas en investigaciones previas, fortaleciendo la confiabilidad del análisis.

En tercer lugar, el estudio permite identificar diversas amenazas a la validez interna y externa que deben ser tomadas en cuenta al interpretar las propiedades psicométricas de los trece instrumentos. En el apartado metodológico se evidencia que, aunque la mayoría de las pruebas cuenta con estudios de consistencia interna, pocos presentan análisis factorial confirmatorio,

elemento fundamental para validar la estructura teórica propuesta. Esto afecta la interpretación de instrumentos como el MCMI-IV, TECA, TAMAI y CUIDA, que requieren modelos factoriales robustos para garantizar que las dimensiones corresponden al constructo evaluado. Asimismo, se observa que muchos estudios carecen de validez de contenido evaluada mediante procedimientos formales con jueces expertos, lo cual dificulta establecer con precisión si los ítems representan adecuadamente los dominios conceptuales.

Además, se evidencia ausencia de estudios de estabilidad temporal (*test-retest*) en pruebas como SENA, LSB-50, TECA y CUIDA, aspecto crítico en escenarios forenses donde se requiere determinar si los cambios observados reflejan modificaciones reales en el estado psicológico o error de medición. Otro elemento interpretativo importante es la heterogeneidad de las muestras empleadas en los estudios revisados, que incluyen adultos, adolescentes, niños, población psiquiátrica, comunitaria y penitenciaria, lo cual introduce variabilidad no controlada y dificulta la comparación de resultados. También se destaca que varios estudios no reportan estrategias para controlar sesgos de respuesta como simulación, disimulación o deseabilidad social, fundamentales al interpretar pruebas como MMPI-2, MMPI-2-RF, PAI, MCMI-IV y SIMS, utilizadas con frecuencia en contextos judiciales.

Finalmente, en instrumentos neuropsicológicos como NEUROPSI y BANFE se identifican factores como nivel educativo, edad y experiencia sociocultural que influyen en el rendimiento cognitivo y deben ser considerados cuidadosamente durante la interpretación. En conjunto, estos elementos permiten concluir que, aunque los trece instrumentos presentan propiedades psicométricas favorables, su interpretación debe realizarse con precaución, considerando las limitaciones metodológicas identificadas.

En cuarto lugar, el análisis de los resultados permite evidenciar que existen restricciones significativas para generalizar el funcionamiento de los instrumentos al contexto colombiano. En los apartados de resultados se observa que la mayor parte de las validaciones proviene de España, México, Chile, Perú, Puerto Rico y Estados Unidos, lo que implica que los datos normativos y los modelos factoriales reflejan características culturales, educativas y lingüísticas distintas a las colombianas. La ausencia de estudios de invarianza factorial para MMPI-2, MMPI-2-RF, PAI, MCMI-IV, TECA, SCL-90-R, SENA y TAMAI dificulta confirmar que estos instrumentos midan los mismos constructos de la misma forma en distintas poblaciones.

Asimismo, la falta de baremos colombianos actualizados limita la precisión diagnóstica, dado que la interpretación de puntajes en pruebas como NEUROPSI, BANFE, SENA o TAMAI está altamente influenciada por nivel educativo, acceso a escolarización, vocabulario, estilos de crianza y experiencias socioculturales. En los análisis también se evidencia que no existen estudios sistemáticos sobre el desempeño de estos instrumentos en población forense colombiana, lo cual restringe la validez externa en áreas como imputabilidad penal, valoración del daño psicológico, custodia infantil, riesgo de reincidencia o idoneidad parental. Aunque instrumentos como MMPI-2, PAI, MCMI-IV y SCL-90-R cuentan con validaciones internacionales extensas, el estudio deja claro que su uso en Colombia debe ir acompañado de procesos de adaptación cultural, normalización local y análisis de validez específica para las poblaciones evaluadas.

Finalmente, el estudio permite identificar implicaciones importantes para el uso de los trece instrumentos en el ámbito forense, la formación profesional y el desarrollo investigativo. En términos prácticos, los resultados indican que, aunque las pruebas presentan propiedades psicométricas adecuadas, su uso debe realizarse con criterios técnicos rigurosos, evitando interpretaciones literales de puntajes obtenidos con baremos extranjeros. En escenarios judiciales, esto implica que los profesionales deben justificar la pertinencia del instrumento, advertir sus limitaciones y contextualizar sus resultados. En el ámbito formativo, los hallazgos subrayan la necesidad de fortalecer la formación en psicometría avanzada, adaptación transcultural y validación de instrumentos, así como de promover competencias para el análisis crítico de la evidencia. A nivel investigativo, los resultados resaltan la urgencia de realizar estudios factoriales confirmatorios, análisis de invarianza, validaciones convergentes y discriminantes, análisis test–retest, así como estudios de sensibilidad y especificidad en población colombiana para instrumentos como SENA, TAMAI, TECA, CUIDA y LSB-50. Finalmente, el estudio muestra la necesidad de que las instituciones judiciales, clínicas y académicas adopten protocolos estandarizados basados en evidencia científica, y fomenten redes de investigación que permitan consolidar una psicometría forense contextualizada y culturalmente pertinente para Colombia.

Referencias

- Abrisqueta-Gómez, J., Ostrosky-Solis, F., Bertolucci, P., & Bueno, O. (2008). Applicability of the abbreviated neuropsychologic battery (NEUROPSI) in Alzheimer disease patients. *Alzheimer Disease and Associated Disorders*, 22(1), 72–78.
<https://doi.org/10.1097/WAD.0b013e3181665397>
- Abuín, M., & de Rivera, L. (2014). La medición de síntomas psicológicos y psicosomáticos: el Listado de Síntomas Breve (LSB-50). *Clínica y Salud*, 25(2), 131–141.
<https://www.redalyc.org/articulo.oa?id=180631191007>
- AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing: National Council on Measurement in Education*. American Educational Research Association.
- Afonso, M., Novo, R., Camilo, C., & González, B. (2024). The bilingual study methodology in translating and adapting personality tests: Equivalence issues in the development of the MMPI-2-RF Portuguese version. *Psychological Test Adaptation and Development*, 5(1), 139–150. <https://doi.org/10.1027/2698-1866/a000075>
- American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct*. <http://www.apa.org/ethics/code/index.aspx>
- American Psychological Association. (2013). *Specialty guidelines for forensic psychology*. American Psychological Association. <https://www.apa.org/practice/guidelines/forensic-psychology>
- Ampudia Rueda, A., Sánchez Crespo, G., & Jiménez Gómez, F. (2018). La contribución del MMPI-2 a la predicción del riesgo de violencia. *Revista de Psicología*, 36 (2), 603–629.
<https://doi.org/10.18800/psico.201802.008>
- APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, 61(4), 271–285. <https://doi.org/10.1037/0003-066X.61.4.271>
- Arce, R., Fariña, F., & Novo, M. (2009). Evaluación psicológica forense: Fundamentos científicos y aplicaciones prácticas. *Papeles del Psicólogo*, 30(2), 139–146.
<https://www.researchgate.net/publication/271133230>
- Archer, R., Wheeler, E. & Vauter, R. (2010). Empirically supported forensic assessment. *Clinical Psychology: Science and Practice*, 17(3), 256–271.
<https://psycnet.apa.org/doi/10.1111/cpsp.12171>

- Arredondo, N., Mejía, C., Acosta, F., Rodríguez, O., & Ramírez, N. (2019). Criterios de idoneidad psicológica para padres en proceso de adopción en Colombia y su validez a través de ecuaciones estructurales. *Anuario de Psicología*, 49(2), 72–84.
<https://doi.org/10.1344/ANPSIC2019.49.8>
- Bartol, C., & Bartol, A. (2015). *Psychology and Law: Research and practice*. SAGE Publications.
- Ben-Porath, Y., & Tellegen, A. (2008). *MMPI-2-RF: Technical manual*. University of Minnesota Press.
<https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft15121-000>
- Blasco Saiz, J., & Pallardó-Durá, L. (2013). Detección de exageración de síntomas mediante el SIMS y el MMPI-2-RF en pacientes diagnosticados de trastorno mixto ansioso-depresivo y adaptativo en el contexto medicolegal: Un estudio preliminar. *Clínica y Salud*, 24(3), 177–183. <https://dx.doi.org/10.5093/cl2013a19>
- Capilla Ramírez, P., González Ordí, H., Santamaría, P., Pérez Nieto, M., & Casado Morales, M. (2013). Fibromialgia: ¿exageración o simulación?. *Clínica y Salud*, 24(3), 184–195.
<https://www.redalyc.org/articulo.oa?id=180628831007>
- Chavira Trujillo, G., & Celis de la Rosa, A. (2022). Propiedades psicométricas del Test de Empatía Cognitiva y Afectiva (TECA) en población mexicana. *Acta de Investigación Psicológica*, 12(1), 62–75. <https://www.redalyc.org/articulo.oa?id=358975435006>
- Córdoba-Viveros, S., Sevilla-Moreno, D., Peña-Rivera, R., Espinosa-Becerra, A., Quiroga-Baquero, L., Perilla-Gamboa, P., Vásquez-Prieto, P., & Loinaz, I. (2024). Caracterización de una muestra de residentes bogotanos que ejercen violencia filio-parental. En A. Espinosa-Becerra & L. Quiroga-Baquero (Comp.), *Tendencias e Investigaciones en Psicología Jurídica* (pp. 20–99). Editorial Ascofapsi.
<https://doi.org/10.61676/9786289532470.01>
- de la Iglesia, G., Liporace, M., & Solano, A. (2015). Psychometric Study of the Main Clinical Scales of the Listado de Síntomas Breve (LSB-50) [Short Checklist of Symptoms] in Argentinean Adolescents. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 22(3), 385–401. <https://doi.org/10.4473/TPM22.3.5>
- del Barrio, A., Jiménez-Huete, A., Toledano, R., García-Morales, I., & Gil-Nagel, A. (2016). Validez de las escalas clínicas y de contenido del Inventario Multifásico de Personalidad

- de Minnesota-2 para el diagnóstico de crisis no epilépticas psicógenas. *Neurología*, 31(2), 106–112. <https://doi.org/10.1016/j.nrleng.2013.12.022>
- Doldán-Peluffo, M., Hamuy Díaz de Bedoya, F., Nacimiento, B., Fleitas, C., Galeano, M., Ramírez, M., González, L., Martínez, M., Paredes-Ramírez, P., & Rolón, M. (2023). Estandarización de la batería neuropsicológica de funciones ejecutivas y lóbulos frontales BaNFE-2 en muestra paraguaya con esclerosis múltiple tipo recaída remisión. *Revista del Instituto de Medicina Tropical*, 18(1), 38–56. <https://doi.org/10.18004/imt/2023.18.1.6>
- Espinosa-Becerra, A., & Quiroga-Baquero, L. (2022). Diseño y validación de un modelo analítico-funcional de formulación de caso forense. En L. González (Ed.), *Perspectivas de investigación en Psicología: Impacto social en diversidad de contextos* (pp. 181-247). Ediciones USTA.
- Espinosa-Becerra, A., Quiroga-Baquero, L., & Jiménez-Molina, R. (2020). Investigación traslacional en psicología jurídica: propuestas, retos y perspectivas. *Journal of Behavior, Health & Social Issues*, 12(2), 1-12. <http://dx.doi.org/10.22201/fesi.20070780e.2020.12.2.76306>
- Fernández-Pinto, I., Santamaría, P., Sánchez-Sánchez, F., Carrasco, M., & Del Barrio, V. (2015). Sistema de evaluación de niños y adolescentes. *SENA. TEA Ediciones*.
- Flores-Lázaro, J., Ostrosky-Solis, F., & Lozano, A. (2014). *Batería Neuropsicológica de Funciones Ejecutivas y Lóbulos Frontales (BANFE-2)*. Manual Moderno.
- Freitas, J., & Miranda, G. (2021). Propriedades psicométricas do Teste de Empatia Cognitiva e Afetiva (TECA) em estudantes universitários brasileiros. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica*, 58(1), 57–66. <https://doi.org/10.21865/RIDEP58.1.05>
- Garrido Amadeo, M., Rosario-Hernández, E., Figueroa Ortiz, J., Medina Vélez, G., & Padovani Rivera, C. (2012). Comparación de las Escalas de Validez, Clínicas y Reestructuradas del MMPI-2 en una Muestra Puertorriqueña. *Revista Puertorriqueña de Psicología*, 23, 119-145. <https://www.redalyc.org/articulo.oa?id=233228917009>
- Gioia, G., Espy, K., & Isquith, P. (2016). BRIEF-P. Evaluación Conductual de la Función Ejecutiva-Versión Infantil (E. Bausela y T. Luque, adaptadoras). TEA Ediciones.
- González-Forteza, C., Solís Torres, C., Jiménez Tapia, A., Hernández Fernández, I., González-González, A., Juárez García, F., Medina-Mora, M. E., & Fernández-Varela Mejía, H.

- (2011). Confiabilidad y validez de la escala de depresión CES-D en un censo de estudiantes de nivel medio superior y superior, en la Ciudad de México. *Salud Mental*, 34(1), 53-59.
- Gorostiaga, A., Balluerka, N., & Soroa, G. (2014). Evaluación de la empatía en el ámbito educativo y su relación con la inteligencia emocional. *Revista de Educación*, 3(64), 12–38. <https://doi.org/10.4438/1988-592X-RE-2014-364-253>
- Guarnera, L., Murrie, D., & Boccaccini, M. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science*, 3(2), 143–152. <https://doi.org/10.1037/tps0000114>
- Hernández-Guanir, P. (2002). *Test Autoevaluativo Multifactorial de Adaptación Infantil (TAMAI). Manual 7 edición revisada*. TEA
- Ignatyev, Y., Fritsch, R., Priebe, S., & Mundt, A. (2016). Psychometric properties of the symptom check-list-90-R in prison inmates. *Psychiatry Research*, (23)9, 226–231. <https://doi.org/10.1016/j.psychres.2016.03.007>
- Jiménez-Ardila, L., Cañas-Serrano, J., Carvajal-Builes, J., Devia-Vega, E., Jiménez-Molina, J., López-Cantero, E., Murillo, C., Quintero-Cárdenas, C., & Quiroga-Baquero, L. (2021). *Subcampos de aplicación de la psicología jurídica*. Universidad Católica de Colombia. ISBN: 978-958-5133-50-1
- Jiménez-Gómez, F., Sánchez-Crespo, G., & Ampudia-Rueda, A. (2013). Is there a social desirability scale in the MMPI-2-RF? *Clínica y Salud*, 24(3), 161-168. <https://www.redalyc.org/articulo.oa?id=180628831004>
- Lally, S. (2003). What tests are acceptable for use in forensic evaluations? A survey of experts. *Professional Psychology: Research and Practice*, 34(5), 491–498. <https://doi.org/10.1037/0735-7028.34.5.491>
- Lemos, V., Vargas Rubilar, J., & López, M. (2022). Validación de una versión breve del Test de Empatía Cognitiva y Afectiva en población universitaria argentina. *Psykhē*, 31(2), 1–18. <https://doi.org/10.7764/psykhe.2020.21747>
- Lobo, A., Espinosa-Becerra, A., Guerrero, A. & Ospina, V. (2016). *Psicología Forense en el proceso penal con tendencia acusatoria. Guía práctica para psicólogos y abogados*. Ediciones USTA – Manual Moderno.

- Londoño, N., Agudelo, D., Martínez, E., Anguila, D., Aguirre, D., & Arias, J. (2018). Validación del cuestionario de 90 síntomas SCL-90-R de Derogatis en una muestra clínica colombiana. *MedUNAB*, 21(2), 45-59.
<https://www.redalyc.org/articulo.oa?id=71964815004>
- López-Pérez, B., Fernández-Pinto, I., & Abad, F. (2008). *TECA. Test de empatía cognitiva y afectiva*. TEA.
- López-Pérez, B., Ambrona, T., & Márquez-González, M. (2014). Adaptación y validación de un instrumento para la evaluación de la empatía en niños y adolescentes: TECA-NA. *Psicología Conductual*, 22(1), 5-18.
<https://www.behavioralpsycho.com/producto/adaptacion-y-validacion-de-un-instrumento-para-la-evaluacion-de-la-empatia-en-ninos-y-adolescentes-teca-na/>
- Lövgren, P., Laake, P., Reitan, S., & Narud, K. (2022). Use of assessment instruments in forensic evaluations of criminal responsibility in Norway. *BMC Psychiatry*, 22, 235.
<https://doi.org/10.1186/s12888-022-03831-4>
- Marín Torices, M., Hidalgo-Ruzzante, N., Daugherty, J., Jiménez-González, P., & Pérez García, M. (2017). Validation of neuropsychological consequences in victims of intimate partner violence in a Spanish population using specific effort tests. *The Journal of Forensic Psychiatry & Psychology*, 29(1), 86-98.
- Martín-Albo, J., Núñez, J., & Navarro, J. (2009). Validation of the Spanish version of the Situational Motivation Scale (EMSI) in the educational context. *The Spanish Journal of Psychology*, 12(2), 799-807. <https://doi.org/10.1017/s113874160000216x>
- Martinez, U., Fernandez del Rio, E., Lopez-Duran, A., & Becona, E. (2017). The utility of the MMPI-2-RF to predict the outcome of a smoking cessation treatment. *Personality and Individual Differences*, 10(6), 172-177. <https://doi.org/10.1016/j.paid.2016.11.019>
- Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., Bouter, L., & de Vet, H. (2010). The COSMIN checklist for assessing methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*, 19(4), 539-549. <https://doi.org/10.1007/s11136-010-9606-8>

- Moreno-Manso, J., García-Baamonde, M., Guerrero-Barona, E., & Blázquez-Alonso, M. (2010). Competencia pragmática y adaptación psicosocial en niños sujetos a medidas de protección infantil. *Salud Mental*, 33(4), 333-340.
<https://www.redalyc.org/articulo.oa?id=58216022005>
- Neal, T., Slobogin, C., Saks, M., Faigman, D., & Geisinger, K. (2020). Psychological assessments in legal contexts: Are courts keeping “junk science” out of the courtroom? *Psychological Science in the Public Interest*, 20(3), 135–164.
<https://doi.org/10.1177/1529100619888860>
- Nijdam-Jones, A., García-López, E., Aparcero, M., & Rosenfeld, B. (2020). How do Latin American professionals approach violence risk assessment? A qualitative exploratory study. *International Journal of Forensic Mental Health*, 19(4), 371–389.
<https://doi.org/10.1080/14999013.2019.1707330>
- Ortiz-Tallo, M., Ferragut, M., Cardenal, V., & Santamaría, P. (2017). Reliability of the Personality Assessment Inventory (PAI) in chilean contexts and differences between general and clinical population. *Revista Mexicana de Psicología*, 34(2), 110–124.
<http://www.redalyc.org/articulo.oa?id=243057743004>
- Ostrosky-Solís, F., Esther Gómez-Pérez, M., Matute, E., Rosselli, M., Ardila, A., & Pineda, D. (2007). NEUROPSI ATTENTION AND MEMORY: A Neuropsychological Test Battery in Spanish with Norms by Age and Educational Level. *Applied Neuropsychology*, 14(3), 156–170. <https://doi.org/10.1080/09084280701508655>
- Osuna, E., López-Martínez, M., Arce, R., & Vázquez, M. (2015). Analysis of response patterns on the MMPI-2 in psychiatric prison inmates. *International Journal of Clinical and Health Psychoiogy*. 15(1), 29–36. <https://doi.org/10.1016/j.ijchp.2014.09.002>
- Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Palmer, A., Borrás, C., Pérez-Pareja, J., Sesé, A., & Vilariño, M. (2013). Are patients with chronic pain and fibromyalgia correctly classified by MMPI-2 validity scales and indexes? *The European Journal of Psychology Applied to Legal Context*, 5(2), 123–129.
<http://dx.doi.org/10.5093/ejpalc2013a1>

- Paulino, M., Moniz, M., Moura, O., Rijo, D., Morey, L., & Simões, M. (2024). Psychometric properties of the Portuguese version of the Personality Assessment Inventory: normative data and reliability. *Frontiers in Psychology, 15*, 1359793. <https://doi.org/10.3389/fpsyg.2024.1359793>
- Penado Abilleira, M., Rodicio-García, M., Corrás Vázquez, T., Ríos de Deus, M., & Iglesias Cortizas, M. (2019). Personality characteristics of a sample of violent adolescents against their partners. *Psicologia, reflexao e critica : revista semestral do Departamento de Psicologia da UFRGS, 32*(1), 11. <https://doi.org/10.1186/s41155-019-0122-7>
- Picasso, H., Ostrosky-Solis, F., & Nicolini, H. (2005). Sensibilidad y especificidad de un instrumento neuropsicológico en la evaluación de subtipos de esquizofrenia: Un estudio con población hispanohablante [Sensitivity and specificity of a neuropsychological instrument in the evaluation of schizophrenia subtypes: A study with a Spanish speaking population]. *Actas Españolas de Psiquiatría, 33*(2), 87–95.
- Poythress, N., Nicholson, R., Otto, R., Edens, J., Bonnie, R., Monahan, J., & Hoge, S. (1999). *The MacArthur Competence Assessment Tool–Criminal Adjudication: Professional manual*. Psychological Assessment Resources.
- Puerta-Lopera, C., Gómez-Camacho, L., & Ceballos-Sánchez, J. (2021). Propiedades psicométricas preliminares de la BANFE en población universitaria colombiana. *Revista Neuropsicología Latinoamericana, 13*(1), 33-49.
- Puente-López, E., Pina, D., Dandachi-FitzGerald, B., Giromini, L., López-Nicolás, R., Nieto-Cañaveras, M., & Merten, T. (2025). Prevalence of symptom overreporting in the Structured Inventory of Malingered Symptomatology (SIMS) in clinical patients: A meta-analysis. *Psicothema, 37*(2), 74-85. <https://doi.org/10.70478/psicothema.2025.37.18>
- Puente-López, E., Pina, D., Ruiz-Hernández, J., & Llor-Esteban, B. (2020). Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in motor vehicle accident patients. *The Journal of Forensic Psychiatry & Psychology, 32*(1), 131–154. <https://doi.org/10.1080/14789949.2020.1833073>
- Puente-López, E., Pina, D., Shura, R., Boskovic, I., Martínez-Jarreta, B., & Merten, T. (2022). The Impact of Different Forms of Coaching on the Structured Inventory of Malingered Symptomatology (SIMS). *Psicothema, 34*(4), 528-536. <https://doi.org/10.7334/psicothema2022.129>

- Puerta-Lopera, I., Dussán-Luberti, C., Montoya-Londoño, D., Landínez-Martínez, D., & Pérez-Parra, L. (2022). Validación y estandarización de pruebas neuropsicológicas para la evaluación de funciones ejecutivas en población universitaria. *Cuadernos Hispanoamericanos de Psicología*, 22(1), 1–17.
<https://revistas.unbosque.edu.co/index.php/CHP/article/view/validacion-estandarizacion-funciones-ejecutivas>
- Querejeta, A., Farías Sarquís, Y., Moreno, M., Crostelli, A., Stecco, J., Venier, A., Godoy, J., & Pilatti, A. (2012). Test Neuropsi: Normas según edad y nivel de instrucción para Argentina. *Cuadernos de Neuropsicología*, 6(2), 48-58.
<https://www.redalyc.org/articulo.oa?id=439643141005>
- Redondo, L., Fariña, F., Seijo, D., Novo, M., & Arce, R. (2019). A meta-analytical review of the responses in the MMPI-2/MMPI-2-RF clinical and restructured scales of parents in child custody dispute. *Annals of Psychology*, 35(1), 156–165.
<http://dx.doi.org/10.6018/analesps.35.1.338381>
- Rodríguez-Aguirre, C., García-Flores, R., Sotelo-Castillo, M., & Meza-Peña, C. (2023). Factores psicosociales y sociodemográficos asociados con el desgaste profesional en médicos residentes de Sonora (México). *Revista Ciencias de la Salud*, 21(3), 1-17.
<https://www.redalyc.org/articulo.oa?id=56278661008>
- Rogers, R. (2008). Rogers's criminal responsibility assessment scales (r-cras). En B. Cutler (Ed.) *Encyclopedia of psychology and law* (pp. 704-704). SAGE Publications, Inc.
<https://doi.org/10.4135/9781412959537.n272>
- Rossignoli-Palomeque, T., & Pérez-Hernández, E. (2022). Validity of the System of Evaluation of Children and Adolescents to measure executive functions: Convergent validity. *Journal of Psychoeducational Assessment*, 40(7), 853–864.
<https://doi.org/10.1177/07342829221104349>
- Sanz, J., Navarro, R., Fausor, R., Altungy, P., Gesteira, C., Morán, N., & Garcia, M. (2018). La escala de discapacidad social de Marlowe-Crowne como instrumento para la medida de la discapacidad social, la sinceridad y otros constructos relacionados en psicología legal y forense. *Psicopatología Clínica, Legal y Forense*, 18(1), 112- 133.
<https://hdl.handle.net/20.500.14352/18918>

- Sánchez, G., Ampudia, A., Jiménez, F., & Amado, B. (2017). Contraste de la eficacia de las escalas de validez del MMPI-2-RF en la detección de la simulación. *Revista Europea de Psicología Aplicada al Contexto Jurídico*, 9(2), 51 -56.
<https://doi.org/10.1016/j.ejpal.2017.03.002>
- Sánchez, O., & Ledesma, R. (2009). Análisis Psicométrico del Inventario de Síntomas Revisado (SCL-90-R) en Población Clínica. *Revista Argentina de Clínica Psicológica*, 18(3), 265-274. <https://www.redalyc.org/articulo.oa?id=281921775007>
- Sandín, B., Valiente, R., Chorot, P., Santed, M., & Lostao, L. (2008). SA-45: Forma abreviada del SCL-90. *Psicothema*, 20(2), 290–296.
<https://www.researchgate.net/publication/5439721>
- Scott, R., & Pampa, W. (2000). The MMPI-2 in Perú: a normative study. *Journal of Personality Assessment*, 74(1), 95–105. <https://doi.org/10.1207/S15327752JPA740107>
- Stover, J., Freiberg- Hoffmann, A., Fernández- Liporace, M., & Castro - Solano, A. (2016). Consistencia interna y dimensionalidad del Inventario de Evaluación de la Personalidad en estudiantes universitarios de Buenos Aires. *Acta Psiquiátrica y Psicológica de América Latina*, 62(3), 153–163.
https://ri.conicet.gov.ar/bitstream/handle/11336/46805/CONICET_Digital_Nro.f6246c20-0afd-47eb-815d-e64b83b099d6_A.pdf?sequence=2&isAllowed=n
- The Campbell Collaboration. (2019). *Campbell systematic reviews: policies and guidelines*. The Campbell Collaboration. <https://doi.org/10.4073/cpg.2016.1>
- Tricco, A., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Levac, D., & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Widows, M., & Smith, G. (2005). SIMS: *Structured Inventory of Malingered Symptomatology*. Professional manual. PAR - Psychological Assessment Resources.