

Experimental Instruments for Measuring Cognitive Biases in Financial Decision-Making: A Systematic Review and Methodological Assessment

Héctor Madariaga Tapia

Corresponding author

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>

Abstract

This study aims to systematically review and evaluate the experimental instruments used to measure cognitive biases in financial decision-making, examining their methodological characteristics and implications for the comparability and transparency of behavioral evidence. Following the PRISMA 2020 guidelines, a systematic review covered the period from 1990 to 2025 and yielded a final corpus of 70 studies. The analysis employed a three-component framework: descriptive extraction (C1), methodological robustness assessment (D1), and procedural reporting audit (E1). Inter-rater agreement was assessed using Cohen's Kappa. Association and temporal analyses incorporated the Holm adjustment for multiple comparisons.

Experimental tasks and simulations accounted for 94.3% of the studies. The methodological dimensions exhibited heterogeneous profiles: several criteria predominantly received High classifications, although five showed severe ceiling effects, while explicitly documented reliability evidence remained limited and reporting of procedural practices varied across studies. Sensitivity analysis indicated that the methodological profile remained heterogeneous after excluding the five criteria affected by severe ceiling effects from the integrative interpretation.

None of the eight contrasts provided sufficient statistical evidence after the Holm adjustment, although some revealed relevant descriptive patterns and association magnitudes. The review relies on the information reported in the primary studies and on the search and eligibility criteria applied. The C1–D1–E1 framework was developed specifically for this review and has not undergone external validation.

More systematic documentation of measurement properties and procedural practices may strengthen the transparency, comparability, and reproducibility of experimental research in

behavioral finance. The study shifts attention from which cognitive biases researchers examine to how they measure such biases experimentally, integrating descriptive extraction, methodological assessment, and procedural auditing. It also provides an empirical foundation for a conceptual proposal encompassing technological, measurement and statistical, and ethical transparency.

Keywords: Cognitive biases; Behavioral finance; Experimental methods; Measurement instruments; Systematic review; Methodological transparency.

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

© Creative Commons Attribution 4.0
Received 24 June, 2026
Accepted 27 September, 2026
Published 2 October, 2026

Introduction

The study of cognitive biases in financial decision-making has expanded considerably over the past decades, becoming a central component of behavioral and experimental economics. A broad body of empirical and theoretical literature has documented systematic deviations from rational choice, identifying biases and heuristics such as overconfidence, anchoring, loss aversion, the herding effect, availability, and representativeness across diverse financial contexts (Barberis & Thaler, 2003; Kahneman & Tversky, 1979; Thaler, 2016; Tversky & Kahneman, 1974). These findings have contributed significantly to understanding how individuals make decisions under conditions of risk and uncertainty.

Despite these advances, one methodological issue has received comparatively less attention: how researchers experimentally measure cognitive biases. In practice, researchers employ diverse tasks, simulations, and decision structures to elicit behavioral responses, ranging from choices between lotteries to portfolio allocations, forecasts, and simulated markets (Biais et al., 2005; Glaser & Weber, 2007; Holt & Laury, 2002). This diversity involves more than differences in format. Methodological evidence indicates that different elicitation procedures can produce measures that do not necessarily demonstrate equivalence, even when they seek to approximate related constructs (Crosetto & Filippin, 2016; Frey et al., 2017; Pedroni et al., 2017). Likewise, a single bias may encompass conceptually distinct manifestations, as demonstrated by the case of overconfidence (Moore & Healy, 2008).

This heterogeneity carries important implications for construct validity and the comparability of evidence. The correspondence between a theoretical construct and its operationalization depends on whether the procedure employed adequately represents the phenomenon it seeks to measure (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995). In experimental research, differences in tasks, incentives, contexts, and procedures can introduce substantial variation in how researchers elicit and observe behavioral responses (Camerer & Hogarth, 1999; Harrison & List, 2004; Hertwig & Ortmann, 2001). These issues become particularly relevant to the accumulation and replication of evidence because comparisons across studies require an

understanding not only of their results but also of the methodological conditions under which researchers obtained those results (Camerer et al., 2016, 2018; Christensen & Miguel, 2018).

Existing systematic reviews in behavioral finance have focused primarily on identifying and classifying biases that affect investment decisions, as well as synthesizing their effects and relationships with investor characteristics (Badola et al., 2024; Kumar & Goyal, 2015; Zahera & Bansal, 2018). In parallel, methodological studies have shown that different procedures used to measure behavioral constructs may differ in their measurement properties and do not necessarily yield equivalent results (Lönnqvist et al., 2015; Pedroni et al., 2017). More recently, De Oliveira Cardoso et al. (2024) extended this discussion through a systematic review focused on the psychometric evidence of instruments used to measure behavioral biases among individual investors. However, systematic assessment specifically targeting the experimental architectures through which researchers operationalize cognitive biases behaviorally remains comparatively underdeveloped, particularly an assessment that jointly examines their instrumental, methodological, and procedural characteristics. This gap supports shifting the focus from which biases researchers identify and what effects they produce toward how researchers experimentally measure them and under which methodological conditions they generate and document such evidence.

From a methodological perspective, the literature provides grounds for examining two types of patterns in this review. First, because experimental designs differ in their levels of control, contextualization, and realism, it becomes relevant to examine whether experimental architectures display differentiated patterns according to participant characteristics and the contexts in which researchers apply them (Harrison & List, 2004; Hertwig & Ortmann, 2001).

Second, changes in the technological possibilities for experimentation and growing attention to transparency and reproducibility warrant an examination of how specific methodological and reporting practices have evolved over time (Camerer et al., 2016, 2018; Christensen & Miguel, 2018). These considerations serve a conceptual and guiding function: they do not constitute confirmatory hypotheses and do not presuppose the existence, direction, or statistical significance of the associations examined.

Against this background, this study aims to identify, classify, and methodologically evaluate the experimental instruments used to measure cognitive biases in individual financial decision-making. Rather than focusing primarily on which biases researchers' study or estimating their aggregate effects, the review examines how researchers operationalize these constructs through different experimental architectures and what methodological and procedural evidence can be evaluated from the published information. To this end, the study employs a three-component analytical framework that distinguishes descriptive information extraction (C1), the assessment of specific dimensions of methodological robustness (D1), and an audit of the reporting of procedural practices (E1). This framework keeps the information explicitly reported by the studies separate from the methodological assessment conducted through predefined criteria and does not treat the framework as a global quality scale.

Building on this purpose, the review addresses three research questions. First, how do researchers characterize and operationalize the experimental instruments used to measure cognitive biases in financial decision-making? Second, what patterns do these instruments

exhibit in terms of their methodological characteristics, measurement evidence, and reporting of procedural practices? Third, what associations emerge between specific instrument characteristics, participant characteristics, and experimental contexts, and how do specific methodological and reporting practices vary over time? These questions integrate descriptive characterization with targeted association and temporal analyses while distinguishing observed patterns from the inferential evidence available to substantiate them.

The study makes three main contributions. First, it provides a structured and comparative characterization of the experimental architectures used to operationalize cognitive biases in financial contexts. Second, it integrates descriptive extraction, methodological assessment, and procedural auditing through the C1–D1–E1 framework, thereby distinguishing observable design characteristics from dimensions whose assessment depends on explicitly documented evidence. Third, it examines the implications of instrument heterogeneity for methodological robustness, transparency, and the comparability of experimental evidence. In this way, the review contributes to a methodological understanding of the experimental measurement of financial biases and provides an empirical basis for identifying areas in which more explicit documentation could strengthen the assessment, reproducibility, and accumulation of evidence in this field.

Methodology

Research Design

This study conducts a systematic review with a methodological orientation aimed at identifying, classifying, and evaluating the experimental instruments used to measure cognitive biases in individual financial decision-making. Unlike reviews designed to estimate aggregate behavioral effects, the analysis focuses on how cognitive biases are operationalized through different experimental designs and instruments and on the methodological robustness that can be assessed from the information reported in primary studies.

The review encompasses eight cognitive biases recognized as highly relevant in behavioral finance: overconfidence, anchoring, herding effect, loss aversion, availability, disposition effect, mental accounting, and representativeness. These biases were selected based on their theoretical relevance and recurrent presence in the literature on financial decision-making (Barberis & Thaler, 2003; Kahneman & Tversky, 1979; Thaler, 2016).

The review process followed the PRISMA 2020 guidelines (Page et al., 2021), with the aim of promoting transparency and traceability throughout study identification, selection, and evaluation. Methodological assessment drew, as complementary references, on principles from the Joanna Briggs Institute (JBI) (Aromataris et al., 2024) and the Critical Appraisal Skills Programme (CASP) (Critical Appraisal Skills Programme, 2018).

Because these frameworks were not specifically designed to evaluate experimental instruments for measuring cognitive biases in financial decision-making, this review developed a dedicated framework by integrating and operationalizing principles from systematic review methodology

and critical appraisal, experimental design and causal inference, measurement theory, experimental economics, reproducibility, external validity, and research ethics.

The framework comprised three components with distinct yet complementary functions: C1, for structured descriptive data extraction; D1, for assessing specific dimensions of methodological robustness; and E1, for auditing the explicit reporting of selected procedural practices. Accordingly, these components should not be interpreted as equivalent quality scales. The rationale and traceability of the criteria, together with their application rules and examples, appear in the Supplementary Material. The reviewers refined the framework through discussion and a preliminary calibration phase; however, no independent external validation was conducted.

Given the heterogeneity of the experimental designs, tasks, and measurement approaches identified, a quantitative meta-analysis of effects was not considered appropriate. Instead, the synthesis followed the principles of Synthesis Without Meta-analysis (SWiM) (Campbell et al., 2020), combining frequency distributions, contingency tables, and targeted inferential analyses of predefined methodological dimensions. This approach enabled systematic comparisons across different experimental architectures without assuming equivalence among heterogeneous measurement instruments or paradigms.

Eligibility criteria were established using a PICOS framework adapted to experimental research, seeking alignment between the review objective and the characteristics of the included evidence. Eligibility did not remain restricted to laboratory experiments or designs involving conventional experimental manipulation; the review also admitted quasi-experimental designs and natural experiments when they met the predefined criteria and enabled the behavioral identification of at least one of the cognitive biases under consideration through individual financial decisions.

Search Strategy

The search strategy followed a systematic and explicit approach by combining three conceptual blocks through Boolean operators:

Cognitive biases: terms corresponding to the eight predefined biases.

Decision-making context: terms related to financial and investment decisions.

Methodological design: terms corresponding to experimental and quasi-experimental approaches.

To focus retrieval on studies involving behavioral measurement, the original strategy used the NOT operator to exclude records retrieved exclusively through terms associated with surveys, questionnaires, or self-reports. Given the potential risk of excluding studies that combined these techniques with behavioral tasks, a sensitivity analysis was subsequently conducted by removing this operator. This search retrieved nine additional records from Web of Science and eleven from Scopus. However, these 20 additional records did not undergo an independent and documented eligibility assessment; therefore, it cannot be determined how many would have met the inclusion criteria or whether their assessment would have altered the composition of the final corpus. Consequently, the original strategy remained the basis of the review, but the

sensitivity analysis introduces residual uncertainty regarding the coverage of the bibliographic retrieval process, which should be considered when interpreting the scope of the corpus.

The search was conducted in two databases: Web of Science, restricted to the Social Sciences Citation Index (SSCI), and Scopus. The review included peer-reviewed journal articles published in English between 1990 and 2025. The complete Boolean search strings and their database-specific adaptations are provided in the Supplementary Material, Appendix A.

The retrieved records were initially exported to Zotero (v. 6.0) for duplicate identification and subsequently managed in Covidence throughout the screening process. The electronic search was supplemented by backward reference checking of relevant studies. The search was conducted on January 3, 2026.

Eligibility Criteria and Study Selection

The study selection process followed a sequential two-stage procedure. In the first stage, titles and abstracts were screened according to the inclusion criteria established through the adapted PICOS framework (Supplementary Material, Appendix C, Table A1). In the second stage, full-text articles were assessed by applying exclusion criteria related to experimental design, the context of individual financial decision-making, and the behavioral measurement of biases (Supplementary Material, Appendix C, Table B1). These criteria allowed the inclusion of experimental and quasi-experimental designs, including natural experiments when they met the predefined eligibility conditions.

A total of 319 records were identified: 279 through database searches (141 from Web of Science and 138 from Scopus) and 40 through citation searching. After duplicate removal, screening, and eligibility assessment, a final database reconciliation identified and removed one additional duplicate record, yielding a final corpus of 70 studies. The complete identification, screening, eligibility assessment, and inclusion process appears in the PRISMA flow diagram (Supplementary Material, Appendix B).

To assess the consistency of the selection process, two reviewers independently made the eligibility decisions. Inter-rater agreement during full-text assessment was evaluated using Cohen's Kappa coefficient (Cohen, 1960), yielding $\kappa = 0.83$, which corresponds to an almost perfect level of agreement according to the conventional criteria proposed by Landis and Koch (1977). The reviewers subsequently resolved discrepancies through discussion and consensus.

Data Extraction

Data extraction followed a structured coding scheme developed ex ante (Supplementary Material, Appendix C, Table C1), designed to record the descriptive, instrumental, and methodological characteristics of each study. The process was implemented in Covidence through independent coding by two reviewers, followed by discrepancy resolution.

To standardize classification decisions, the reviewers used a detailed coding guide (Supplementary Material, Appendix E) specifying the purpose, response categories, decision rules, and illustrative examples for each C1 variable.

The extracted information was subsequently exported to Microsoft Excel to construct the analytical database, frequency distributions, and contingency tables. The extraction variables were organized into three groups: general study characteristics, including publication year, region, sample, and experimental context; instrumental and technical characteristics, including the bias measured, instrument type, task type, and form of operationalization; and basic methodological characteristics, including comparison conditions, incentive structure, and level of methodological detail. Instrument types encompassed experimental tasks, economic games, simulations, and induced-value designs, the latter drawing on the induced-value tradition in experimental economics (Smith, 1976). The corresponding structure and coding rules are documented in the Supplementary Material.

This phase served a strictly descriptive function. Coding relied exclusively on information explicitly reported in the primary studies, without inferring undocumented practices, properties, or procedures. When the information required for a variable was unavailable, the category “Not reported” was assigned. This rule distinguished factual extraction from the subsequent methodological assessment.

The consistency of independent extraction was assessed using the unweighted Cohen’s Kappa coefficient, yielding $\kappa = 0.76$, which corresponds to a substantial level of agreement (Cohen, 1960; Landis & Koch, 1977). The operational rules and criteria specified in the coding guide standardized the application of the categories and reduced ambiguity during the extraction process.

The resulting database enabled systematic comparisons across the different experimental architectures and facilitated the identification of patterns in how cognitive biases were operationalized and measured.

Methodological Robustness Assessment and Procedural Reporting Audit

The assessment followed a three-component framework designed to distinguish descriptive extraction from methodological evaluation:

C1: descriptive data extraction.

D1: assessment of specific dimensions of methodological robustness.

E1: audit of procedural practice reporting.

C1 recorded the characteristics explicitly reported by the studies without assigning a quality rating. D1 assessed specific dimensions of the methodological robustness of the experimental instruments using the ordinal categories High, Moderate, and Low. The criteria covered objective clarity, causal appropriateness of the design, structural control, experimental manipulation, conceptual precision of the instrument, replicability, validity evidence, reliability

evidence, incentives, sample–target alignment, discussion of representativeness, and reporting of limitations (Supplementary Material, Appendix C, Table D1). Each criterion underwent an independent assessment; consequently, D1 did not serve to construct an overall robustness score, nor did the assessment assume that a weakness in one dimension automatically determined the rating assigned to any other dimension.

E1 complemented this assessment by auditing the explicit reporting of three procedural practices: piloting, technological support, and ethical protocol and anonymity (Supplementary Material, Appendix C, Table E1). The reviewers coded these indicators independently and did not interpret them as an aggregate quality scale. In particular, technological support represented a documentable characteristic of experimental implementation rather than, in itself, evidence of greater methodological quality.

The analysis maintained a fundamental distinction between extraction and evaluation. In C1, the absence of explicit information was recorded as “Not reported.” In D1, by contrast, the Low category indicates that the evidence available in the article did not provide sufficient support for a higher rating on the criterion assessed. Therefore, a Low classification should not automatically be interpreted as evidence that a practice or methodological property was absent during the original execution of the study. This distinction enabled the assessment of methodological robustness supported by the published information without equating absence of evidence with evidence of absence.

The reviewers established the decision rules for each criterion *ex ante* and applied them through coding guides specifying the purpose of each criterion, response categories, and conditions for its application. The methodological rationale and traceability of the C1–D1–E1 criteria appear in Supplementary Material, Appendix D, whereas the complete coding guides for C1, D1, and E1, including decision rules and illustrative examples, appear in Appendices E, F, and G, respectively. In particular, the assessment of incentives under D1–28 considered their salience and connection to experimental decisions, consistent with the methodological relevance of incentives in experimental economics (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith, 1976), without interpreting the mere presence of monetary incentives as a global indicator of quality.

Before fully implementing the framework, the reviewers conducted an internal calibration using five randomly selected studies to examine the interpretation of the criteria and harmonize their application. This procedure constituted a calibration stage for the coding process rather than an external validation of the framework.

Two reviewers independently conducted the assessment. Inter-rater consistency was estimated using the unweighted Cohen’s Kappa coefficient, yielding $\kappa = 0.76$ for D1 and $\kappa = 0.80$ for E1 (Cohen, 1960), corresponding to substantial levels of agreement according to the conventional criteria proposed by Landis and Koch (1977). These coefficients provide an overall measure of consistency in framework application between the reviewers and do not constitute evidence of external validation of the assessment instrument. For D1, given the ordinal nature of the categories, unweighted Kappa also does not account for the magnitude of discrepancies between categories. The reviewers subsequently resolved the identified discrepancies through discussion and consensus.

Evidence Integration and Analysis

The analytical strategy integrated the three complementary components: C1, D1, and E1, while maintaining a clear distinction between descriptive information and methodological assessment. First, frequency distributions and contingency tables were calculated to characterize the experimental instruments, participant profiles, application contexts, biases examined, and methodological dimensions assessed.

On this basis, four targeted association analyses (A1–A4) were defined: experimental instrument type $\times$ participant profile (A1); region $\times$ primary cognitive bias (A2); experimental context $\times$ discussion of representativeness (A3); and experimental instrument type $\times$ sample–target alignment (A4). These analyses used the final corpus of 70 studies, except for A2, whose analytical universe comprised 65 studies after excluding two studies classified as transregional and three studies without regional information, in accordance with the predefined rules. The specifications and derived categories used in these analyses appear in the Supplementary Material.

Four temporal analyses (T1–T4) were also defined to examine the distribution of selected methodological and reporting practices throughout the study period: publication period $\times$ piloting (T1); period $\times$ reliability evidence (T2); period $\times$ ethical protocol and anonymity (T3); and period $\times$ technological support (T4). For these analyses, the studies were grouped into four predefined time intervals. Percentage variations across periods were interpreted jointly with the inferential results and association magnitudes rather than treated independently as evidence of a temporal trend.

For contingency tables, the Pearson χ^2 test was used when its assumptions were adequately met. When expected frequencies proved insufficient in $R \times C$ tables, the Fisher–Freeman–Halton exact test was applied. Association magnitude was estimated using Cramér’s V . p -values and association magnitudes were interpreted jointly, avoiding the reduction of evidence to a dichotomous classification based exclusively on the statistical significance threshold.

The eight inferential contrasts, A1–A4 and T1–T4, constituted a single primary analytical family, and their p -values were jointly adjusted using the sequential Holm procedure with a two-sided $\alpha = .05$. Additionally, a sensitivity analysis applied the same Holm procedure separately to two conceptual families: A1–A4 and T1–T4. This specification enabled an assessment of the stability of the conclusions under an alternative definition of the contrast family. Unadjusted p -values were retained for transparency, whereas the primary inferential conclusions relied on the adjusted p -values.

Given the high concentration of High classifications observed for some D1 criteria, a further descriptive sensitivity analysis examined their influence on the integrative interpretation of methodological robustness. A proportion exceeding 90% of High classifications was descriptively defined as a severe ceiling effect. Criteria exceeding this threshold remained unchanged in D1, the coded database, and the original descriptive results, but the analysis did not use them as cumulative evidence in the integrative interpretation of methodological

robustness. No overall score was constructed, and no recoding was performed. The specification and results of this analysis appear in the Supplementary Material.

The inferential analyses were conducted in R version 4.6.1, using only base R functions. The Fisher–Freeman–Halton exact tests were computed with `fisher.test()` without Monte Carlo simulation; Pearson χ^2 tests with `chisq.test()`; and multiplicity adjustment with `p.adjust(..., method = "holm")`. Cramér’s V was calculated from the χ^2 statistic and the dimensions of the contingency table. The script reproduced the analytical transformations, contingency tables, statistical contrasts, and association magnitudes from the final database. The code, reproducibility guide, and execution-environment information are provided in the Supplementary Material.

Results

Evolution of Experimental Contexts

The 70 included studies were published between 1996 and 2025, within the predefined search period of 1990–2025. Their temporal distribution shows a greater concentration of studies in the more recent periods: 12 corresponded to 1996–2005, 17 to 2006–2014, 19 to 2015–2020, and 22 to 2021–2025.

The laboratory represented the predominant context during the first two periods, accounting for 9 of 12 studies in 1996–2005 and 12 of 17 in 2006–2014. In the subsequent periods, the descriptive presence of online experiments increased, accounting for 7 of 19 studies in 2015–2020 and 11 of 22 in 2021–2025. Across the full corpus, 34 studies (48.6%) were conducted in laboratory settings, 23 (32.9%) online, 8 (11.4%) used mixed contexts, and 5 (7.1%) constituted field studies. Overall, this distribution indicates greater descriptive diversification of experimental contexts in the more recent periods, although these variations should not, in isolation, be interpreted as evidence of a temporal trend.

Table 1. Evolution of Experimental Context by Publication Period

Period	Studies	Laboratory	Online	Mixed	Field
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

Source: Authors’ elaboration based on the final database, N = 70.

Experimental Instruments and Participant Profiles

The instruments used to measure cognitive biases were concentrated in two main architectures. Experimental tasks accounted for 36 studies (51.4%), while simulations accounted for 30 (42.9%); induced-value designs and economic games jointly represented four studies (5.7%). Regarding the specific type of experimental decision, simulated markets predominated (31 studies), followed by portfolio allocation (15), decisions under risk (8), other tasks (7), lottery choice (6), and financial forecasts (3). Measurement relied primarily on observed behavior (58 studies), compared with 11 studies that quantified bias intensity and one that used a presence measure.

Participants also exhibited a heterogeneous composition. Students constituted the most frequent group, appearing in 32 studies (45.7%), followed by the general population with 21 (30.0%), mixed samples with 9 (12.9%), and investors or professionals with 8 (11.4%). Analysis A1 examined the association between this distribution and the type of instrument used.

Table 2. Experimental Instrument Type by Participant Profile (A1)

Instrument type	Students	Investors / professionals	General population	Mixed	Total
Experimental task	20	0	10	6	36
Simulation	10	7	10	3	30
Other	2	1	1	0	4
Total	32	8	21	9	70

Source: Authors' elaboration based on the final database, N = 70.

The Fisher–Freeman–Halton exact test for A1 yielded an unadjusted p-value of .025, with a moderate association magnitude (Cramér's $V = .288$). However, after jointly considering the eight contrasts A1–A4 and T1–T4 and applying the Holm procedure, the association no longer provided sufficient statistical evidence at the $\alpha = .05$ level (adjusted $p = .2008$).

Descriptively, experimental tasks clustered particularly among student samples; among the eight studies involving investors or professionals, seven employed simulations, while one fell into the Other category. Accordingly, A1 reveals a descriptive pattern accompanied by a moderate association magnitude, but the inferential evidence does not support the conclusion that participant profile associates with instrument type after accounting for multiplicity adjustment.

Regional Distribution and Cognitive Biases

The evidence reveals considerable geographic concentration. Europe accounts for 28 of the 70 studies (40.0%), while North America accounts for 21 (30.0%), meaning that these two regions jointly represent 70.0% of the corpus. Asia/Oceania comprises 13 studies (18.6%), and Latin

America accounts for three (4.3%). In addition, two studies were classified as transregional, while three did not report sufficient information for regional classification.

Regarding the primary cognitive bias, the disposition effect accounted for 27 studies (38.6%), whereas loss aversion accounted for 20 (28.6%). The remaining 23 studies (32.9%) addressed the other biases considered. To examine the relationship between region and primary bias, A2 grouped these remaining biases into a common category and excluded from the inferential analysis only the two transregional studies and the three studies without reported regional information; accordingly, the analysis used $N = 65$.

No studies examined availability or mental accounting as the primary cognitive bias within the eligibility and coding criteria applied.

Table 3. Region by Primary Cognitive Bias (A2)

Region	Disposition effect	Loss aversion	Other biases	Total
North America	6	9	6	21
Europe	11	8	9	28
Asia/Oceania	8	2	3	13
Latin America	2	0	1	3
Total	27	19	19	65

Source: Authors' elaboration based on the final database, $N = 65$.

Although descriptive differences can be observed, for example, a relatively greater presence of loss aversion in North America and of the disposition effect in Asia/Oceania, the Fisher–Freeman–Halton exact test for A2 yielded an unadjusted p -value of .465 and an association magnitude of Cramér's $V = .214$. Following the joint adjustment of the eight contrasts A1–A4 and T1–T4 using the Holm procedure, the adjusted p -value reached 1.0000. Therefore, the observed descriptive differences do not provide sufficient statistical evidence to support an association between region and primary cognitive bias.

Methodological Assessment of the Instruments

The D1 assessment reveals differences across the dimensions examined. Five criteria directly related to experimental architecture and documentation show a high concentration in the High category: objective clarity, 68 studies (97.1%); structural control, 64 (91.4%); experimental manipulation, 69 (98.6%); conceptual precision of the instrument, 69 (98.6%); and instrument replicability, 64 (91.4%). Causal appropriateness shows a more differentiated distribution: 51 studies (72.9%) received a high classification, 17 (24.3%) a Moderate classification, and 2 (2.9%) a Low classification.

Table 4. Methodological Assessment of Experimental Instruments (D1)

Criterion	High	Moderate	Low
Objective clarity	68	2	0
Causal appropriateness	51	17	2
Structural control	64	6	0
Experimental manipulation	69	1	0
Conceptual precision of the instrument	69	1	0
Instrument replicability	64	6	0
Validity evidence	59	11	0
Reliability evidence	7	0	63
Incentives	47	6	17
Sample–target alignment	56	13	1
Discussion of representativeness	51	16	3
Reported limitations	38	30	2

Source: Authors' elaboration based on D1, N = 70.

Given the concentration exceeding 90% of High classifications for objective clarity, structural control, experimental manipulation, conceptual precision of the instrument, and replicability, these five criteria were considered dimensions with severe ceiling effects in the descriptive sensitivity analysis. In accordance with the predefined procedure, their original coding and results were retained without modification; however, these criteria were not used as cumulative evidence in the integrative interpretation of methodological robustness.

When this interpretation focused on the remaining criteria: causal appropriateness, validity evidence, reliability evidence, incentives, sample–target alignment, discussion of representativeness, and reported limitations, the methodological profile remained heterogeneous: some dimensions showed a high proportion of favorable ratings, whereas others, particularly reliability evidence, showed limited availability of explicitly documented evidence. Therefore, the integrative interpretation of the corpus's methodological profile does not depend on the accumulation of criteria with highly concentrated distributions. The complete sensitivity analysis appears in the Supplementary Material.

Sample-related dimensions also showed a concentration of High ratings. Sample–target alignment received a high classification in 56 studies (80.0%), Moderate in 13 (18.6%), and Low in one (1.4%). Discussion of representativeness yielded 51 High classifications (72.9%), 16 Moderate (22.9%), and 3 Low (4.3%). However, the two criteria follow different rationales: sample–target alignment assesses the correspondence between the sample used and the study objective or intended inference, whereas representativeness depends on the authors' explicit discussion of scope, external validity, or generalizability limitations. This distinction forms part of the final D1 decision rules.

Analysis A3 specifically examined whether discussion of representativeness differed according to experimental context. Among the 34 laboratory studies, 26 received a high classification and

8 received a Moderate/Low classification; among the 23 online studies, the corresponding frequencies were 15 and 8, respectively; all five field studies received a high classification; and among the eight studies using mixed contexts, five received a high classification and three received a Moderate/Low classification. The Fisher–Freeman–Halton exact test yielded an unadjusted p-value of .402, with Cramér’s $V = .214$. Following the joint adjustment of the eight contrasts using the Holm procedure, the adjusted p-value reached 1.0000. Accordingly, the observed descriptive differences do not provide sufficient statistical evidence to support an association between experimental context and discussion of representativeness.

Similarly, A4 assessed the relationship between instrument type and sample–target alignment. High classifications occurred in 29 of the 36 experimental tasks, 24 of the 30 simulations, and 3 of the 4 instruments grouped under Other. The Fisher–Freeman–Halton exact test yielded an unadjusted p-value of 1.000, with Cramér’s $V = .031$; the Holm-adjusted p-value was 1.0000. Therefore, the observed pattern does not provide sufficient statistical evidence to support an association between instrument type and sample–target alignment.

Validity, Reliability, and Incentive Evidence

The most pronounced differences within D1 emerge when comparing the reported evidence of validity and reliability. Validity evidence received a high classification in 59 studies (84.3%) and a Moderate classification in 11 (15.7%). In contrast, only 7 studies (10.0%) provided sufficient reliability evidence to receive a high classification, whereas the remaining 63 (90.0%) received a Low classification.

This difference should be interpreted in light of the specific nature of both criteria. In particular, a Low classification for reliability evidence does not demonstrate that an instrument lacks reliability. D1-27 exclusively assesses the level of reliability evidence explicitly documented in the article; therefore, the concentration in the Low category primarily reflects the limited availability of reported evidence that could support a higher classification. Consistent with this interpretation, in C1, 63 studies were coded as “Not reported” and only 7 as “Yes” regarding the existence of reliability evidence.

Incentives showed a more heterogeneous distribution: 47 studies (67.1%) received a high classification, 6 (8.6%) a Moderate classification, and 17 (24.3%) a Low classification. In the C1 descriptive extraction, 55 studies reported actual monetary incentives, 11 used hypothetical incentives, and 4 did not report the incentive scheme. The difference between these two distributions arises because D1-28 does not merely record the presence of incentives; rather, it methodologically assesses their salience and connection to experimental decisions.

Taken together, these findings show that distinct methodological dimensions exhibit different assessment profiles. Whereas validity evidence predominantly receives High classifications, explicitly documented reliability evidence remains limited, and incentives display a more heterogeneous distribution. This pattern reinforces the importance of distinguishing between the robustness that can be assessed from the available information and the extent to which specific methodological properties receive explicit documentation, without equating the two concepts.

Procedural Reporting Audit

The E1 audit reveals an uneven distribution across the three practices examined. Technological support was reported in 66 of the 70 studies (94.3%), whereas piloting was reported in 13 (18.6%) and ethical protocol and anonymity in 34 (48.6%).

Table 5. Procedural Practice Reporting Audit (E1)

Procedural dimension	Yes	No	Yes (%)
Piloting	13	57	18.6
Technological support	66	4	94.3
Ethical protocol and anonymity	34	36	48.6

Source: Authors' elaboration based on E1, N = 70.

These figures describe the evidence identified in the articles according to the established coding rules. Accordingly, a “No” classification for piloting or ethical protocol does not demonstrate that the practice necessarily remained absent during the original execution of the experiment; rather, it indicates that the practice could not be identified in the reported information. Likewise, technological support constitutes a documented procedural practice and should not, in itself, be interpreted as evidence of greater methodological quality or as evidence of instrument reliability. This distinction proves particularly relevant given the descriptive contrast between the high reporting rate of technological support (94.3%) and the proportion of studies receiving a high classification for reliability evidence (10.0%).

Temporal Evolution of Methodological Practices

Analyses T1–T4 examined whether specific methodological and reporting practices were associated with publication period. The results revealed distinct descriptive patterns across the four dimensions examined.

Table 6. Temporal Evolution of Selected Methodological Dimensions

Dimension	1996–2005	2006–2014	2015–2020	2021–2025
Piloting: Yes	0/12 (0.0%)	3/17 (17.6%)	2/19 (10.5%)	8/22 (36.4%)
Reliability evidence: High	0/12 (0.0%)	2/17 (11.8%)	1/19 (5.3%)	4/22 (18.2%)
Ethical protocol and anonymity: Yes	2/12 (16.7%)	7/17 (41.2%)	12/19 (63.2%)	13/22 (59.1%)
Technological support: Yes	10/12 (83.3%)	17/17 (100.0%)	17/19 (89.5%)	22/22 (100.0%)

Source: Authors' elaboration based on D1 and E1, N = 70.

Inferential contrasts provide context for these descriptive variations. In T1, the proportion of studies reporting piloting was higher in the most recent period, reaching 36.4% compared with 0.0% in the earliest period. The Fisher–Freeman–Halton exact test yielded an unadjusted p-value of .0508 and Cramér’s $V = .342$. After jointly adjusting the eight contrasts A1–A4 and T1–T4 using the Holm procedure, the adjusted p-value was .3437. Accordingly, although the descriptive pattern and association magnitude suggest differences across periods that warrant consideration, the adjusted inferential evidence does not support a statistically substantiated temporal association.

For T2, the Fisher–Freeman–Halton exact test yielded an unadjusted p-value of .4368, with Cramér’s $V = .224$; following the Holm adjustment, the p-value reached 1.0000. Although 2021–2025 presented the highest proportion of High classifications (18.2%), the sequence 0.0%–11.8%–5.3%–18.2% does not follow a monotonic pattern, and the inferential evidence does not support an association between publication period and reported reliability evidence.

For T3, reporting of ethical protocol and anonymity increased descriptively from 16.7% in 1996–2005 to 41.2% in 2006–2014 and 63.2% in 2015–2020, reaching 59.1% in 2021–2025. The Pearson χ^2 test yielded $\chi^2(3) = 7.86$, with an unadjusted p-value of .0491 and Cramér’s $V = .335$. However, following the joint Holm adjustment, the adjusted p-value was .3437. Thus, the data show a higher descriptive frequency of reporting these safeguards in the more recent periods compared with the initial period, accompanied by an association of appreciable magnitude; however, the adjusted inferential evidence does not support a statistically substantiated temporal association. The pattern also does not follow a strictly monotonic trajectory because reporting declined slightly between the final two periods.

Finally, T4 shows that reporting of technological support remained high throughout virtually the entire study period: 83.3%, 100.0%, 89.5%, and 100.0%, respectively. The Fisher–Freeman–Halton exact test yielded an unadjusted p-value of .0603 and Cramér’s $V = .289$; following the Holm adjustment, the adjusted p-value was .3437. Therefore, the adjusted inferential evidence does not support a temporal association in a context characterized by a high frequency of technological support reporting across all four periods.

Integrated Synthesis of Results

Overall, the findings reveal an experimental literature characterized by a high prevalence of experimental tasks and simulations, descriptive diversification of application contexts in more recent periods, and geographic concentration primarily in Europe and North America. Across analyses A1–A4, none of the four contrasts provided sufficient statistical evidence of an association after multiplicity adjustment. A1 nevertheless revealed a descriptive pattern between instrument type and participant profile, accompanied by a moderate association magnitude (Cramér’s $V = .288$), although this evidence did not remain inferentially supported after the Holm adjustment. Analyses A2–A4 likewise provided insufficient statistical evidence to support associations between region and primary cognitive bias, experimental context and discussion of representativeness, or instrument type and sample–target alignment, respectively.

The D1 assessment revealed differentiated profiles across methodological dimensions. Five criteria related to experimental architecture and documentation showed concentrations exceeding 90% of High classifications and were therefore examined through a descriptive sensitivity analysis. After excluding these criteria solely from the integrative interpretation, without modifying their coding or original results, the remaining criteria continued to exhibit a heterogeneous methodological profile: validity evidence predominantly received High classifications, whereas explicitly documented reliability evidence showed a marked concentration in the Low category. Other dimensions, such as incentives, sample–target alignment, discussion of representativeness, and reported limitations, displayed more differentiated distributions. These findings indicate that interpretation of the corpus’s methodological profile does not depend on the accumulation of criteria affected by ceiling effects.

The procedural reporting audit further reinforces this heterogeneity: technological support was widely documented, whereas piloting and ethical protocol and anonymity appeared less frequently. The temporal analyses T1–T4 likewise provided insufficient statistical evidence of associations after joint multiplicity adjustment. Nevertheless, T1 and T3 revealed descriptive patterns and association magnitudes that warrant consideration: piloting was reported more frequently in the most recent period (Cramér’s $V = .342$), while reporting of ethical protocol or anonymity occurred more frequently in recent periods than in the initial period (Cramér’s $V = .335$), although neither pattern retained sufficient inferential evidence after the Holm adjustment.

Taken together, the findings reveal differences between the methodological robustness that can be assessed from published information and the extent to which specific methodological practices and properties receive explicit documentation, without providing evidence of a generalized temporal improvement in these dimensions. These findings provide the basis for their conceptual interpretation in the discussion.

Discussion

From Psychometrics to Decision Architecture

The results show that the experimental measurement of cognitive biases in financial decision-making relies predominantly on architectures designed to observe behavioral choices and responses. Experimental tasks and simulations jointly account for 66 of the 70 studies analyzed (94.3%), encompassing simulated markets, portfolio decisions, lotteries, risk-based tasks, and other configurations. This predominance aligns with the tradition of experimental economics, which investigates decision-making mechanisms through structured environments and controlled conditions (Harrison & List, 2004; Hertwig & Ortmann, 2001; Levitt & List, 2007).

However, the predominance of these architectures does not imply measurement homogeneity. Different elicitation procedures can yield measures that do not necessarily demonstrate equivalence, even when they seek to approximate related constructs (Crosetto & Filippin, 2016; Frey et al., 2017; Pedroni et al., 2017). From a construct validity perspective, this heterogeneity matters because correspondence between a theoretical construct and its operationalization

depends on whether the procedure adequately represents the phenomenon it seeks to measure (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995). This issue becomes particularly evident in biases such as overconfidence, whose distinct manifestations do not necessarily represent a unitary phenomenon (Moore & Healy, 2008).

Accordingly, experimental instruments can be understood not merely as recording mechanisms but also as components of the architecture through which biases become observable. The findings of this review do not establish that one architecture produces more valid measurements than another; however, they show that operationalization varies across studies. Consequently, comparing evidence requires consideration not only of the bias under investigation but also of the task, context, and rules through which researchers elicit behavioral responses.

Gap Between Methodological Robustness and Transparency

The methodological assessment reveals differences between dimensions that researchers can evaluate from the described design and those that depend more heavily on explicitly documented evidence. Experimental manipulation and conceptual precision received High classifications in 98.6% of the studies, while structural control and replicability reached 91.4%. However, these four dimensions, together with objective clarity, showed concentrations exceeding 90% in the High category, indicating severe ceiling effects and potentially limited discriminative capacity within the corpus. For this reason, the sensitivity analysis excluded them only from the integrative interpretation, without modifying their coding or original results. When the interpretation focuses on the remaining criteria, a heterogeneous methodological profile persists; for example, causal appropriateness received a high classification in 72.9% of the studies, whereas other dimensions displayed more differentiated distributions.

The pattern becomes particularly pronounced for certain measurement properties. Validity evidence received a high classification in 84.3% of the studies, whereas only 10.0% provided sufficient reliability evidence to reach that category. The remaining 90.0% received a Low classification, but this finding does not imply that the instruments lack reliability. Under the D1 decision rules, it indicates that the published information did not provide sufficient explicit evidence to support a higher classification. This distinction proves particularly important given the heterogeneity of the instruments identified, as prior or repeated use of a task does not substitute for documentation of its measurement properties (Cronbach & Meehl, 1955; Flake & Fried, 2020; Messick, 1995).

A similar asymmetry emerges in the reporting of procedural practices. Technological support appeared in 94.3% of the studies, whereas piloting was reported in 18.6% and ethical protocol and anonymity in 48.6%. Again, the absence of reporting does not demonstrate that these practices did not occur. Likewise, technological support can facilitate standardized administration and decision recording, but it does not, in itself, constitute evidence of reliability or greater methodological quality.

Taken together, these findings reveal a distinction between the dimensions of methodological robustness that can be assessed from published information and the extent to which specific methodological properties and practices receive explicit documentation. This interpretation

remains consistent when the D1 criteria affected by ceiling effects are considered separately and therefore does not depend on the accumulation of favorable classifications in dimensions with limited discriminative capacity. This distinction does not constitute a global measure of quality, nor does it allow a methodological deficiency to be inferred when particular information has not been reported. The temporal analyses complement this interpretation by providing no evidence of a generalized and uniform improvement across the dimensions examined.

Experimental Design, Participants, and Ecological Validity

Analysis A1 revealed a descriptive pattern between instrument type and participant profile: experimental tasks were used primarily in studies involving students, whereas simulations accounted for most studies involving investors or professionals. The association magnitude was moderate (Cramér's $V = .288$), and the unadjusted p -value was .025; however, the inferential evidence did not remain supported after multiplicity correction using the Holm procedure (adjusted $p = .2008$). Therefore, the observed pattern holds methodological relevance for characterizing the corpus but does not support a statistically substantiated association between instrument type and participant profile.

Experimental literature has emphasized the balance between control and contextualization: laboratory environments allow researchers to isolate mechanisms under controlled conditions, whereas more contextualized designs can approximate specific decision-making environments (Falk & Heckman, 2009; Harrison & List, 2004; Levitt & List, 2007). Within this framework, the observed pattern between participants and instruments could align with different contextualization requirements; however, our data do not allow us to attribute such an intention to the researchers or establish the mechanism that might explain this distribution.

Likewise, A3 provided insufficient statistical evidence of an association between experimental context and discussion of representativeness (Cramér's $V = .214$), while A4 provided insufficient evidence of an association between instrument type and sample–target alignment (Cramér's $V = .031$). Accordingly, the pattern observed in A1 should not be extrapolated to systematic differences in these other dimensions.

Experimental contexts also show a relatively greater presence of online studies in more recent periods. However, this pattern remains descriptive and should not be interpreted as a statistically demonstrated trend. Complementarily, reporting of technological support remained high throughout virtually the entire study period. Although T4 yielded an association magnitude of Cramér's $V = .289$, the inferential evidence did not reach sufficient support after multiplicity adjustment. Taken together, these findings do not support a linear interpretation whereby experimental infrastructure progressively advanced while other practices remained lagging.

Toward a Tripartite Framework of Transparency

The identified patterns allow methodological transparency to be organized into three complementary dimensions: technological, measurement and statistical, and ethical transparency. This proposal does not constitute a validated scale or normative standard but

rather an analytical perspective for identifying the information needed to understand and evaluate the implementation of experimental instruments.

Technological transparency encompasses documentation of the platforms, software, and procedures used to administer tasks and record decisions. This dimension received the most extensive procedural reporting, although its presence should not be conflated with evidence concerning measurement properties.

Measurement and statistical transparency encompass documentation of instrument operationalization, evidence supporting its measurement properties, particularly validity and reliability, and the relevant statistical procedures used to analyze and interpret the results. The marked difference observed between these dimensions shows that a technically reproducible experimental architecture does not guarantee equally comprehensive documentation of the properties of the measures employed. Reproducibility research has emphasized the importance of explicitly reporting the procedures and evidence required to assess and replicate experimental findings (Camerer et al., 2016, 2018; Christensen & Miguel, 2018; Open Science Collaboration, 2015).

Ethical transparency encompasses reporting of consent, anonymity, confidentiality, and other safeguards applicable to participants. The temporal analyses did not provide sufficient statistical evidence of an association after multiplicity adjustment for any of the dimensions examined. Nevertheless, differentiated descriptive patterns emerged. In T3, reporting of ethical protocol and anonymity occurred more frequently in recent periods than in the initial period, with an association magnitude of Cramér's $V = .335$. Similarly, T1 showed a higher frequency of piloting reports in the most recent period and a comparable association magnitude (Cramér's $V = .342$). In both cases, the inferential evidence did not remain supported after the Holm adjustment. T2 and T4 likewise provided insufficient statistical evidence of temporal associations. Therefore, the results do not support a generalized temporal improvement in transparency, although they reveal differentiated descriptive patterns across the dimensions examined.

Taken together, these three dimensions complement the assessment of experimental architecture by distinguishing how an instrument is implemented, what evidence supports its measurement, and which procedural safeguards researcher's document. The tripartite framework should therefore be understood as a conceptual proposal derived from the identified patterns and subject to subsequent validation.

To translate this framework into a more operational reporting guide, Table 7 summarizes the information that future experimental studies should consider documenting within each transparency dimension. These elements provide prospective reporting guidance rather than a validated assessment scale or normative standard; accordingly, they should not be interpreted as an extension of the D1–E1 instruments used in this review.

Table 7. Proposed Reporting Guide for the Tripartite Transparency Framework

Transparency Dimension	Information That Should Be Documented
Technological Transparency	Experimental platform or software; mode of administration; relevant technological configuration of the experiment; technological procedures used to control or monitor participants.
Measurement and Statistical Transparency	Operationalization of the cognitive bias; scoring or classification rules; evidence or justification of validity and reliability, when applicable; piloting or calibration procedures; relevant statistical procedures.
Ethical Transparency	Informed consent procedures; safeguards for anonymity or confidentiality; data protection; ethical review or approval, when applicable; treatment of incentives and potential risks to participants.

Source: Authors' elaboration.

Contribution to the Literature

This review makes three main contributions. First, it shifts the focus from identifying cognitive biases to examining how they are operationalized experimentally. This approach shows that the same construct can be studied through different tasks, contexts, and populations, making measurement architecture a relevant element for interpreting and comparing the evidence.

Second, it identifies differences between methodological robustness dimensions that can be assessed from published information and the degree to which certain methodological properties and practices are explicitly documented. This interpretation remains consistent when considering separately the D1 criteria affected by ceiling effects and reveals differentiated profiles across dimensions such as causal appropriateness, validity and reliability evidence, incentives, sample–target alignment, discussion of representativeness, and reported limitations. The A1–A4 and T1–T4 analyses also distinguish descriptive patterns from inferential evidence: none of the eight contrasts provided sufficient statistical evidence of association after multiplicity adjustment, although A1, T1, and T3 showed descriptive patterns and association magnitudes relevant to characterizing and interpreting the corpus.

Third, the C1–D1–E1 structure separates descriptive extraction, methodological assessment, and procedural auditing, thereby reducing the risk of automatically interpreting the absence of reported information as evidence concerning the practices actually implemented. The framework was developed specifically for this review through the integration and operationalization of established methodological principles from systematic review methodology, experimental design, measurement, experimental economics, reproducibility, and research ethics. Its application yielded substantial levels of inter-rater agreement; however, the framework did not undergo formal external validation, and its results should therefore be interpreted in light of this characteristic.

Limitations and Future Research

This review has several limitations. First, the assessment depends on the information reported in the published studies. Therefore, the absence of evidence concerning reliability, piloting, ethical safeguards, or other practices may reflect either their actual absence or an omission in reporting. Accordingly, the corresponding categories should be interpreted as limitations in evaluability and transparency rather than automatically as deficiencies in study implementation.

Second, the search was restricted to articles published in English between 1990 and 2025 and indexed in Web of Science–SSCI or Scopus. Consequently, studies indexed exclusively in other specialized bibliographic databases, such as EconLit or PsycINFO, may not have been retrieved. In addition, the search retained a NOT operator intended to exclude studies based exclusively on surveys or self-reports. The sensitivity analysis conducted without this operator retrieved nine additional records from Web of Science and eleven from Scopus; because these records did not undergo an independently documented eligibility assessment, it cannot be determined whether any would ultimately have met the inclusion criteria. This decision constitutes a limitation of the retrieval strategy; therefore, the conclusions should be interpreted in relation to the corpus effectively identified through the search strategy applied.

Third, C1–D1–E1 was developed specifically for this review and, although the framework underwent internal calibration and was independently applied by two reviewers, it has not undergone external validation. The reported Kappa coefficients represent unweighted measures of overall agreement and, for ordinal assessments, do not distinguish the magnitude of discrepancies between categories. In addition, five D1 criteria showed concentrations above 90% in the High category, limiting their discriminant capacity within the corpus. The descriptive sensitivity analysis showed that the integrative interpretation remains consistent when these criteria are considered separately, without modifying the original codings; nevertheless, ceiling effects should be considered when interpreting the performance of the framework and represent an aspect that future applications and validation processes should examine.

Fourth, the A1–A4 and T1–T4 analyses examine associations between study characteristics and cannot establish causal mechanisms. Moreover, because multiple contrasts were performed, inferential evidence was evaluated using the Holm procedure. None of the eight analyses retained sufficient statistical evidence of association after this adjustment, although some showed descriptive patterns and association magnitudes of interest. Accordingly, these findings should be interpreted by integrating adjusted inferential evidence, association magnitudes, and observed descriptive patterns rather than through a dichotomous classification based exclusively on a significance threshold.

Future research could directly compare different instruments designed to measure the same bias under equivalent conditions, assessing their stability across tasks, contexts, and participant profiles. It would also be relevant to extend the analysis to approaches deliberately excluded from this review, such as instruments based exclusively on self-reports, and to externally evaluate both the C1–D1–E1 framework and the proposed tripartite transparency framework. In particular, future applications of the framework could examine the discriminant capacity of the

D1 criteria and evaluate their properties through independent validation procedures. Such work would help determine the extent to which the heterogeneity of experimental architectures identified in this review translates into actual differences in the validity, reliability, and comparability of experimental measures.

Conclusions

This study provides a systematic and methodological assessment of experimental instruments used to measure cognitive biases in financial decision-making. Based on the analysis of 70 studies, the findings show a clear predominance of experimental tasks and simulations, as well as differentiated profiles across the methodological dimensions assessed. Some dimensions of experimental architecture and documentation showed a high concentration of High classifications; however, five of them exhibited severe ceiling effects, and their contribution to the integrative interpretation was therefore examined through a sensitivity analysis.

When considering this sensitivity, marked heterogeneity remains across the remaining methodological dimensions. While some dimensions predominantly receive favorable ratings, others that depend on explicitly reported evidence or documentation show greater limitations, particularly reliability evidence. These differences do not necessarily imply deficiencies in practices that were not documented, nor do they support the construction of an overall measure of methodological quality. Likewise, the analyses conducted do not provide evidence of a generalized and uniform temporal evolution of the methodological and reporting practices examined.

Based on these findings, the study proposes a tripartite transparency framework that integrates technological, measurement and statistical, and ethical dimensions as complementary perspectives for examining the methodological information provided by experimental studies. This framework constitutes a conceptual and analytical proposal rather than a validated normative standard. It distinguishes documentation of experimental implementation, evidence supporting measurement properties, and reporting of safeguards applicable to participants.

Overall, the main contribution of this study consists of shifting the analytical focus from which cognitive biases researchers' study to how they measure them experimentally. The findings show that instruments do not constitute homogeneous measurement mechanisms but rather methodological architectures through which researchers operationalize and observe cognitive biases. Therefore, interpreting and comparing the evidence requires consideration not only of the bias under study but also of the characteristics of the tasks, experimental contexts, and procedures used to obtain behavioral responses.

Accordingly, strengthening transparency in the documentation of experimental instruments and procedures can facilitate the assessment of their methodological properties and improve the comparability, reproducibility, and accumulation of evidence in experimental behavioral finance. In particular, more explicit documentation of measurement properties and procedural practices would allow researchers to assess more precisely the similarities and differences among the various architectures used to study cognitive biases.

References

- 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>