

Instrumentos Experimentales para medir los Sesgos Cognitivos en la Toma de Decisiones Financieras: una Revisión Sistemática y Evaluación Metodológica

Héctor Madariaga Tapia
Autor de correspondencia
hmadariaga@pucp.edu.pe
Pontificia Universidad Católica del Perú
<https://orcid.org/0009-0002-9215-3778>

Marta Tostes Vieira
mtostes@pucp.edu.pe
Pontificia Universidad Católica del Perú
<https://orcid.org/0000-0003-0855-7463>

Resumen

El presente estudio tiene como objetivo revisar y evaluar sistemáticamente los instrumentos experimentales utilizados para medir los sesgos cognitivos en la toma de decisiones financieras, examinando sus características metodológicas y sus implicaciones para la comparabilidad y transparencia de la evidencia conductual. Siguiendo las directrices PRISMA 2020, se realizó una revisión sistemática que abarcó el período comprendido entre 1990 y 2025 y dio lugar a un corpus final de 70 estudios. El análisis empleó un marco de tres componentes: extracción descriptiva (C1), evaluación de la robustez metodológica (D1) y auditoría del reporte procedimental (E1). La concordancia entre evaluadores se determinó mediante el coeficiente Kappa de Cohen. Los análisis de asociación y temporales incorporaron el ajuste de Holm para comparaciones múltiples.

Las tareas experimentales y las simulaciones representaron el 94,3 % de los estudios. Las dimensiones metodológicas presentaron perfiles heterogéneos: varios criterios recibieron predominantemente clasificaciones Altas, aunque cinco mostraron efectos techo severos, mientras que la evidencia de confiabilidad documentada explícitamente continuó siendo limitada y el reporte de prácticas procedimentales varió entre los estudios. El análisis de sensibilidad indicó que el perfil metodológico mantuvo su heterogeneidad después de excluir de la interpretación integradora los cinco criterios afectados por efectos techo severos.

Ninguno de los ocho contrastes proporcionó evidencia estadística suficiente después del ajuste de Holm, aunque algunos revelaron patrones descriptivos relevantes y magnitudes de asociación de interés. La revisión se basa en la información reportada en los estudios primarios y en los

criterios de búsqueda y elegibilidad aplicados. El marco C1–D1–E1 fue desarrollado específicamente para esta revisión y no ha sido sometido a validación externa.

Una documentación más sistemática de las propiedades de medición y de las prácticas procedimentales podría fortalecer la transparencia, comparabilidad y reproducibilidad de la investigación experimental en finanzas conductuales. El estudio desplaza la atención desde cuáles sesgos cognitivos examinan los investigadores hacia cómo miden experimentalmente dichos sesgos, integrando la extracción descriptiva, la evaluación metodológica y la auditoría de los procedimientos. Asimismo, proporciona una base empírica para una propuesta conceptual que abarca la transparencia tecnológica, de medición y estadística, y ética.

Palabras clave: Sesgos cognitivos; Finanzas conductuales; Métodos experimentales; Instrumentos de medición; Revisión sistemática; Transparencia metodológica.

GECONTEC Vol. 14(2). 2026
ISSN 2255-5684
gecontec.org

© Creative Commons Attribution 4.0
Recibido 24 June, 2026
Aceptado 27 Septiembre, 2026
Publicado 2 Octubre, 2026

Introducción

El estudio de los sesgos cognitivos en la toma de decisiones financieras se ha expandido considerablemente durante las últimas décadas, convirtiéndose en un componente central de la economía conductual y experimental. Un amplio cuerpo de literatura empírica y teórica ha documentado desviaciones sistemáticas respecto de la elección racional, identificando sesgos y heurísticas como el exceso de confianza, el anclaje, la aversión a las pérdidas, el efecto de comportamiento gregario, la disponibilidad y la representatividad en diversos contextos financieros (Barberis & Thaler, 2003; Kahneman & Tversky, 1979; Thaler, 2016; Tversky & Kahneman, 1974). Estos hallazgos han contribuido significativamente a comprender cómo los individuos toman decisiones bajo condiciones de riesgo e incertidumbre.

A pesar de estos avances, una cuestión metodológica ha recibido comparativamente menor atención: cómo los investigadores miden experimentalmente los sesgos cognitivos. En la práctica, los investigadores emplean diversas tareas, simulaciones y estructuras de decisión para elicitación respuestas conductuales, desde elecciones entre loterías hasta asignaciones de carteras, pronósticos y mercados simulados (Biais et al., 2005; Glaser & Weber, 2007; Holt & Laury, 2002). Esta diversidad implica algo más que diferencias de formato. La evidencia metodológica indica que distintos procedimientos de elicitación pueden producir medidas que no necesariamente demuestran equivalencia, incluso cuando buscan aproximarse a constructos relacionados (Crosetto & Filippin, 2016; Frey et al., 2017; Pedroni et al., 2017). Asimismo, un mismo sesgo puede abarcar manifestaciones conceptualmente diferenciadas, como demuestra el caso del exceso de confianza (Moore & Healy, 2008).

Esta heterogeneidad tiene importantes implicaciones para la validez de constructo y la comparabilidad de la evidencia. La correspondencia entre un constructo teórico y su operacionalización depende de si el procedimiento empleado representa adecuadamente el fenómeno que busca medir (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995). En

la investigación experimental, las diferencias en las tareas, los incentivos, los contextos y los procedimientos pueden introducir una variación sustancial en la forma en que los investigadores elicitán y observan las respuestas conductuales (Camerer & Hogarth, 1999; Harrison & List, 2004; Hertwig & Ortmann, 2001). Estas cuestiones adquieren especial relevancia para la acumulación y replicación de la evidencia, ya que las comparaciones entre estudios requieren comprender no solo sus resultados, sino también las condiciones metodológicas bajo las cuales se obtuvieron dichos resultados (Camerer et al., 2016, 2018; Christensen & Miguel, 2018).

Las revisiones sistemáticas existentes en el ámbito de las finanzas conductuales se han centrado principalmente en identificar y clasificar los sesgos que afectan las decisiones de inversión, así como en sintetizar sus efectos y sus relaciones con las características de los inversores (Badola et al., 2024; Kumar & Goyal, 2015; Zahera & Bansal, 2018). Paralelamente, estudios metodológicos han demostrado que los distintos procedimientos utilizados para medir constructos conductuales pueden diferir en sus propiedades de medición y no necesariamente producir resultados equivalentes (Lönngqvist et al., 2015; Pedroni et al., 2017). Más recientemente, De Oliveira Cardoso et al. (2024) ampliaron esta discusión mediante una revisión sistemática centrada en la evidencia psicométrica de los instrumentos utilizados para medir sesgos conductuales entre inversores individuales. Sin embargo, la evaluación sistemática específicamente orientada a las arquitecturas experimentales mediante las cuales los investigadores operacionalizan conductualmente los sesgos cognitivos continúa siendo relativamente limitada, particularmente cuando dicha evaluación examina conjuntamente sus características instrumentales, metodológicas y procedimentales. Esta brecha justifica desplazar el foco desde cuáles sesgos identifican los investigadores y qué efectos producen hacia cómo los miden experimentalmente y bajo qué condiciones metodológicas generan y documentan dicha evidencia.

Desde una perspectiva metodológica, la literatura proporciona fundamentos para examinar dos tipos de patrones en esta revisión. En primer lugar, debido a que los diseños experimentales difieren en sus niveles de control, contextualización y realismo, resulta pertinente examinar si las arquitecturas experimentales presentan patrones diferenciados según las características de los participantes y los contextos en los que los investigadores las aplican (Harrison & List, 2004; Hertwig & Ortmann, 2001). En segundo lugar, los cambios en las posibilidades tecnológicas para la experimentación y la creciente atención a la transparencia y la reproducibilidad justifican examinar cómo han evolucionado a lo largo del tiempo determinadas prácticas metodológicas y de reporte de información (Camerer et al., 2016, 2018; Christensen & Miguel, 2018). Estas consideraciones cumplen una función conceptual y orientadora: no constituyen hipótesis confirmatorias ni presuponen la existencia, dirección o significancia estadística de las asociaciones examinadas.

En este contexto, el presente estudio tiene como objetivo identificar, clasificar y evaluar metodológicamente los instrumentos experimentales utilizados para medir los sesgos cognitivos en la toma de decisiones financieras individuales. En lugar de centrarse principalmente en cuáles sesgos estudian los investigadores o en estimar sus efectos agregados, la revisión examina cómo los investigadores operacionalizan estos constructos mediante diferentes arquitecturas experimentales y qué evidencia metodológica y procedimental puede evaluarse a partir de la información publicada. Para ello, el estudio emplea un marco analítico de tres componentes que distingue la extracción de información descriptiva (C1), la evaluación de dimensiones específicas

de robustez metodológica (D1) y una auditoría del reporte de prácticas procedimentales (E1). Este marco mantiene separada la información explícitamente reportada por los estudios de la evaluación metodológica realizada mediante criterios predefinidos y no considera el marco como una escala global de calidad.

A partir de este propósito, la revisión aborda tres preguntas de investigación. En primer lugar, ¿cómo caracterizan y operacionalizan los investigadores los instrumentos experimentales utilizados para medir los sesgos cognitivos en la toma de decisiones financieras? En segundo lugar, ¿qué patrones presentan estos instrumentos en términos de sus características metodológicas, evidencia de medición y reporte de prácticas procedimentales? En tercer lugar, ¿qué asociaciones surgen entre características específicas de los instrumentos, características de los participantes y contextos experimentales, y cómo varían determinadas prácticas metodológicas y de reporte de información a lo largo del tiempo? Estas preguntas integran la caracterización descriptiva con análisis específicos de asociación y temporalidad, distinguiendo los patrones observados de la evidencia inferencial disponible para sustentarlos.

El estudio realiza tres contribuciones principales. En primer lugar, proporciona una caracterización estructurada y comparativa de las arquitecturas experimentales utilizadas para operacionalizar los sesgos cognitivos en contextos financieros. En segundo lugar, integra la extracción descriptiva, la evaluación metodológica y la auditoría procedimental mediante el marco C1–D1–E1, distinguiendo así las características observables del diseño de aquellas dimensiones cuya evaluación depende de evidencia explícitamente documentada. En tercer lugar, examina las implicaciones de la heterogeneidad de los instrumentos para la robustez metodológica, la transparencia y la comparabilidad de la evidencia experimental. De este modo, la revisión contribuye a una comprensión metodológica de la medición experimental de los sesgos financieros y proporciona una base empírica para identificar ámbitos en los que una documentación más explícita podría fortalecer la evaluación, reproducibilidad y acumulación de evidencia en este campo.

Methodología

Diseño de la investigación

Este estudio realiza una revisión sistemática con una orientación metodológica dirigida a identificar, clasificar y evaluar los instrumentos experimentales utilizados para medir los sesgos cognitivos en la toma de decisiones financieras individuales. A diferencia de las revisiones diseñadas para estimar efectos conductuales agregados, el análisis se centra en cómo se operacionalizan los sesgos cognitivos mediante diferentes diseños e instrumentos experimentales y en la robustez metodológica que puede evaluarse a partir de la información reportada en los estudios primarios.

La revisión abarca ocho sesgos cognitivos reconocidos como altamente relevantes en las finanzas conductuales: exceso de confianza, anclaje, efecto de comportamiento gregario, aversión a las pérdidas, disponibilidad, efecto disposición, contabilidad mental y representatividad. Estos sesgos fueron seleccionados en función de su relevancia teórica y su presencia recurrente en la literatura sobre toma de decisiones financieras (Barberis & Thaler, 2003; Kahneman & Tversky, 1979; Thaler, 2016).

El proceso de revisión siguió las directrices PRISMA 2020 (Page et al., 2021), con el objetivo de promover la transparencia y trazabilidad durante la identificación, selección y evaluación de los estudios. La evaluación metodológica tomó como referencias complementarias los principios del Joanna Briggs Institute (JBI) (Aromataris et al., 2024) y del Critical Appraisal Skills Programme (CASP) (Critical Appraisal Skills Programme, 2018). Dado que estos marcos no fueron diseñados específicamente para evaluar instrumentos experimentales destinados a medir sesgos cognitivos en la toma de decisiones financieras, esta revisión desarrolló un marco específico mediante la integración y operacionalización de principios procedentes de la metodología de revisiones sistemáticas y la evaluación crítica, el diseño experimental y la inferencia causal, la teoría de la medición, la economía experimental, la reproducibilidad, la validez externa y la ética de la investigación.

El marco comprendió tres componentes con funciones diferenciadas pero complementarias: C1, destinado a la extracción estructurada de datos descriptivos; D1, orientado a evaluar dimensiones específicas de la robustez metodológica; y E1, destinado a auditar el reporte de determinadas prácticas procedimentales. En consecuencia, estos componentes no deben interpretarse como escalas de calidad equivalentes. La justificación y trazabilidad de los criterios, junto con sus reglas de aplicación y ejemplos, se presentan en el Material Suplementario. Los revisores perfeccionaron el marco mediante discusión y una fase preliminar de calibración; sin embargo, no se realizó una validación externa independiente.

Dada la heterogeneidad de los diseños experimentales, las tareas y los enfoques de medición identificados, no se consideró apropiado realizar un metaanálisis cuantitativo de los efectos. En su lugar, la síntesis siguió los principios de Synthesis Without Meta-analysis (SWiM) (Campbell et al., 2020), combinando distribuciones de frecuencia, tablas de contingencia y análisis inferenciales específicos de dimensiones metodológicas predefinidas. Este enfoque permitió realizar comparaciones sistemáticas entre diferentes arquitecturas experimentales sin asumir equivalencia entre instrumentos o paradigmas de medición heterogéneos.

Los criterios de elegibilidad se establecieron mediante un marco PICOS adaptado a la investigación experimental, buscando garantizar la correspondencia entre el objetivo de la revisión y las características de la evidencia incluida. La elegibilidad no se restringió a los experimentos de laboratorio ni a los diseños que involucraban manipulación experimental convencional; la revisión también admitió diseños cuasiexperimentales y experimentos naturales cuando cumplieran los criterios predefinidos y permitían identificar conductualmente al menos uno de los sesgos cognitivos considerados mediante decisiones financieras individuales.

Estrategia de búsqueda

La estrategia de búsqueda siguió un enfoque sistemático y explícito mediante la combinación de tres bloques conceptuales a través de operadores booleanos:

Sesgos cognitivos: términos correspondientes a los ocho sesgos predefinidos.

Contexto de toma de decisiones: términos relacionados con decisiones financieras y de inversión.

Diseño metodológico: términos correspondientes a enfoques experimentales y cuasiexperimentales.

Con el propósito de centrar la recuperación en estudios que involucraran medición conductual, la estrategia original utilizó el operador NOT para excluir los registros recuperados exclusivamente mediante términos asociados con encuestas, cuestionarios o autoinformes. Debido al riesgo potencial de excluir estudios que combinaran estas técnicas con tareas conductuales, posteriormente se realizó un análisis de sensibilidad eliminando dicho operador. Esta búsqueda recuperó nueve registros adicionales en Web of Science y once en Scopus. Sin embargo, estos 20 registros adicionales no fueron sometidos a una evaluación independiente y documentada de elegibilidad; por lo tanto, no es posible determinar cuántos habrían cumplido los criterios de inclusión ni si su evaluación habría modificado la composición del corpus final. En consecuencia, la estrategia original se mantuvo como base de la revisión, pero el análisis de sensibilidad introduce una incertidumbre residual respecto de la cobertura del proceso de recuperación bibliográfica, aspecto que debe considerarse al interpretar el alcance del corpus.

La búsqueda se realizó en dos bases de datos: Web of Science, restringida al Social Sciences Citation Index (SSCI), y Scopus. La revisión incluyó artículos de revistas sometidos a revisión por pares, publicados en inglés entre 1990 y 2025. Las cadenas de búsqueda booleanas completas y sus adaptaciones específicas para cada base de datos se presentan en el Material Suplementario, Apéndice A.

Los registros recuperados fueron exportados inicialmente a Zotero (v. 6.0) para la identificación de duplicados y posteriormente gestionados en Covidence durante todo el proceso de cribado. La búsqueda electrónica se complementó mediante la revisión retrospectiva de las referencias bibliográficas de los estudios relevantes. La búsqueda se realizó el 3 de enero de 2026.

Criterios de elegibilidad y selección de estudios

El proceso de selección de los estudios siguió un procedimiento secuencial de dos etapas. En la primera etapa, se realizó el cribado de títulos y resúmenes de acuerdo con los criterios de inclusión establecidos mediante el marco PICOS adaptado (Material Suplementario, Apéndice C, Tabla A1). En la segunda etapa, los artículos a texto completo fueron evaluados mediante la aplicación de criterios de exclusión relacionados con el diseño experimental, el contexto de la toma de decisiones financieras individuales y la medición conductual de los sesgos (Material Suplementario, Apéndice C, Tabla B1). Estos criterios permitieron incluir diseños experimentales y cuasiexperimentales, incluidos experimentos naturales cuando cumplían las condiciones de elegibilidad predefinidas.

Se identificaron un total de 319 registros: 279 mediante búsquedas en bases de datos (141 procedentes de Web of Science y 138 de Scopus) y 40 mediante búsqueda de citas. Después de eliminar los duplicados, realizar el cribado y evaluar la elegibilidad, una conciliación final de las bases de datos permitió identificar y eliminar un registro duplicado adicional, lo que dio lugar a un corpus final de 70 estudios. El proceso completo de identificación, cribado, evaluación de elegibilidad e inclusión se presenta en el diagrama de flujo PRISMA (Material Suplementario, Apéndice B).

Para evaluar la consistencia del proceso de selección, dos revisores tomaron de manera independiente las decisiones de elegibilidad. La concordancia entre evaluadores durante la

evaluación de los textos completos se determinó mediante el coeficiente Kappa de Cohen (Cohen, 1960), obteniéndose $\kappa = 0,83$, valor que corresponde a un nivel de concordancia casi perfecto según los criterios convencionales propuestos por Landis y Koch (1977). Posteriormente, los revisores resolvieron las discrepancias mediante discusión y consenso.

Extracción de datos

La extracción de datos siguió un esquema de codificación estructurado desarrollado ex ante (Material Suplementario, Apéndice C, Tabla C1), diseñado para registrar las características descriptivas, instrumentales y metodológicas de cada estudio. El proceso se implementó en Covidence mediante la codificación independiente realizada por dos revisores, seguida de la resolución de discrepancias.

Para estandarizar las decisiones de clasificación, los revisores utilizaron una guía de codificación detallada (Material Suplementario, Apéndice E) que especificaba el propósito, las categorías de respuesta, las reglas de decisión y ejemplos ilustrativos para cada variable C1.

Posteriormente, la información extraída se exportó a Microsoft Excel para construir la base de datos analítica, las distribuciones de frecuencia y las tablas de contingencia. Las variables de extracción se organizaron en tres grupos: características generales de los estudios, incluyendo año de publicación, región, muestra y contexto experimental; características instrumentales y técnicas, incluyendo el sesgo medido, el tipo de instrumento, el tipo de tarea y la forma de operacionalización; y características metodológicas básicas, incluyendo las condiciones de comparación, la estructura de incentivos y el nivel de detalle metodológico. Los tipos de instrumentos comprendieron tareas experimentales, juegos económicos, simulaciones y diseños de valor inducido, estos últimos basados en la tradición del valor inducido de la economía experimental (Smith, 1976). La estructura correspondiente y las reglas de codificación se documentan en el Material Suplementario.

Esta fase cumplió una función estrictamente descriptiva. La codificación se basó exclusivamente en la información explícitamente reportada en los estudios primarios, sin inferir prácticas, propiedades o procedimientos que no hubieran sido documentados. Cuando la información requerida para una variable no estaba disponible, se asignó la categoría «No reportado». Esta regla permitió distinguir la extracción factual de la evaluación metodológica posterior.

La consistencia de la extracción independiente se evaluó mediante el coeficiente Kappa de Cohen no ponderado, obteniéndose $\kappa = 0,76$, lo que corresponde a un nivel de concordancia sustancial (Cohen, 1960; Landis & Koch, 1977). Las reglas operativas y los criterios especificados en la guía de codificación permitieron estandarizar la aplicación de las categorías y reducir la ambigüedad durante el proceso de extracción.

La base de datos resultante permitió realizar comparaciones sistemáticas entre las diferentes arquitecturas experimentales y facilitó la identificación de patrones en la forma en que los sesgos cognitivos fueron operacionalizados y medidos.

Evaluación de la robustez metodológica y auditoría del reporte procedimental

La evaluación siguió un marco de tres componentes diseñado para distinguir la extracción descriptiva de la evaluación metodológica:

C1: extracción de datos descriptivos.

D1: evaluación de dimensiones específicas de la robustez metodológica.

E1: auditoría del reporte de prácticas procedimentales.

C1 registró las características explícitamente reportadas por los estudios sin asignar una valoración de calidad. D1 evaluó dimensiones específicas de la robustez metodológica de los instrumentos experimentales mediante las categorías ordinales Alta, Moderada y Baja. Los criterios abarcaron la claridad de los objetivos, la adecuación causal del diseño, el control estructural, la manipulación experimental, la precisión conceptual del instrumento, la replicabilidad, la evidencia de validez, la evidencia de confiabilidad, los incentivos, la correspondencia entre la muestra y la población objetivo, la discusión de la representatividad y el reporte de limitaciones (Material Suplementario, Apéndice C, Tabla D1). Cada criterio fue evaluado de manera independiente; por consiguiente, D1 no se utilizó para construir una puntuación global de robustez, ni la evaluación asumió que una debilidad en una dimensión determinara automáticamente la valoración asignada a cualquier otra dimensión.

E1 complementó esta evaluación mediante la auditoría del reporte explícito de tres prácticas procedimentales: pilotaje, soporte tecnológico y protocolo ético y anonimato (Material Suplementario, Apéndice C, Tabla E1). Los revisores codificaron estos indicadores de manera independiente y no los interpretaron como una escala agregada de calidad. En particular, el soporte tecnológico representó una característica documentable de la implementación experimental y no, por sí mismo, evidencia de una mayor calidad metodológica.

El análisis mantuvo una distinción fundamental entre extracción y evaluación. En C1, la ausencia de información explícita se registró como «No reportado». En D1, por el contrario, la categoría Baja indica que la evidencia disponible en el artículo no proporcionaba suficiente respaldo para asignar una valoración superior al criterio evaluado. Por lo tanto, una clasificación Baja no debe interpretarse automáticamente como evidencia de que una práctica o propiedad metodológica estuvo ausente durante la ejecución original del estudio. Esta distinción permitió evaluar la robustez metodológica sustentada por la información publicada sin equiparar la ausencia de evidencia con evidencia de ausencia.

Los revisores establecieron ex ante las reglas de decisión para cada criterio y las aplicaron mediante guías de codificación que especificaban el propósito de cada criterio, las categorías de respuesta y las condiciones para su aplicación. La fundamentación metodológica y la trazabilidad de los criterios C1–D1–E1 se presentan en el Material Suplementario, Apéndice D, mientras que las guías de codificación completas para C1, D1 y E1, incluidas las reglas de decisión y los ejemplos ilustrativos, se presentan en los Apéndices E, F y G, respectivamente. En particular, la evaluación de los incentivos en D1-28 consideró su relevancia y vinculación con las decisiones experimentales, en consonancia con la importancia metodológica de los incentivos en la

economía experimental (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith, 1976), sin interpretar la mera presencia de incentivos monetarios como un indicador global de calidad.

Antes de implementar plenamente el marco, los revisores realizaron una calibración interna utilizando cinco estudios seleccionados aleatoriamente con el propósito de examinar la interpretación de los criterios y armonizar su aplicación. Este procedimiento constituyó una etapa de calibración del proceso de codificación, no una validación externa del marco.

Dos revisores realizaron la evaluación de manera independiente. La consistencia entre evaluadores se estimó mediante el coeficiente Kappa de Cohen no ponderado, obteniéndose $\kappa = 0,76$ para D1 y $\kappa = 0,80$ para E1 (Cohen, 1960), valores que corresponden a niveles de concordancia sustancial según los criterios convencionales propuestos por Landis y Koch (1977). Estos coeficientes proporcionan una medida global de la consistencia en la aplicación del marco entre los revisores y no constituyen evidencia de validación externa del instrumento de evaluación. En el caso de D1, dada la naturaleza ordinal de las categorías, el Kappa no ponderado tampoco tiene en cuenta la magnitud de las discrepancias entre categorías. Posteriormente, los revisores resolvieron las discrepancias identificadas mediante discusión y consenso.

Integración y análisis de la evidencia

La estrategia analítica integró los tres componentes complementarios: C1, D1 y E1, manteniendo al mismo tiempo una distinción clara entre la información descriptiva y la evaluación metodológica. En primer lugar, se calcularon distribuciones de frecuencia y tablas de contingencia para caracterizar los instrumentos experimentales, los perfiles de los participantes, los contextos de aplicación, los sesgos examinados y las dimensiones metodológicas evaluadas.

Sobre esta base, se definieron cuatro análisis de asociación específicos (A1–A4): tipo de instrumento experimental $\times$ perfil de los participantes (A1); región $\times$ sesgo cognitivo principal (A2); contexto experimental $\times$ discusión de la representatividad (A3); y tipo de instrumento experimental $\times$ correspondencia entre la muestra y la población objetivo (A4). Estos análisis utilizaron el corpus final de 70 estudios, excepto A2, cuyo universo analítico comprendió 65 estudios después de excluir dos estudios clasificados como transregionales y tres estudios sin información regional, de acuerdo con las reglas predefinidas. Las especificaciones y categorías derivadas utilizadas en estos análisis se presentan en el Material Suplementario.

También se definieron cuatro análisis temporales (T1–T4) para examinar la distribución de determinadas prácticas metodológicas y de reporte a lo largo del período de estudio: período de publicación $\times$ pilotaje (T1); período $\times$ evidencia de confiabilidad (T2); período $\times$ protocolo ético y anonimato (T3); y período $\times$ soporte tecnológico (T4). Para estos análisis, los estudios se agruparon en cuatro intervalos temporales predefinidos. Las variaciones porcentuales entre períodos se interpretaron conjuntamente con los resultados inferenciales y las magnitudes de asociación, en lugar de considerarlas de manera independiente como evidencia de una tendencia temporal.

Para las tablas de contingencia, se utilizó la prueba χ^2 de Pearson cuando sus supuestos se cumplían adecuadamente. Cuando las frecuencias esperadas resultaron insuficientes en tablas $R \times C$, se aplicó la prueba exacta de Fisher–Freeman–Halton. La magnitud de la asociación se

estimó mediante V de Cramér. Los valores p y las magnitudes de asociación se interpretaron conjuntamente, evitando reducir la evidencia a una clasificación dicotómica basada exclusivamente en el umbral de significancia estadística.

Los ocho contrastes inferenciales, A1–A4 y T1–T4, constituyeron una única familia analítica primaria, y sus valores p se ajustaron conjuntamente mediante el procedimiento secuencial de Holm, con un α bilateral = .05. Adicionalmente, un análisis de sensibilidad aplicó el mismo procedimiento de Holm por separado a dos familias conceptuales: A1–A4 y T1–T4. Esta especificación permitió evaluar la estabilidad de las conclusiones bajo una definición alternativa de la familia de contrastes. Los valores p no ajustados se conservaron por transparencia, mientras que las conclusiones inferenciales principales se basaron en los valores p ajustados.

Dada la elevada concentración de clasificaciones Altas observada en algunos criterios D1, se realizó un análisis de sensibilidad descriptivo adicional para examinar su influencia en la interpretación integradora de la robustez metodológica. Una proporción superior al 90 % de clasificaciones Altas se definió descriptivamente como un efecto techo severo. Los criterios que superaron este umbral se mantuvieron sin modificaciones en D1, en la base de datos codificada y en los resultados descriptivos originales, pero el análisis no los utilizó como evidencia acumulativa en la interpretación integradora de la robustez metodológica. No se construyó ninguna puntuación global ni se realizó recodificación alguna. La especificación y los resultados de este análisis se presentan en el Material Suplementario.

Los análisis inferenciales se realizaron en R versión 4.6.1, utilizando únicamente funciones de R base. Las pruebas exactas de Fisher–Freeman–Halton se calcularon mediante `fisher.test()` sin simulación de Monte Carlo; las pruebas χ^2 de Pearson mediante `chisq.test()`; y el ajuste por multiplicidad mediante `p.adjust(..., method = "holm")`. La V de Cramér se calculó a partir del estadístico χ^2 y de las dimensiones de la tabla de contingencia. El script permitió reproducir las transformaciones analíticas, las tablas de contingencia, los contrastes estadísticos y las magnitudes de asociación a partir de la base de datos final. El código, la guía de reproducibilidad y la información sobre el entorno de ejecución se proporcionan en el Material Suplementario.

Resultados

Evolución de los contextos experimentales

Los 70 estudios incluidos fueron publicados entre 1996 y 2025, dentro del período de búsqueda predefinido de 1990–2025. Su distribución temporal muestra una mayor concentración de estudios en los períodos más recientes: 12 correspondieron a 1996–2005, 17 a 2006–2014, 19 a 2015–2020 y 22 a 2021–2025.

El laboratorio representó el contexto predominante durante los dos primeros períodos, con 9 de 12 estudios en 1996–2005 y 12 de 17 en 2006–2014. En los períodos posteriores, aumentó la presencia descriptiva de los experimentos en línea, que representaron 7 de 19 estudios en 2015–2020 y 11 de 22 en 2021–2025.

En el corpus completo, 34 estudios (48,6 %) se realizaron en contextos de laboratorio, 23 (32,9 %) en línea, 8 (11,4 %) utilizaron contextos mixtos y 5 (7,1 %) correspondieron a estudios de campo.

En conjunto, esta distribución indica una mayor diversificación descriptiva de los contextos experimentales en los períodos más recientes, aunque estas variaciones no deben interpretarse, por sí solas, como evidencia de una tendencia temporal.

Tabla 1. Evolución del contexto experimental por período de publicación

Período	Estudios	Laboratorio	En línea	Mixto	Campo
1996–2005	12	9	1	1	1
2006–2014	17	12	4	1	0
2015–2020	19	7	7	3	2
2021–2025	22	6	11	3	2
Total	70	34	23	8	5

Fuente: Elaboración de los autores a partir de la base de datos final, N = 70.

Instrumentos experimentales y perfiles de los participantes

Los instrumentos utilizados para medir los sesgos cognitivos se concentraron en dos arquitecturas principales. Las tareas experimentales representaron 36 estudios (51,4 %), mientras que las simulaciones representaron 30 (42,9 %); los diseños de valor inducido y los juegos económicos representaron conjuntamente cuatro estudios (5,7 %).

En cuanto al tipo específico de decisión experimental, predominaron los mercados simulados (31 estudios), seguidos de la asignación de carteras (15), las decisiones bajo riesgo (8), otras tareas (7), la elección entre loterías (6) y los pronósticos financieros (3).

La medición se basó principalmente en el comportamiento observado (58 estudios), frente a 11 estudios que cuantificaron la intensidad del sesgo y uno que utilizó una medida de presencia.

Los participantes también presentaron una composición heterogénea. Los estudiantes constituyeron el grupo más frecuente, presente en 32 estudios (45,7 %), seguidos por la población general con 21 estudios (30,0 %), las muestras mixtas con 9 (12,9 %) y los inversores o profesionales con 8 (11,4 %). El análisis A1 examinó la asociación entre esta distribución y el tipo de instrumento utilizado.

Tabla 2. Tipo de instrumento experimental según perfil de participantes (A1)

Tipo de instrumento	Estudiantes	Inversionistas / profesionales	Población general	Mixto	Total
Tarea experimental	20	0	10	6	36
Simulación	10	7	10	3	30
Otros	2	1	1	0	4
Total	32	8	21	9	70

Fuente: Elaboración de los autores a partir de la base de datos final, N = 70.

La prueba exacta de Fisher–Freeman–Halton para A1 arrojó un valor p no ajustado de .025, con una magnitud de asociación moderada (V de Cramér = .288). Sin embargo, al considerar conjuntamente los ocho contrastes A1–A4 y T1–T4 y aplicar el procedimiento de Holm, la asociación dejó de proporcionar evidencia estadística suficiente al nivel $\alpha = .05$ (p ajustado = .2008).

Desde una perspectiva descriptiva, las tareas experimentales se concentraron particularmente en las muestras de estudiantes; de los ocho estudios que involucraron inversores o profesionales, siete emplearon simulaciones, mientras que uno se clasificó en la categoría Otros. En consecuencia, A1 revela un patrón descriptivo acompañado de una magnitud de asociación moderada, pero la evidencia inferencial no respalda la conclusión de que el perfil de los participantes esté asociado con el tipo de instrumento después de considerar el ajuste por multiplicidad.

Distribución regional y sesgos cognitivos

La evidencia revela una considerable concentración geográfica. Europa representa 28 de los 70 estudios (40,0 %), mientras que América del Norte representa 21 (30,0 %), lo que significa que ambas regiones conjuntamente constituyen el 70,0 % del corpus. Asia/Oceanía comprende 13 estudios (18,6 %), y América Latina representa tres (4,3 %). Además, dos estudios fueron clasificados como transregionales, mientras que tres no reportaron información suficiente para su clasificación regional.

En cuanto al sesgo cognitivo principal, el efecto disposición estuvo presente en 27 estudios (38,6 %), mientras que la aversión a las pérdidas estuvo presente en 20 (28,6 %). Los 23 estudios restantes (32,9 %) abordaron los demás sesgos considerados. Para examinar la relación entre la región y el sesgo principal, A2 agrupó estos sesgos restantes en una categoría común y excluyó del análisis inferencial únicamente los dos estudios transregionales y los tres estudios sin información regional reportada; en consecuencia, el análisis utilizó $N = 65$. Ningún estudio examinó la disponibilidad o la contabilidad mental como sesgo cognitivo principal dentro de los criterios de elegibilidad y codificación aplicados.

Tabla 3. Región según sesgo cognitivo principal (A2)

Región	Efecto disposición	Aversión a la pérdida	Otros sesgos	Total
América del Norte	6	9	6	21
Europa	11	8	9	28
Asia/Oceanía	8	2	3	13
Latinoamérica	2	0	1	3
Total	27	19	19	65

Fuente: Elaboración de los autores a partir de la base de datos final, $N = 65$.

Aunque pueden observarse diferencias descriptivas, como una presencia relativamente mayor de la aversión a las pérdidas en América del Norte y del efecto disposición en Asia/Oceanía, la prueba exacta de Fisher–Freeman–Halton para A2 arrojó un valor p no ajustado de .465 y una magnitud de asociación de V de Cramér = .214. Tras el ajuste conjunto de los ocho contrastes A1–A4 y T1–T4 mediante el procedimiento de Holm, el valor p ajustado alcanzó 1.0000. Por lo tanto, las diferencias descriptivas observadas no proporcionan evidencia estadística suficiente para respaldar una asociación entre la región y el sesgo cognitivo principal.

Evaluación metodológica de los instrumentos

La evaluación D1 revela diferencias entre las dimensiones examinadas. Cinco criterios directamente relacionados con la arquitectura experimental y su documentación muestran una elevada concentración en la categoría Alta: claridad de los objetivos, 68 estudios (97,1 %); control estructural, 64 (91,4 %); manipulación experimental, 69 (98,6 %); precisión conceptual del instrumento, 69 (98,6 %); y replicabilidad del instrumento, 64 (91,4 %).

La adecuación causal presenta una distribución más diferenciada: 51 estudios (72,9 %) recibieron una clasificación Alta, 17 (24,3 %) una clasificación Moderada y 2 (2,9 %) una clasificación Baja.

Tabla 4. Evaluación metodológica de los instrumentos experimentales (D1)

Criterio	Alto	Moderado	Bajo
Claridad del objetivo	68	2	0
Adecuación causal	51	17	2
Control estructural	64	6	0
Manipulación experimental	69	1	0
Precisión conceptual del instrumento	69	1	0
Replicabilidad del instrumento	64	6	0
Evidencia de validez	59	11	0
Evidencia de confiabilidad	7	0	63
Incentivos	47	6	17
Coherencia muestra–objetivo	56	13	1
Discusión de la representatividad	51	16	3
Limitaciones reportadas	38	30	2

Fuente: Elaboración de los autores a partir de D1, N = 70.

Dado que la concentración de clasificaciones Altas superó el 90 % para la claridad de los objetivos, el control estructural, la manipulación experimental, la precisión conceptual del instrumento y la replicabilidad, estos cinco criterios se consideraron dimensiones con efectos techo severos en el análisis de sensibilidad descriptivo. De acuerdo con el procedimiento predefinido, su codificación y resultados originales se conservaron sin modificaciones; sin embargo, estos criterios no se utilizaron como evidencia acumulativa en la interpretación integradora de la robustez metodológica. Cuando esta interpretación se centró en los criterios restantes: adecuación causal, evidencia de validez, evidencia de confiabilidad, incentivos, correspondencia entre la muestra y la población objetivo, discusión de la representatividad y limitaciones reportadas, el perfil metodológico mantuvo su heterogeneidad: algunas dimensiones mostraron una elevada proporción de valoraciones favorables, mientras que otras, particularmente la evidencia de confiabilidad, presentaron una disponibilidad limitada de evidencia explícitamente documentada. Por lo tanto, la interpretación integradora del perfil metodológico del corpus no depende de la acumulación de criterios con distribuciones altamente concentradas. El análisis de sensibilidad completo se presenta en el Material Suplementario.

Las dimensiones relacionadas con la muestra también mostraron una concentración de clasificaciones Altas. La correspondencia entre la muestra y la población objetivo recibió una clasificación Alta en 56 estudios (80,0 %), Moderada en 13 (18,6 %) y Baja en uno (1,4 %). La discusión de la representatividad produjo 51 clasificaciones Altas (72,9 %), 16 Moderadas (22,9 %) y 3 Bajas (4,3 %). Sin embargo, ambos criterios siguen fundamentos diferentes: la correspondencia entre la muestra y la población objetivo evalúa la relación entre la muestra

utilizada y el objetivo del estudio o la inferencia prevista, mientras que la representatividad depende de la discusión explícita de los autores sobre el alcance, la validez externa o las limitaciones de generalización. Esta distinción forma parte de las reglas de decisión finales de D1.

El análisis A3 examinó específicamente si la discusión de la representatividad difería según el contexto experimental. Entre los 34 estudios de laboratorio, 26 recibieron una clasificación Alta y 8 una clasificación Moderada/Baja; entre los 23 estudios en línea, las frecuencias correspondientes fueron 15 y 8, respectivamente; los cinco estudios de campo recibieron una clasificación Alta; y, entre los ocho estudios que utilizaron contextos mixtos, cinco recibieron una clasificación Alta y tres una clasificación Moderada/Baja. La prueba exacta de Fisher–Freeman–Halton arrojó un valor p no ajustado de .402, con V de Cramér = .214. Tras el ajuste conjunto de los ocho contrastes mediante el procedimiento de Holm, el valor p ajustado alcanzó 1.0000. En consecuencia, las diferencias descriptivas observadas no proporcionan evidencia estadística suficiente para respaldar una asociación entre el contexto experimental y la discusión de la representatividad.

De manera similar, el análisis A4 evaluó la relación entre el tipo de instrumento y la correspondencia entre la muestra y la población objetivo. Se observaron clasificaciones Altas en 29 de las 36 tareas experimentales, 24 de las 30 simulaciones y 3 de los 4 instrumentos agrupados en la categoría Otros. La prueba exacta de Fisher–Freeman–Halton arrojó un valor p no ajustado de 1.000, con V de Cramér = .031; el valor p ajustado mediante Holm fue 1.0000. Por lo tanto, el patrón observado no proporciona evidencia estadística suficiente para respaldar una asociación entre el tipo de instrumento y la correspondencia entre la muestra y la población objetivo.

Evidencia de validez, confiabilidad e incentivos

Las diferencias más pronunciadas dentro de D1 surgen al comparar la evidencia reportada de validez y confiabilidad. La evidencia de validez recibió una clasificación Alta en 59 estudios (84,3 %) y una clasificación Moderada en 11 (15,7 %). En contraste, solo 7 estudios (10,0 %) proporcionaron evidencia de confiabilidad suficiente para recibir una clasificación Alta, mientras que los 63 restantes (90,0 %) recibieron una clasificación Baja.

Esta diferencia debe interpretarse teniendo en cuenta la naturaleza específica de ambos criterios. En particular, una clasificación Baja para la evidencia de confiabilidad no demuestra que un instrumento carezca de confiabilidad. D1-27 evalúa exclusivamente el nivel de evidencia de confiabilidad documentado explícitamente en el artículo; por lo tanto, la concentración en la categoría Baja refleja principalmente la limitada disponibilidad de evidencia reportada que pudiera sustentar una clasificación superior. De manera coherente con esta interpretación, en C1, 63 estudios fueron codificados como «No reportado» y solo 7 como «Sí» respecto de la existencia de evidencia de confiabilidad.

Los incentivos mostraron una distribución más heterogénea: 47 estudios (67,1 %) recibieron una clasificación Alta, 6 (8,6 %) una clasificación Moderada y 17 (24,3 %) una clasificación Baja. En la extracción descriptiva C1, 55 estudios reportaron incentivos monetarios reales, 11 utilizaron incentivos hipotéticos y 4 no reportaron el esquema de incentivos. La diferencia entre ambas

distribuciones se debe a que D1-28 no registra simplemente la presencia de incentivos, sino que evalúa metodológicamente su relevancia y vinculación con las decisiones experimentales.

En conjunto, estos hallazgos muestran que las distintas dimensiones metodológicas presentan perfiles de evaluación diferentes. Mientras que la evidencia de validez recibe predominantemente clasificaciones Altas, la evidencia de confiabilidad explícitamente documentada continúa siendo limitada y los incentivos presentan una distribución más heterogénea. Este patrón refuerza la importancia de distinguir entre la robustez que puede evaluarse a partir de la información disponible y el grado en que determinadas propiedades metodológicas reciben una documentación explícita, sin equiparar ambos conceptos.

Auditoría del reporte procedimental

La auditoría E1 revela una distribución desigual entre las tres prácticas examinadas. El soporte tecnológico fue reportado en 66 de los 70 estudios (94,3 %), mientras que el pilotaje fue reportado en 13 (18,6 %) y el protocolo ético y anonimato en 34 (48,6 %).

Tabla 5. Auditoría del reporte de prácticas procedimentales (E1)

Dimensión procedimental	Sí	No	% Sí
Pilotaje	13	57	18.6
Soporte tecnológico	66	4	94.3
Protocolo ético y anonimato	34	36	48.6

Fuente: Elaboración de los autores a partir de E1, N = 70.

Estas cifras describen la evidencia identificada en los artículos de acuerdo con las reglas de codificación establecidas. En consecuencia, una clasificación «No» para el pilotaje o el protocolo ético no demuestra que la práctica necesariamente haya estado ausente durante la ejecución original del experimento; más bien, indica que la práctica no pudo identificarse en la información reportada. Del mismo modo, el soporte tecnológico constituye una práctica procedimental documentada y no debe interpretarse, por sí mismo, como evidencia de una mayor calidad metodológica ni como evidencia de la confiabilidad del instrumento. Esta distinción resulta particularmente relevante debido al contraste descriptivo entre la elevada tasa de reporte del soporte tecnológico (94,3 %) y la proporción de estudios que recibieron una clasificación Alta en cuanto a evidencia de confiabilidad (10,0 %).

Evolución temporal de las prácticas metodológicas

Los análisis T1–T4 examinaron si determinadas prácticas metodológicas y de reporte estaban asociadas con el período de publicación. Los resultados revelaron patrones descriptivos diferenciados entre las cuatro dimensiones examinadas.

Tabla 6. Evolución temporal de dimensiones metodológicas seleccionadas

Dimensión	1996–2005	2006–2014	2015–2020	2021–2025
Pilotaje: Sí	0/12 (0.0%)	3/17 (17.6%)	2/19 (10.5%)	8/22 (36.4%)
Evidencia de confiabilidad: Alto	0/12 (0.0%)	2/17 (11.8%)	1/19 (5.3%)	4/22 (18.2%)
Protocolo ético y anonimato: Sí	2/12 (16.7%)	7/17 (41.2%)	12/19 (63.2%)	13/22 (59.1%)
Soporte tecnológico: Sí	10/12 (83.3%)	17/17 (100.0%)	17/19 (89.5%)	22/22 (100.0%)

Fuente: Elaboración de los autores a partir de D1 y E1, N = 70.

Los contrastes inferenciales proporcionan contexto para estas variaciones descriptivas. En T1, la proporción de estudios que reportaron pilotaje fue mayor en el período más reciente, alcanzando el 36,4 %, frente al 0,0 % en el período más temprano. La prueba exacta de Fisher–Freeman–Halton arrojó un valor p no ajustado de .0508 y V de Cramér = .342. Después de ajustar conjuntamente los ocho contrastes A1–A4 y T1–T4 mediante el procedimiento de Holm, el valor p ajustado fue .3437. En consecuencia, aunque el patrón descriptivo y la magnitud de la asociación sugieren diferencias entre los períodos que merecen consideración, la evidencia inferencial ajustada no respalda una asociación temporal estadísticamente sustentada.

Para T2, la prueba exacta de Fisher–Freeman–Halton arrojó un valor p no ajustado de .4368, con V de Cramér = .224; después del ajuste de Holm, el valor p alcanzó 1.0000. Aunque el período 2021–2025 presentó la mayor proporción de clasificaciones Altas (18,2 %), la secuencia 0,0 %–11,8 %–5,3 %–18,2 % no sigue un patrón monotónico, y la evidencia inferencial no respalda una asociación entre el período de publicación y la evidencia de confiabilidad reportada.

Para T3, el reporte del protocolo ético y el anonimato aumentó descriptivamente del 16,7 % en 1996–2005 al 41,2 % en 2006–2014 y al 63,2 % en 2015–2020, alcanzando el 59,1 % en 2021–2025. La prueba χ^2 de Pearson produjo $\chi^2(3) = 7.86$, con un valor p no ajustado de .0491 y V de Cramér = .335. Sin embargo, después del ajuste conjunto de Holm, el valor p ajustado fue .3437. Por lo tanto, los datos muestran una mayor frecuencia descriptiva de reporte de estas salvaguardas en los períodos más recientes en comparación con el período inicial, acompañada de una asociación de magnitud apreciable; sin embargo, la evidencia inferencial ajustada no respalda una asociación temporal estadísticamente sustentada. El patrón tampoco sigue una trayectoria estrictamente monotónica, ya que el reporte disminuyó ligeramente entre los dos últimos períodos.

Finalmente, T4 muestra que el reporte del soporte tecnológico se mantuvo elevada durante prácticamente todo el período de estudio: 83,3 %, 100,0 %, 89,5 % y 100,0 %, respectivamente. La prueba exacta de Fisher–Freeman–Halton arrojó un valor p no ajustado de .0603 y V de Cramér = .289; después del ajuste de Holm, el valor p ajustado fue .3437. Por lo tanto, la evidencia inferencial ajustada no respalda una asociación temporal en un contexto caracterizado por una elevada frecuencia de reporte del soporte tecnológico en los cuatro períodos.

Síntesis integrada de los resultados

En conjunto, los hallazgos revelan una literatura experimental caracterizada por una elevada prevalencia de tareas experimentales y simulaciones, una diversificación descriptiva de los contextos de aplicación en los períodos más recientes y una concentración geográfica principalmente en Europa y América del Norte. En los análisis A1–A4, ninguno de los cuatro contrastes proporcionó evidencia estadística suficiente de una asociación después del ajuste por multiplicidad. No obstante, A1 reveló un patrón descriptivo entre el tipo de instrumento y el perfil de los participantes, acompañado de una magnitud de asociación moderada (V de Cramér = .288), aunque esta evidencia no mantuvo respaldo inferencial después del ajuste de Holm. Los análisis A2–A4 también proporcionaron evidencia estadística insuficiente para respaldar asociaciones entre la región y el sesgo cognitivo principal, el contexto experimental y la discusión

de la representatividad, o el tipo de instrumento y la correspondencia entre la muestra y la población objetivo, respectivamente.

La evaluación D1 reveló perfiles diferenciados entre las dimensiones metodológicas. Cinco criterios relacionados con la arquitectura experimental y su documentación mostraron concentraciones superiores al 90 % de clasificaciones Altas y, por ello, fueron examinados mediante un análisis de sensibilidad descriptivo. Después de excluir estos criterios únicamente de la interpretación integradora, sin modificar su codificación ni sus resultados originales, los criterios restantes continuaron mostrando un perfil metodológico heterogéneo: la evidencia de validez recibió predominantemente clasificaciones Altas, mientras que la evidencia de confiabilidad explícitamente documentada mostró una marcada concentración en la categoría Baja. Otras dimensiones, como los incentivos, la correspondencia entre la muestra y la población objetivo, la discusión de la representatividad y las limitaciones reportadas, presentaron distribuciones más diferenciadas. Estos hallazgos indican que la interpretación del perfil metodológico del corpus no depende de la acumulación de criterios afectados por efectos techo.

La auditoría del reporte de prácticas procedimentales refuerza esta heterogeneidad: el soporte tecnológico estuvo ampliamente documentado, mientras que el pilotaje y el protocolo ético y anonimato aparecieron con menor frecuencia. Los análisis temporales T1–T4 también proporcionaron evidencia estadística insuficiente de asociaciones después del ajuste conjunto por multiplicidad. No obstante, T1 y T3 revelaron patrones descriptivos y magnitudes de asociación que merecen consideración: el pilotaje se reportó con mayor frecuencia en el período más reciente (V de Cramér = .342), mientras que el reporte del protocolo ético o el anonimato ocurrió con mayor frecuencia en los períodos recientes que en el período inicial (V de Cramér = .335), aunque ninguno de estos patrones mantuvo evidencia inferencial suficiente después del ajuste de Holm.

En conjunto, los hallazgos revelan diferencias entre la robustez metodológica que puede evaluarse a partir de la información publicada y el grado en que determinadas prácticas y propiedades metodológicas reciben una documentación explícita, sin proporcionar evidencia de una mejora temporal generalizada en estas dimensiones. Estos resultados constituyen la base para su interpretación conceptual en la discusión.

Discusión

De la psicometría a la arquitectura de las decisiones

Los resultados muestran que la medición experimental de los sesgos cognitivos en la toma de decisiones financieras se basa predominantemente en arquitecturas diseñadas para observar elecciones y respuestas conductuales. Las tareas experimentales y las simulaciones representan conjuntamente 66 de los 70 estudios analizados (94,3 %), e incluyen mercados simulados, decisiones de cartera, loterías, tareas basadas en riesgo y otras configuraciones. Este predominio es coherente con la tradición de la economía experimental, que investiga los mecanismos de toma de decisiones mediante entornos estructurados y condiciones controladas (Harrison & List, 2004; Hertwig & Ortmann, 2001; Levitt & List, 2007).

Sin embargo, el predominio de estas arquitecturas no implica homogeneidad en la medición. Diferentes procedimientos de elicitación pueden generar medidas que no necesariamente demuestran equivalencia, incluso cuando buscan aproximarse a constructos relacionados (Crosetto & Filippin, 2016; Frey et al., 2017; Pedroni et al., 2017). Desde la perspectiva de la validez de constructo, esta heterogeneidad resulta relevante porque la correspondencia entre un constructo teórico y su operacionalización depende de si el procedimiento representa adecuadamente el fenómeno que busca medir (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995). Esta cuestión resulta particularmente evidente en sesgos como el exceso de confianza, cuyas distintas manifestaciones no necesariamente representan un fenómeno unitario (Moore & Healy, 2008).

En consecuencia, los instrumentos experimentales pueden entenderse no simplemente como mecanismos de registro, sino también como componentes de la arquitectura mediante la cual los sesgos se hacen observables. Los resultados de esta revisión no establecen que una arquitectura produzca mediciones más válidas que otra; sin embargo, muestran que la operacionalización varía entre los estudios. Por consiguiente, la comparación de la evidencia requiere considerar no solo el sesgo investigado, sino también la tarea, el contexto y las reglas mediante las cuales los investigadores eliciten las respuestas conductuales.

Brecha entre robustez metodológica y transparencia

La evaluación metodológica revela diferencias entre las dimensiones que pueden evaluarse a partir del diseño descrito y aquellas que dependen en mayor medida de evidencia documentada explícitamente. La manipulación experimental y la precisión conceptual recibieron clasificaciones Altas en el 98,6 % de los estudios, mientras que el control estructural y la replicabilidad alcanzaron el 91,4 %. Sin embargo, estas cuatro dimensiones, junto con la claridad de los objetivos, mostraron concentraciones superiores al 90 % en la categoría Alta, lo que indica la presencia de efectos techo severos y una capacidad potencialmente limitada para discriminar dentro del corpus. Por esta razón, el análisis de sensibilidad excluyó estos criterios únicamente de la interpretación integradora, sin modificar sus codificaciones ni sus resultados originales. Cuando la interpretación se centra en los criterios restantes, persiste un perfil metodológico heterogéneo; por ejemplo, la adecuación causal recibió una clasificación Alta en el 72,9 % de los estudios, mientras que otras dimensiones presentaron distribuciones más diferenciadas.

El patrón resulta particularmente pronunciado en determinadas propiedades de medición. La evidencia de validez recibió una clasificación Alta en el 84,3 % de los estudios, mientras que solo el 10,0 % proporcionó evidencia de confiabilidad suficiente para alcanzar dicha categoría. El 90,0 % restante recibió una clasificación Baja, pero este hallazgo no implica que los instrumentos carezcan de confiabilidad. De acuerdo con las reglas de decisión de D1, indica que la información publicada no proporcionó evidencia explícita suficiente para sustentar una clasificación superior. Esta distinción resulta especialmente importante dada la heterogeneidad de los instrumentos identificados, ya que el uso previo o repetido de una tarea no sustituye la documentación de sus propiedades de medición (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995).

Una asimetría similar surge en el reporte de prácticas procedimentales. El soporte tecnológico apareció en el 94,3 % de los estudios, mientras que el pilotaje fue reportado en el 18,6 % y el protocolo ético y anonimato en el 48,6 %. Nuevamente, la ausencia de información reportada

no demuestra que estas prácticas no se hayan realizado. Del mismo modo, el soporte tecnológico puede facilitar la administración estandarizada y el registro de las decisiones, pero no constituye, por sí mismo, evidencia de confiabilidad ni de una mayor calidad metodológica.

En conjunto, estos hallazgos revelan una distinción entre las dimensiones de la robustez metodológica que pueden evaluarse a partir de la información publicada y el grado en que determinadas propiedades y prácticas metodológicas reciben una documentación explícita. Esta interpretación se mantiene cuando los criterios D1 afectados por efectos techo se consideran por separado y, por lo tanto, no depende de la acumulación de clasificaciones favorables en dimensiones con capacidad discriminativa limitada. Esta distinción no constituye una medida global de calidad ni permite inferir una deficiencia metodológica cuando determinada información no ha sido reportada. Los análisis temporales complementan esta interpretación al no proporcionar evidencia de una mejora generalizada y uniforme en las dimensiones examinadas.

Diseño experimental, participantes y validez ecológica

El análisis A1 reveló un patrón descriptivo entre el tipo de instrumento y el perfil de los participantes: las tareas experimentales se utilizaron principalmente en estudios que involucraron estudiantes, mientras que las simulaciones representaron la mayoría de los estudios que involucraron inversores o profesionales. La magnitud de la asociación fue moderada (V de Cramér = .288) y el valor p no ajustado fue .025; sin embargo, la evidencia inferencial no mantuvo respaldo después de la corrección por multiplicidad mediante el procedimiento de Holm (p ajustado = .2008*). Por lo tanto, el patrón observado resulta metodológicamente relevante para caracterizar el corpus, pero no respalda una asociación estadísticamente sustentada entre el tipo de instrumento y el perfil de los participantes.

La literatura experimental ha destacado el equilibrio entre control y contextualización: los entornos de laboratorio permiten aislar mecanismos bajo condiciones controladas, mientras que los diseños más contextualizados pueden aproximarse a determinados entornos de toma de decisiones (Falk & Heckman, 2009; Harrison & List, 2004; Levitt & List, 2007). Dentro de este marco, el patrón observado entre participantes e instrumentos podría ser compatible con diferentes requerimientos de contextualización; sin embargo, nuestros datos no permiten atribuir dicha intención a los investigadores ni establecer el mecanismo que podría explicar esta distribución. Asimismo, A3 proporcionó evidencia estadística insuficiente de una asociación entre el contexto experimental y la discusión de la representatividad (V de Cramér = .214), mientras que A4 proporcionó evidencia insuficiente de una asociación entre el tipo de instrumento y la correspondencia entre la muestra y la población objetivo (V de Cramér = .031). En consecuencia, el patrón observado en A1 no debe extrapolarse a diferencias sistemáticas en estas otras dimensiones.

Los contextos experimentales también muestran una presencia relativamente mayor de estudios en línea en los períodos más recientes. Sin embargo, este patrón sigue siendo descriptivo y no debe interpretarse como una tendencia estadísticamente demostrada. De manera complementaria, el reporte del soporte tecnológico se mantuvo elevada durante prácticamente todo el período de estudio. Aunque T4 produjo una magnitud de asociación de V de Cramér = .289, la evidencia inferencial no alcanzó suficiente respaldo después del ajuste por multiplicidad.

En conjunto, estos hallazgos no respaldan una interpretación lineal según la cual la infraestructura experimental habría avanzado progresivamente mientras otras prácticas permanecían rezagadas.

Hacia un marco tripartito de transparencia

Los patrones identificados permiten organizar la transparencia metodológica en tres dimensiones complementarias: transparencia tecnológica, transparencia de la medición y estadística, y transparencia ética. Esta propuesta no constituye una escala validada ni un estándar normativo, sino una perspectiva analítica para identificar la información necesaria para comprender y evaluar la implementación de los instrumentos experimentales.

La transparencia tecnológica comprende la documentación de las plataformas, el software y los procedimientos utilizados para administrar las tareas y registrar las decisiones. Esta dimensión presentó la mayor extensión en cuanto a su reporte procedimental, aunque su presencia no debe confundirse con evidencia relativa a las propiedades de medición. La transparencia de la medición y estadística comprende la documentación de la operacionalización del instrumento, la evidencia que respalda sus propiedades de medición, particularmente la validez y la confiabilidad, y los procedimientos estadísticos pertinentes utilizados para analizar e interpretar los resultados. La marcada diferencia observada entre estas dimensiones muestra que una arquitectura experimental técnicamente reproducible no garantiza una documentación igualmente exhaustiva de las propiedades de las medidas empleadas. La investigación sobre reproducibilidad ha destacado la importancia de reportar explícitamente los procedimientos y la evidencia necesarios para evaluar y replicar los hallazgos experimentales (Camerer et al., 2016, 2018; Christensen & Miguel, 2018; Open Science Collaboration, 2015).

La transparencia ética comprende el reporte del consentimiento, el anonimato, la confidencialidad y otras salvaguardas aplicables a los participantes. Los análisis temporales no proporcionaron evidencia estadística suficiente de una asociación, después del ajuste por multiplicidad, para ninguna de las dimensiones examinadas. No obstante, surgieron patrones descriptivos diferenciados. En T3, el reporte del protocolo ético y el anonimato ocurrió con mayor frecuencia en los períodos recientes que en el período inicial, con una magnitud de asociación de V de Cramér = .335. De manera similar, T1 mostró una mayor frecuencia de reporte de pilotaje en el período más reciente y una magnitud de asociación comparable (V de Cramér = .342). En ambos casos, la evidencia inferencial no mantuvo respaldo después del ajuste de Holm. T2 y T4 también proporcionaron evidencia estadística insuficiente de asociaciones temporales. Por lo tanto, los resultados no respaldan una mejora temporal generalizada de la transparencia, aunque revelan patrones descriptivos diferenciados entre las dimensiones examinadas.

En conjunto, estas tres dimensiones complementan la evaluación de la arquitectura experimental al distinguir cómo se implementa un instrumento, qué evidencia sustenta su medición y qué salvaguardas procedimentales documentan los investigadores. Por lo tanto, el marco tripartito debe entenderse como una propuesta conceptual derivada de los patrones identificados y sujeta a validación posterior.

Para traducir este marco en una guía de reporte más operativa, la Tabla 7 resume la información que los futuros estudios experimentales deberían considerar documentar dentro de cada dimensión de transparencia. Estos elementos constituyen una orientación prospectiva para el reporte, no una escala de evaluación validada ni un estándar normativo; por consiguiente, no deben interpretarse como una extensión de los instrumentos D1–E1 utilizados en esta revisión.

Tabla 7. Guía de reporte propuesta para el marco tripartito de transparencia

Dimensión de transparencia	Información que debería documentarse
Transparencia tecnológica	Plataforma o software experimental; modalidad de administración; configuración tecnológica relevante del experimento; procedimientos tecnológicos utilizados para controlar o monitorear a los participantes.
Transparencia de medición y estadística	Operacionalización del sesgo cognitivo; reglas de puntuación o clasificación; evidencia o justificación de validez y confiabilidad, cuando corresponda; procedimientos de pilotaje o calibración; procedimientos estadísticos relevantes.
Transparencia ética	Procedimientos de consentimiento informado; salvaguardas de anonimato o confidencialidad; protección de datos; revisión o aprobación ética, cuando corresponda; tratamiento de incentivos y posibles riesgos para los participantes.

Fuente: Elaboración de los autores.

Contribución a la literatura

Esta revisión realiza tres contribuciones principales. En primer lugar, desplaza el foco desde la identificación de los sesgos cognitivos hacia el examen de cómo estos se operacionalizan experimentalmente. Este enfoque muestra que un mismo constructo puede estudiarse mediante diferentes tareas, contextos y poblaciones, lo que convierte a la arquitectura de medición en un elemento relevante para interpretar y comparar la evidencia.

En segundo lugar, identifica diferencias entre las dimensiones de robustez metodológica que pueden evaluarse a partir de la información publicada y el grado en que determinadas propiedades y prácticas metodológicas se documentan explícitamente. Esta interpretación se mantiene cuando los criterios D1 afectados por efectos techo se consideran por separado y revela perfiles diferenciados entre dimensiones como la adecuación causal, la evidencia de validez y confiabilidad, los incentivos, la correspondencia entre la muestra y la población objetivo, la discusión de la representatividad y las limitaciones reportadas. Los análisis A1–A4 y T1–T4 también distinguen los patrones descriptivos de la evidencia inferencial: ninguno de los ocho contrastes proporcionó evidencia estadística suficiente de asociación después del ajuste por multiplicidad, aunque A1, T1 y T3 mostraron patrones descriptivos y magnitudes de asociación relevantes para caracterizar e interpretar el corpus.

En tercer lugar, la estructura C1–D1–E1 separa la extracción descriptiva, la evaluación metodológica y la auditoría procedimental, reduciendo así el riesgo de interpretar automáticamente la ausencia de información reportada como evidencia sobre las prácticas efectivamente implementadas. El marco fue desarrollado específicamente para esta revisión mediante la integración y operacionalización de principios metodológicos consolidados procedentes de la metodología de revisiones sistemáticas, el diseño experimental, la medición, la economía experimental, la reproducibilidad y la ética de la investigación. Su aplicación produjo niveles sustanciales de concordancia entre evaluadores; sin embargo, el marco no fue sometido

a una validación externa formal y, por lo tanto, sus resultados deben interpretarse teniendo en cuenta esta característica.

Limitaciones y futuras líneas de investigación

Esta revisión presenta varias limitaciones. En primer lugar, la evaluación depende de la información reportada en los estudios publicados. Por lo tanto, la ausencia de evidencia sobre la confiabilidad, el pilotaje, las salvaguardas éticas u otras prácticas puede reflejar tanto su ausencia efectiva como una omisión en el reporte de la información. En consecuencia, las categorías correspondientes deben interpretarse como limitaciones en la evaluabilidad y transparencia, y no automáticamente como deficiencias en la implementación de los estudios.

En segundo lugar, la búsqueda se restringió a artículos publicados en inglés entre 1990 y 2025 e indexados en Web of Science–SSCI o Scopus. En consecuencia, es posible que no se hayan recuperado estudios indexados exclusivamente en otras bases de datos bibliográficas especializadas, como EconLit o PsycINFO. Además, la estrategia de búsqueda mantuvo un operador NOT destinado a excluir estudios basados exclusivamente en encuestas o autoinformes. El análisis de sensibilidad realizado sin este operador recuperó nueve registros adicionales de Web of Science y once de Scopus; debido a que estos registros no fueron sometidos a una evaluación de elegibilidad documentada de manera independiente, no puede determinarse si alguno de ellos habría cumplido finalmente los criterios de inclusión. Esta decisión constituye una limitación de la estrategia de recuperación; por lo tanto, las conclusiones deben interpretarse en relación con el corpus efectivamente identificado mediante la estrategia de búsqueda aplicada.

En tercer lugar, C1–D1–E1 fue desarrollado específicamente para esta revisión y, aunque el marco fue sometido a una calibración interna y aplicado de manera independiente por dos revisores, no ha sido objeto de validación externa. Los coeficientes Kappa reportados representan medidas no ponderadas de concordancia global y, en el caso de las evaluaciones ordinales, no distinguen la magnitud de las discrepancias entre categorías. Además, cinco criterios D1 mostraron concentraciones superiores al 90 % en la categoría Alta, lo que limita su capacidad discriminante dentro del corpus. El análisis de sensibilidad descriptivo mostró que la interpretación integradora se mantiene consistente cuando estos criterios se consideran por separado, sin modificar las codificaciones originales; no obstante, los efectos techo deben tenerse en cuenta al interpretar el desempeño del marco y representan un aspecto que futuras aplicaciones y procesos de validación deberían examinar.

En cuarto lugar, los análisis A1–A4 y T1–T4 examinan asociaciones entre características de los estudios y no pueden establecer mecanismos causales. Asimismo, debido a que se realizaron múltiples contrastes, la evidencia inferencial se evaluó mediante el procedimiento de Holm. Ninguno de los ocho análisis mantuvo evidencia estadística suficiente de asociación después de este ajuste, aunque algunos mostraron patrones descriptivos y magnitudes de asociación de interés. En consecuencia, estos hallazgos deben interpretarse integrando la evidencia inferencial ajustada, las magnitudes de asociación y los patrones descriptivos observados, en lugar de recurrir a una clasificación dicotómica basada exclusivamente en un umbral de significación.

Las futuras investigaciones podrían comparar directamente diferentes instrumentos diseñados para medir un mismo sesgo bajo condiciones equivalentes, evaluando su estabilidad entre tareas, contextos y perfiles de participantes. También sería pertinente ampliar el análisis a enfoques deliberadamente excluidos de esta revisión, como los instrumentos basados exclusivamente en autoinformes, así como evaluar externamente tanto el marco C1–D1–E1 como el marco de transparencia tripartito propuesto. En particular, futuras aplicaciones del marco podrían examinar la capacidad discriminante de los criterios D1 y evaluar sus propiedades mediante procedimientos de validación independientes. Este tipo de investigación contribuiría a determinar en qué medida la heterogeneidad de las arquitecturas experimentales identificada en esta revisión se traduce en diferencias efectivas en la validez, la confiabilidad y la comparabilidad de las medidas experimentales.

Conclusiones

Este estudio proporciona una evaluación sistemática y metodológica de los instrumentos experimentales utilizados para medir los sesgos cognitivos en la toma de decisiones financieras. A partir del análisis de 70 estudios, los hallazgos muestran un claro predominio de las tareas experimentales y las simulaciones, así como perfiles diferenciados entre las dimensiones metodológicas evaluadas. Algunas dimensiones de la arquitectura experimental y de su documentación mostraron una alta concentración de clasificaciones en la categoría Alta; sin embargo, cinco de ellas presentaron efectos techo severos y, por lo tanto, su contribución a la interpretación integradora fue examinada mediante un análisis de sensibilidad.

Al considerar este análisis de sensibilidad, persiste una marcada heterogeneidad entre las dimensiones metodológicas restantes. Si bien algunas dimensiones reciben predominantemente valoraciones favorables, otras que dependen de evidencia o documentación reportada explícitamente muestran mayores limitaciones, particularmente en lo que respecta a la evidencia de confiabilidad. Estas diferencias no implican necesariamente deficiencias en las prácticas que no fueron documentadas, ni respaldan la construcción de una medida global de calidad metodológica. Asimismo, los análisis realizados no proporcionan evidencia de una evolución temporal generalizada y uniforme de las prácticas metodológicas y de reporte de información examinadas.

A partir de estos hallazgos, el estudio propone un marco de transparencia tripartito que integra las dimensiones tecnológicas, de medición y estadística, y ética como perspectivas complementarias para examinar la información metodológica proporcionada por los estudios experimentales. Este marco constituye una propuesta conceptual y analítica, más que un estándar normativo validado. Asimismo, distingue entre la documentación de la implementación experimental, la evidencia que sustenta las propiedades de medición y el reporte de información sobre las salvaguardas aplicables a los participantes.

En términos generales, la principal contribución de este estudio consiste en desplazar el foco analítico desde qué sesgos cognitivos estudian los investigadores hacia cómo los miden experimentalmente. Los hallazgos muestran que los instrumentos no constituyen mecanismos de medición homogéneos, sino arquitecturas metodológicas mediante las cuales los investigadores operacionalizan y observan los sesgos cognitivos. Por lo tanto, la interpretación y comparación de la evidencia requiere considerar no solo el sesgo objeto de estudio, sino

también las características de las tareas, los contextos experimentales y los procedimientos utilizados para obtener las respuestas conductuales.

En consecuencia, fortalecer la transparencia en la documentación de los instrumentos y procedimientos experimentales puede facilitar la evaluación de sus propiedades metodológicas y mejorar la comparabilidad, reproducibilidad y acumulación de evidencia en las finanzas conductuales experimentales. En particular, una documentación más explícita de las propiedades de medición y de las prácticas procedimentales permitiría a los investigadores evaluar con mayor precisión las similitudes y diferencias entre las diversas arquitecturas utilizadas para estudiar los sesgos cognitivos.

Referencias

- Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., & Jordan, Z. (Eds.). (2024). JBI manual for evidence synthesis. JBI. <https://doi.org/10.46658/JBIMES-24-01>
- Badola, S., Sahu, A. K., & Adlakha, A. (2024). A systematic review on behavioral biases affecting individual investment decisions. *Qualitative Research in Financial Markets*, 16(3), 448-476. <https://doi.org/10.1108/QRFM-05-2022-0095>
- Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. En G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), *Handbook of the Economics of Finance* (Vol. 1, pp. 1053-1128). Elsevier. [https://doi.org/10.1016/S1574-0102\(03\)01027-6](https://doi.org/10.1016/S1574-0102(03)01027-6)
- Biais, B., Hilton, D., Mazurier, K., & Pouget, S. (2005). Judgemental overconfidence, self-monitoring and trading performance in an experimental financial market. *The Review of Economic Studies*, 72(2), 287-312. <https://doi.org/10.1111/j.1467-937X.2005.00333.x>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436. <https://doi.org/10.1126/science.aaf0918>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644. <https://doi.org/10.1038/s41562-018-0399-z>
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7-42. <https://doi.org/10.1023/A:1007850605129>
- Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 368, 16890. <https://doi.org/10.1136/bmj.l6890>
- Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920-980. <https://doi.org/10.1257/jel.20171350>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://doi.org/10.1177/001316446002000104>
- Critical Appraisal Skills Programme. (2018). CASP checklists. <https://casp-uk.net/casp-tools-checklists/>

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/h0040957>
- Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19, 613-641. <https://doi.org/10.1007/s10683-015-9457-9>
- De Oliveira Cardoso, N., Zorgi Salvador, E., Broch, G., Mette, F. M. B., Yoshinaga, C. E., & De Lara Machado, W. (2024). Measuring behavioral biases in individual investors decision-making and sociodemographic correlations: A systematic review. *Qualitative Research in Financial Markets*, 16(4), 636-659. <https://doi.org/10.1108/QRFM-05-2022-0090>
- Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952), 535-538. <https://doi.org/10.1126/science.1168244>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456-465. <https://doi.org/10.1177/2515245920952393>
- Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10), e1701381. <https://doi.org/10.1126/sciadv.1701381>
- Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review*, 32(1), 1-36. <https://doi.org/10.1007/s10713-007-0003-3>
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009-1055. <https://doi.org/10.1257/0022051043004577>
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24(3), 383-403. <https://doi.org/10.1017/S0140525X01004149>
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644-1655. <https://doi.org/10.1257/000282802762024700>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-292. <https://doi.org/10.2307/1914185>
- Kumar, S., & Goyal, N. (2015). Behavioural biases in investment decision making – a systematic literature review. *Qualitative Research in Financial Markets*, 7(1), 88-108. <https://doi.org/10.1108/QRFM-07-2014-0022>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2), 153-174. <https://doi.org/10.1257/jep.21.2.153>
- Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization*, 119, 254-266. <https://doi.org/10.1016/j.jebo.2015.08.003>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502-517. <https://doi.org/10.1037/0033-295X.115.2.502>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11), 803-809. <https://doi.org/10.1038/s41562-017-0219-x>

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2), 274-279. <http://www.jstor.org/stable/1817233>

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review*, 106(7), 1577-1600. <https://doi.org/10.1257/aer.106.7.1577>

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Zahera, S. A., & Bansal, R. (2018). Do investors exhibit behavioral biases in investment decision making? A systematic review. *Qualitative Research in Financial Markets*, 10(2), 210-251. <https://doi.org/10.1108/QRFM-04-2017-0028>