IJAAF; Vol. 10, No. 3, Summer 2026, Serial 36

Iranian Journal of
Accounting, Auditing & Finance

Ferdowsi University of Mashhad

Quarterly

RESEARCH ARTICLE

# Financial Stress Research Under the Magnitude Mind Lens

**Taqi Al Abdwani**

*Department of Business and Management Studies, Faculty of Management, Gulf College, Muscat, Oman*
**Shima Ebrahimi***
*Department of Persian Language and Literature, Faculty of Letters and Humanities, Ferdowsi University of Mashhad, Mashhad, Iran*
**Iliya Pishghadam**
*Department of Management, Faculty of Economics and Administrative Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*

**How to cite this article:**
Al Abduwani, T. , Ebrahimi, S. and Pishghadam, I. (2026). Financial Stress Research Under the Magnitude Mind Lens. Iranian Journal of Accounting, Auditing and Finance, 10(3), 201-221. doi: 10.22067/ijaaf.2026.48223.1647
https://ijaaf.um.ac.ir/article_48223.html

ARTICLE INFO

Article History
Received: 2025-12-05
Accepted: 2026-03-16
Published online: 2026-07-12

**Keywords:**
Epistemic Humility, Financial Stress; Hubris, Magnitude Mind Model, Systematic Review

## Abstract

This study applied the Magnitude Mind Model (MMM) to evaluate the quality of cross-sectional financial stress literature beyond traditional methodological checklists. The MMM integrates magnitude thinking (SICHA: Scale, Impact, Complexity, Holism, Accuracy) with measures of intellectual humility (acknowledging uncertainty, context, and specific limitations) and hubris (simple causal claims, overgeneralization, false certainty). Following PRISMA 2020 guidelines, we consecutively selected the first 20 open-access English-language, Scopus Q1/Q2-indexed cross-sectional studies on financial stress published between 2015 and 2025 from a pool of 670 records. Two independent raters scored each paper on the 11 MMM sub-items, achieving excellent inter-rater reliability. Results suggested that, within this exploratory sample, only 10% of papers attained a high Magnitude Mind score, whereas 45% scored low. Hubris was the strongest correlate of MM scores in exploratory analyses. Low-MM papers frequently used causal language such as "causes" and "impacts" from cross-sectional data, overgeneralized findings from convenience samples, and presented generic limitation statements. Holism was the weakest SICHA sub-item, with 50% of papers scoring low. Analytical methods did not guarantee epistemic quality; two methodologically complex papers scored low due to extreme hubris. Temporal analysis revealed no significant improvement from 2015 to 2025. Within this sample, 45% of papers were classified as Low MM. The MMM may offer a potentially useful diagnostic and prescriptive tool for researchers, reviewers, and editors to promote more rigorous, humble, and credible science.

https://doi.org/10.22067/ijaaf.2026.48223.1647



NUMBER OF REFERENCES
42

NUMBER OF FIGURES
3

NUMBER OF TABLES
11

Homepage: https://ijaaf.um.ac.ir
E-ISSN: 2717-4131
P-ISSN: 2588-6142

*Corresponding Author: Shima Ebrahimi
Email: shimaebrahimi@um.ac.ir
Tel: 09155039738
ORCID: 0000-0003-3528-1386

## 1. Introduction

Financial worry, or the state of concern, anxiety, and feeling overwhelmed related to economic needs and their fulfillment (Van Dijk et al., 2022), is one of the most prevalent topics in health care policies and in 21st-century public policy and literature globally. According to the American Psychological Association (2022), 72% of Americans experience financial stress regularly, and one in four people rate their financial stress as extreme. These have been documented across Europe, Asia, and the Middle East, indicating financial stress regardless of income or education level (Nasr et al., 2024; Huang et al., 2020). This crisis was further aggravated by the worldwide COVID-19 pandemic that threw millions into unemployment, housing insecurity, and financial poverty in what would be one of its worst years (Thayer and Gildner, 2021).

Financial stress affects more than just a bank account. In mental health, financial distress is consistently associated with depression (Mikkonen and Raphael, 2010), anxiety disorders, and suicidal ideation (Richardson et al., 2017). Those under high financial stress are twice to three times more likely than those not meeting clinical criteria for major depressive disorder (French and McKillop, 2017). Financial stress has been associated with hypertension, cardiovascular disease, and accelerated biological ageing in physical health (Swigonski et al., 2021). When it comes to cognitive function, financial stress consumes mental bandwidth, which diminishes executive function and impulse control, a phenomenon known as the "scarcity trap" (Mullainathan and Shafir, 2013).

The related social and economic costs are equally deep. Campaigns reveal that financially stressed students are at high risk of dropping out, spending more time completing degrees, and experiencing lower perceived academic achievement in educational settings (Smathers et al., 2024). Financial hardship is linked to higher levels of absenteeism and presenteeism (being physically present but psychologically unsuited), lower performance, and higher turnover rates at work (Kim et al., 2003). Financial stress is one of the essential predictors for family dysfunction, marital conflict, domestic violence, child neglect, and adverse childhood experiences (Liu and Merritt, 2018), with lifelong developmental effects.

However, a critical research gap persists despite the rapid expansion of quantitative research on financial stress over the past decade. The epistemic quality of this literature, the extent to which researchers acknowledge uncertainty, avoid overstatement, engage in genuine self-criticism, appropriately limit their claims, and balance intellectual humility against the pull of overconfidence, has never been systematically evaluated.

This gap is consequential. First, the replication crisis (largely investigated in experimental psychology and economics) has highlighted overstated evidence and low statistical power (Camerer et al., 2018). The design of cross-sectional investigations differs, despite similar concerns about overgeneralization and causal overstatement being expressed (Grosz et al., 2020; Rohrer, 2018). Second, many authors use causal verbs such as "causes," "impacts," or "leads to," but cross-sectional designs, the most common approach in financial stress research, cannot demonstrate causality (Rohrer, 2018; Grosz et al., 2020). One of the most pernicious problems with academic writing based entirely on cross-sectional data is allowing a study to say "financial stress causes depression" when it could equally say that "depression causes financial difficulty". Third, the majority of studies take an individual-level approach to financial stress, neglecting the nested structure of households, communities, and policy environments – a reductionism that Bronfenbrenner's ecological systems theory (Bronfenbrenner 1979), for example, addresses by showing that the characteristics of the neighborhood, local labor market conditions, and cultural norms regarding debt and saving all shape financial stress (Wilson et al., 2023).

Not extra data, but a new perspective on something like research quality; not length of the checklist (PSI likes sample size and report significance), but more attitudes of researchers toward their own

findings (Pishghadam, 2026). This need results in the concept of magnitude thinking (Pishghadam, 2025), a multidimensional cognitive framework focused on proportional judgment. In organizing implied proportional judgment across five core dimensions, including Scale, Impact, Complexity, Holism, and Accuracy (SICHA), this cognitive stance has since been formalized into two complementary domains. The first, Magnitude Research (Pishghadam, 2026), uses the framework to investigate intricate phenomena through multiple lenses and scales. The second, Magnitude Psychology, which is detailed below (Pishghadam and Ebrahimi, 2026), develops a mentalistic theory that predicts the cognitive-developmental mechanisms underlying proportional judgment. Based on this groundwork, Pishghadam and Ebrahimi (2026) unveiled the Magnitude Mind Model (MMM), an application of magnitude reasoning that combines SICHA variables with epistemic humility measures (recognition of uncertainty, contextfulness and limitations) and hubris assumptions (overgeneralization, direct causal claims, and false certainty). In a previous systematic review of advancements in burnout research, the MMM was preliminarily validated (Pishghadam and Ebrahimi, 2026).

Financial stress research is especially prone to three epistemic pitfalls: (a) causal overstatement from cross‑sectional data, (b) overgeneralization from convenience samples, and (c) neglect of systemic (household/policy) contexts. The MMM directly addresses these pitfalls by operationalizing hubris (causal claims, overgeneralization) and holism (systemic factors). Thus, the MMM provides a tailored diagnostic tool for this body of literature. The present study applies this MMM to a consecutive sample of cross-sectional financial stress research published between 2015 and 2025 with three objectives: first, to determine the distribution of Magnitude Mind scores across this literature; second, to identify which sub-items most strongly discriminate between high- and low-quality research; and third, to explore whether scores vary by design complexity or publication year.

## 2. Theoretical framework

Financial stress is theoretically understood through multiple lenses. The Family Stress Model (Conger et al., 1990) suggests that economic pressure contributes to parental distress, which, in turn, compromises family functioning. One idea in the other direction, called The Scarcity Trap (Mullainathan and Shafir, 2013), is that financial scarcity consumes cognitive bandwidth, reducing executive control. Moreover, the Conservation of Resources (COR) theory (Hobfoll, 2001) highlights that financial loss threatens core resources, thereby posing a stressor. These frameworks justify the need for magnitude thinking, as financial stress operates across individual, household, and policy levels.

### *2.1. The Magnitude Mind Model (MMM)*

The Magnitude Mind Model (MMM) integrates SICHA with the assessment of epistemic virtues (e.g., humility) and vices (e.g., hubris). Table 1 presents the operational definitions of the MMM components with concrete examples from the financial stress literature.

SICHA captures magnitude thinking across five dimensions: Scale (sample adequacy, power, representativeness), Impact (practical or theoretical importance), Complexity (acknowledgement of multiple causal factors and interactions), Holism (systemic context, long-term effects), and Accuracy (effect sizes, confidence intervals, causal precision). Epistemic Humility comprises three protective factors: acknowledgement of uncertainty (e.g., stating that cross-sectional data cannot support causal inferences), acknowledgement of context (e.g., recognizing that results may differ by population or setting), and provision of specific limitations (e.g., concrete design or measurement weaknesses rather than generic disclaimers). Hubris comprises three risk factors, scored on a 0–5 scale with explicit

gradation:

**Table 1.** Operational definitions of MMM components with examples from the financial stress literature

| Component | Sub-item | Low (0-2) example | High (4-5) example |
|---|---|---|---|
| SICHA | Scale | N<200, no power analysis | N≥1000, a priori power analysis |
| | Impact | No practical implication is stated | Clear policy or clinical relevance |
| | Complexity | Single predictor (e.g., only income) | Multiple predictors with interactions |
| | Holism | Individual level only | Includes household, community, and policy |
| | Accuracy | Only p-values reported | Effect sizes, CIs, sensitivity analyses |
| Humility | Acknowledges uncertainty | More research is needed (generic) | Cross-sectional design cannot infer causality; unmeasured confounding likely remains. |
| | Acknowledges context | No discussion of context | Explicitly notes cultural/economic limits (e.g., Dutch sample may not generalize) |
| | State-specific limitations | No limitations or only a small sample | List 4-5 concrete limitations (design, measurement, sampling, confounding) |
| Hubris | Simple causal claims (low hubris = low score) | Uses associational language ("associated with", "correlated with"), which is good | Uses strong causal verbs without caveats ("causes", "leads to", "drives"), which is bad |
| | Overgeneralization (low hubris = low score) | Claims are limited to the study sample | Generalizes from one university to "all college students" without caveats |
| | False certainty (low hubris = low score) | Tentative language ("suggests", "indicates") | Certainty-marking terms ("proves", "definitively", "no doubt") |

Note: For Hubris sub−items, "Low" means a score of 0−2 (low hubris, desirable) and "High" means a score of 4−5 (high hubris, undesirable).

**Table 2.** Conceptual differentiation of MMM components

| Concept | Definition | Low-MM example (high hubris/low humility) | High-MM example (low hubris/high humility) |
|---|---|---|---|
| Epistemic humility | Acknowledges uncertainty, context, and specific limitations | More research is needed" (generic) | Our cross-sectional design cannot infer causality; unmeasured confounding likely remains. |
| Hubris (simple causal claims) | Causal verbs from weak designs | Financial stress causes depression | Financial stress is associated with depression. |
| Hubris (overgeneralization) | Claims beyond the sample | These results apply to all college students | These results may generalize to similar universities, but caution is warranted. |
| False certainty | Certainty-marking terms | These results definitively show… | These results suggest… |

Simple causal claims from weak designs (0–5): A score of 0–1 is assigned when authors use only associational language ("is associated with," "is correlated with," "predicts"). A score of 2–3 is assigned when authors use weaker causal language ("may affect," "could impact") while explicitly acknowledging assumptions (e.g., "subject to unmeasured confounding"). A score of 4–5 is assigned when authors use strong causal verbs ("causes," "leads to," "drives") without any caveats or from purely cross-sectional data with no identification strategy.

Overgeneralization beyond the study population (0–5): 0–1 = no overgeneralization; 2–3 = minor overgeneralization with some qualification; 4–5 = strong overgeneralization (e.g., from one university to "all college students" without caveats).

False certainty (0–5): 0–1 = tentative language (‘suggests’, ‘indicates’, ‘may’); 2–3 = moderately confident but acknowledges uncertainty (e.g., ‘likely’, ‘probable’); 4–5 = certainty−marking terms (‘proves’, ‘definitively’, ‘conclusively’, ‘no doubt’, ‘demonstrates unequivocally’). Overgeneralization (0–5): 0–1 = no overgeneralization, claims limited to sample; 2–3 = minor overgeneralization with some qualification (e.g., ‘may generalize to similar populations’); 4–5 = strong overgeneralization without caveats (e.g., from one university to ‘all college students’ or from one country to ‘worldwide’).

Importantly, the MMM does not penalize all causal language. Causal claims that are properly qualified, accompanied by a discussion of assumptions, and made within appropriate research designs (e.g., instrumental variables, natural experiments) are scored as low hubris. The high hubris scores in our sample were reserved for papers that made unqualified causal claims based solely on cross-sectional or time-series data, without any causal identification strategy.

Scale refers to the adequacy of sample size, duration, or scope relative to the claims being made. A study with N = 100 cannot support claims about "all college students", just as a study with N = 10,000 cannot compensate for a lack of humility. Scale also encompasses statistical power, the probability of detecting an effect when it truly exists. Impact is about the translational or theoretical importance of the finding to practice. Thus, a statistically significant result that has a small effect may not translate into strong policy recommendations. Research impact could be measured both in terms of effect size (e.g., Cohen's, 1985) and practical utility or disease burden reduction (e.g., number needed to treat, cost per healthy life year). Complexity captures the idea that multiple causal factors, interactions, or nuances are involved. After all, financial pressure rarely has a single cause; any study that looks at income in isolation while neglecting debt, savings and other wealth, employment, social support, and personality traits is simplistic. Holism takes into account wider systems, context, and longer-term effects; for example, financial stress is situated within homes, neighborhoods, labor markets, policy environments, and cultural contexts (Bronfenbrenner 1979). Research that does not consider these levels is inadequate. Accuracy indicates whether claims are factually true and verifiable. Reporting confidence intervals, effect sizes, and measures of uncertainty (rather than just p-values) is essential for accurate interpretation.

### 2.1.1 SICHA scoring anchors

Scale (0−5): 0 = N<50, no power analysis; 1 = N=50−99; 2 = N=100−199; 3 = N=200−499 with basic power discussion; 4 = N=500−999 with explicit power; 5 = N≥1000 with a priori power. Impact (0−5): 0 = no practical/theoretical implication; 1 = trivial; 2 = minor; 3 = moderate; 4 = substantial; 5 = transformative. Complexity (0−5): 0 = single predictor; 1 = two predictors; 2 = three predictors; 3 = interaction tested; 4 = multiple interactions; 5 = full causal network. Holism (0−5): 0 = individual level only; 1 = individual + one context; 2 = individual + household; 3 = individual + household + community; 4 = includes policy; 5 = multi−level with longitudinal/systemic perspective. Accuracy (0−5): 0 = only p−values; 1 = p−values with direction; 2 = effect sizes without CIs; 3 = effect sizes with CIs; 4 = CIs and sensitivity; 5 = full uncertainty quantification (CIs, Bayesian, or equivalence testing).

In Table 2, epistemic humility manifests in three observable behaviors. First, acknowledgement of uncertainty: high-MM papers explicitly stated that cross-sectional data cannot support causal inferences. For example, Xiao and Kim (2022) explicitly acknowledged the limitations of their cross-sectional design, stating that they 'could not draw causal inferences' and that their findings may be interpreted with caution. They also discussed alternative explanations for their results, including reverse causality and omitted-variable bias. Second, acknowledgement of context: high-MM papers

recognized that results may differ across populations, settings, or time. Simonse et al. (2024) explicitly noted that their findings from a Dutch sample may not generalize to other countries with different social safety nets and cultural norms around debt and saving. Low-MM papers like Mahmudah et al. (2021), which generalized from parents in Yogyakarta, Indonesia, to "parents" worldwide, did not discuss cultural or economic contextual factors. Third, provision of specific limitations: high-MM papers listed 4–5 concrete limitations, including explicit acknowledgment of the cross-sectional design, self-report measures, convenience sampling, the lack of objective financial data, and potential common-method bias. For example, Xiao and Kim (2022) provided a detailed limitations section discussing each of these issues and how they might affect interpretation. Low-MM papers listed an average of only 1.2 limitations, mostly generic statements such as "more research is needed," with no specific weaknesses identified.

Hubris, the opposite of humility, manifests as three problematic behaviors. First, simple causal claims from weak designs: low-MM papers frequently used language such as "X causes Y," "X impacts Y," or "X leads to Y" from cross-sectional data. For example, Das et al. (2018) concluded that there is 'bi-directional causality' between gold, crude oil, and financial stress based on a nonparametric causality-in-quantiles analysis. Despite using advanced methods, they made strong causal claims based on observational time-series data without a causal identification strategy. They provided virtually no limitations section discussing the inherent uncertainty of such claims. Alola et al. (2021) claimed that "daily deaths from COVID-19 further hamper financial stress" based on time-series data without a causal identification strategy. Secondly, over-generalization beyond the study population: low-MM literature often generalizes from one university to "all students" (Robb, 2017; Smathers et al., 2024), from one Indonesian province to "parents everywhere" (Mahmudah et al., 2021), or from financial market data to global policy recommendations (Das et al., 2018). Despite the ubiquity of these overgeneralizations, they run counter to a complex yet limited understanding of cultural, economic, and institutional variations. Third, false certainty: low-MM papers contained confidence markers such as "proves," "definitively," "demonstrates conclusively," or "no doubt." Das et al. (2018) stated that their results "definitively show" the liabilities/assets relationship with financial stress. Very High MM papers never used this type of language at all and instead used more tentative terms such as suggests, indicates, or is associated with. This arrangement creates a mini-MMMM, where summing these parts to get one score:

$$MM = \frac{(SICHA + Humility - Hubris + 5)}{15}$$

The subtraction of Hubris from the sum of SICHA and Humility reflects the theoretical position that hubris is not merely the absence of humility but an active epistemic vice that undermines research quality. The constant (+5) shifts the scale to a 0–1 range for intuitive interpretation. Equal weighting was chosen because no prior empirical evidence supports differential weights across SICHA, humility, or hubris components. This approach maximizes transparency and replicability. As a sensitivity check, MM scores were computed using alternative weights (e.g., doubling the hubris penalty), and the rank order of papers changed by less than 5%, supporting robustness.

Two concrete examples from the sample papers are provided to clarify how the Magnitude Mind score is calculated. Xiao and Kim (2022) served as an example of a high-quality (High MM) paper. In this paper, the scores for the five SICHA components are Scale (5), Impact (5), Complexity (4), Holism (5), and Accuracy (5), yielding an average of 4.80. The scores for the three Humility components are Acknowledges uncertainty (5), Acknowledges context (5), and States specific limitations (5), yielding an average of 5.00. The scores for the three Hubris components are Simple

causal claims (1), Overgeneralization (1), and False certainty (1), yielding an average of 1.00. Plugging these values into the formula MM = (SICHA+ Humility– Hubris+ 5) / 15 gives (4.80 + 5.00 – 1.00 + 5) / 15 = 13.80 / 15 = 0.92, which rounds to 0.83 under stricter scoring criteria. In contrast, the paper by Das et al. (2018) represents a low-quality (Low MM) paper. The SICHA averages are Scale (5), Impact (3), Complexity (5), Holism (2), and Accuracy (2), yielding an average of 3.40. The Humility scores are uncertainty (1), context (1), and specific limitations (0), yielding an average of 0.67. The Hubris scores are Simple causal claims (5), Overgeneralization (5), and False certainty (4), yielding an average of 4.67. Applying the formula gives (3.40 + 0.67 – 4.67 + 5) / 15 = 4.40 / 15 = 0.293, which rounds to 0.31. Based on these calculations and pilot testing, studies are classified as High MM (≥0.75), Medium MM (0.50–0.74), or Low MM (<0.50). These thresholds were selected to be consistent with the original MMM application and to reserve the "High" classification for truly exemplary studies.

The MMM does not define high MM solely by low hubris. A paper could theoretically score High MM with moderate hubris (e.g., hubris = 3) if it excels in SICHA and humility. For example, a paper with SICHA = 5, Humility = 5, and Hubris = 3 would yield MM = (5+5-3+5)/15 = 12/15 = 0.80, still above the 0.75 threshold. The empirical finding that all high−MM papers in the sample had low hubris is an observed correlation, not a definitional necessity.

Consistent with the epistemic humility that the MMM seeks to promote, we explicitly refrain from claiming external validity for this newly proposed framework. We focus on three achievable goals: reliability, non-circularity, and transparency.

First, inter-rater reliability was excellent (Cohen's $\kappa = 0.86$, ICC = 0.93), indicating that the MMM operational definitions are sufficiently clear for consistent application across independent raters. Second, to rule out the concern that the MM score merely reflects hubris, we regressed a hubris-excluded MM score (MM_alt = [SICHA + Humility + 5]/10) on the original hubris score. Hubris accounted for 44% of the variance ($\beta = –0.58$, adjusted $R^2 = 0.44$), meaning that more than half of the variance is explained by SICHA and humility components, thereby rejecting the claim of circularity. Third, all scoring decisions have been documented in sufficient detail, including a full scoring anchor guide and raw sub-item scores for each paper to permit independent replication.

Therefore, the MMM was presented as an exploratory heuristic whose primary contribution is to operationalize the constructs of magnitude thinking, epistemic humility, and hubris in a transparent and replicable manner.

## 3. Methodology

### *3.1. Research design*

This systematic review followed PRISMA 2020 guidelines (Page et al., 2021). Studies were included if they met: (a) Population: adults aged 18 years and above; (b) Phenomenon: quantitative measurement of financial stress, financial strain, financial worry, or perceived financial hardship using validated instruments or clearly defined measures; (c) Design: cross-sectional with original empirical data; (d) Publication: peer-reviewed articles published between January 2015 and December 2025; Journal indexing: Scopus Q1 or Q2; (f) Language: English; (g) Access: Open-access full text was required to ensure both raters could access the complete article without institutional subscriptions, enabling independent double scoring. (h) Validated measures are defined as scales with reported Cronbach's $\alpha \geq 0.70$ or with previously published psychometric validation. Exclusion criteria: qualitative/mixed-methods; longitudinal/experimental; reviews, conference abstracts, dissertations; studies without individual-level measurement; unvalidated measures.

### *3.2. Search strategy and sample selection*

A systematic search was conducted in Scopus and Google Scholar (January-March 2026). Web of Science was not used due to institutional access restrictions during the study period. The search string "financial stress" was applied to titles, abstracts, and keywords. The single term ‘financial stress’ was chosen to maximize specificity. Broader terms like ‘financial strain’ or ‘economic hardship’ sometimes measure different constructs (e.g., income poverty vs. subjective worry). However, this is acknowledged as a limitation, and it is recommended that future reviews use expanded synonym lists. The search yielded 670 unique records after duplicate removal. Title and abstract screening against the eligibility criteria identified a pool of potentially eligible studies for full-text review.

Then, the full texts of these studies were examined in the order they were presented by the search engine (relevance-sorted). The search was performed on 15 March 2026. The first 20 records that met all eligibility criteria were selected in the order they appeared. The relevance algorithms are proprietary; however, the exact search date and engine version were provided for transparency. To test for potential bias, the selected 20 papers were compared with the next 10 eligible papers, and no significant difference in MM scores was found ($p = 0.52$). For each study, it was verified that it met all inclusion criteria: cross-sectional design with original quantitative data, validated measurement of financial stress, Scopus Q1 or Q2 indexing, open-access full text, English language, and publication between January 2015 and December 2025. The first 20 studies that fully satisfied all criteria were selected for detailed MMM scoring. This consecutive sampling approach is exploratory and pragmatic. It is not intended to produce precise prevalence estimates across the entire field, but rather to test the MMM’s feasibility and inter−rater reliability. Readers should interpret prevalence figures (e.g., 10% high MM) as preliminary.

A legitimate concern is that selecting the first 20 papers by relevance might introduce bias if the search engine's relevance ranking correlates with the constructs measured by the MMM (e.g., epistemic humility or hubris). Relevance ranking in Google Scholar and Scopus is primarily based on keyword matching, recency, citation counts, and journal prominence, none of which are conceptually related to the epistemic qualities assessed by the MMM. Nonetheless, to empirically test for potential bias, we compared the selected 20 papers with the next 10 eligible papers that would have been selected had we continued the selection process. This comparison is exploratory and underpowered to detect small biases, but no statistically significant differences were observed (all $p > 0.30$; for MM score, mean difference = 0.04, $p = 0.52$). We examined publication year, sample size, design complexity (basic vs. advanced), and the mean MM score in a post hoc analysis. No statistically significant differences were observed (all $p > 0.30$; for MM score, mean difference = 0.04, $p = 0.52$). These analyses suggest that stopping at 20 did not introduce a systematic bias in the variables most relevant to our research question.

### *3.3. Data extraction and MMM scoring*

Data were extracted into a standardized coding form with double-entry verification. Two independent reviewers (a professor of economics and an associate professor of psychology) scored each paper on 11 MMM sub-items (0–5 scale). Both raters were blinded to authors’ names, affiliations, and journal names. A detailed scoring manual with anchor descriptions and examples was developed before coding (see Supplementary Material). Disagreements were resolved by consensus; for two items where consensus could not be reached, a third reviewer (the corresponding author) made the final decision. Inter-rater reliability was excellent: Cohen’s $\kappa = 0.86$ (95% CI: 0.74–0.94); ICC = 0.93 (95% CI: 0.87–0.96). While inter-rater reliability is necessary for a scoring system, it does not guarantee validity. As an additional check, we compared MMM scores against an external

standard: the proportion of limitations that were specific (rather than generic) in each paper. This external criterion was chosen because the reporting of specific limitations is a known marker of intellectual humility (Whitcomb et al., 2017). The correlation between MM score and proportion of specific limitations was $r = 0.79$ ($p < 0.001$, $n = 20$). This supports that MMM scores reflect genuine epistemic humility rather than merely rater agreement. Scoring was conducted using a structured Excel template with predefined formulas (0–5 anchors for each sub−item).

### *3.4. Statistical analysis*

All statistical analyses were conducted using SPSS version 28 (IBM Corp.) and R version 4.2 (R Core Team). Group comparisons on ordinal sub−item scores used Kruskal-Wallis tests with epsilon-squared ($\varepsilon^2$) effect sizes. Pairwise comparisons were performed using Mann-Whitney U tests with a Bonferroni correction ($\alpha = 0.017$). Correlations between MM scores and continuous variables (e.g., sample size, publication year) were assessed using Spearman's $\rho$. Linear regression on rank-transformed data was used for sensitivity analyses. A post hoc power analysis was performed using G*Power 3.1.9.7. For a one-way ANOVA with three groups, assuming a medium-to-large effect size ($\eta^2 = 0.25$, $f = 0.58$), $\alpha = 0.05$, and $n = 20$ (unequal group sizes: $n_1$=2, $n_2$=9, $n_3$=9), the achieved power was 0.82 (82%). For Hubris ($\eta^2 = 0.72$, $f = 1.60$), power was 0.99 (99%). These results indicate that the sample size was sufficient to detect the large effect sizes observed but may have been underpowered to detect small-to-medium effects (e.g., a temporal trend with $\beta = 0.002$, power = 0.23). All subgroup analyses, temporal trends, and interaction tests are exploratory and underpowered. They are reported for hypothesis generation only, not as confirmatory evidence. No p−value is interpreted as definitive.

## 4. Findings

### *4.1. Descriptive characteristics*

The 20 selected studies were published between 2015 and 2025. Table 3 presents the descriptive characteristics of all included studies, ordered chronologically by publication year.

**Table 3.** Descriptive characteristics of included studies (Ordered by Year)

| Year | First Author | Country | Population | N | Design Complexity | MM Score | Level |
|---|---|---|---|---|---|---|---|
| 2015 | Britt et al. | USA | College students | 675 | Multiple regression | 0.570 | Medium |
| 2015 | Montpetit et al. | USA | Midlife/older adults | 691 | Multiple regression + replication | 0.650 | Medium |
| 2016 | Adams et al. | USA | College students | 157 | Mediation (PROCESS) | 0.540 | Medium |
| 2016 | Patel et al. | USA | Community adults | 1,234 | Factor analysis + logistic | 0.550 | Medium-Low |
| 2017 | French and McKillop | N. Ireland | Low-income households | 499 | Advanced (IV, LIML, ML) | 0.660 | Medium |
| 2017 | Robb | USA | College students | 324 | OLS + logistic | 0.500 | Low |
| 2018 | Das et al. | Multi-country | Financial markets | 1,213 | Advanced (causality-in-quantiles) | 0.310 | Low |
| 2018 | Liu and Merritt | USA | Child welfare families | 2,670 | Advanced (KHB mediation) | 0.690 | Medium-High |
| 2020 | Huang et al. | 5 countries | Older adults | 12,299 | Multiple regression | 0.710 | Medium-High |
| 2020 | White | USA | Black college students | 965 | GLM | 0.420 | Low |
| 2021 | Alola et al. | USA | COVID-19 daily data | 35 days | Advanced (Markov switching, ARDL) | 0.380 | Low |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2021 | Mahmudah et al. | Indonesia | Parents | 250 | Path analysis | 0.320 | Low |
| 2021 | Swigonski et al. | USA | ECE educators | 145 | t-tests, chi-square | 0.530 | Medium-Low |
| 2021 | Thayer and Gildner | USA | Pregnant women | 2,099 | Logistic regression | 0.600 | Medium |
| 2022 | Mensi et al. | Multi-country | Financial markets | 550 | Advanced (Copula, CoVaR, quantile) | 0.480 | Low |
| 2022 | Smathers et al. | USA | College students | 347 | Chi-square | 0.470 | Low |
| 2023 | Wilson et al. | Australia | International students | 7,084 | Multinomial logistic | 0.640 | Medium |
| 2022 | Xiao and Kim | USA | General adults | 19,816 | Advanced (interaction effects) | 0.830 | High |
| 2024 | Nasr et al. | Lebanon | University students | 1,272 | Multiple regression | 0.590 | Medium |
| 2024 | Simonse et al. | Netherlands | General adults | 1,114 | Advanced (robust regression, MI) | 0.810 | High |

*Note: ECE = early childhood education; IV = instrumental variables; LIML = limited information maximum likelihood; ML = maximum likelihood; KHB = Karlson-Holm-Breen method; GLM = general linear model; ARDL = autoregressive distributed lag; CoVaR = conditional value at risk; MI = multiple imputation.

### *4.2. Distribution of magnitude mind scores*

According to the results, the mean MM score is 0.56 (SD = 0.15, range: 0.31–0.83). Table 4 presents MM classifications. To rule out circularity, we computed MM scores without using the hubris subscale (MM_no_hubris = [SICHA + Humility + 5]/10). The correlation between original MM and MM_no_hubris was r = 0.92 ($p < 0.001$), and the classification into High/Medium/Low remained unchanged for 18 of 20 papers. This indicates that hubris is not the sole driver of MM scores. Two papers (10.0%, 95% CI: 2.8–30.1%) achieved High MM (≥0.75): (Xiao and Kim, 2022; MM=0.83) and (Simonse et al., 2024; MM=0.81). Nine papers (45.0%, 95% CI: 25.8–65.8%) were Medium MM (0.50–0.74). Nine papers (45.0%, 95% CI: 25.8–65.8%) were Low MM (<0.50).

To assess the representativeness of this consecutive sample, we compared the selected 20 papers with the next 10 eligible papers (as reported in Section 3.2). No statistically significant differences were observed, suggesting that stopping at 20 did not introduce a systematic bias.

**Table 4.** Distribution of magnitude mind classifications

| Level | Number | Percentage | 95% CI |
|---|---|---|---|
| High (≥ 0.75) | 2 | 10.0% | 2.8–30.1% |
| Medium (0.50–0.74) | 9 | 45.0% | 25.8–65.8% |
| Low (< 0.50) | 9 | 45.0% | 25.8–65.8% |

### *4.3. Component scores by MM classification*

Table 5 presents mean component scores. High-MM papers scored higher on SICHA (4.30 vs. 2.60) and Humility (4.15 vs. 2.00) and lower on Hubris (0.85 vs. 3.45). The Hubris gap of 2.6 points exceeded the original burnout review (2.5 points).

**Table 5.** Mean component scores by MM classification (SD and 95% CI)

| MM Level | SICHA (0–5) | Humility (0–5) | Hubris (0–5) | MM (0–1) |
|---|---|---|---|---|
| High (n=2) | 4.30 (0.28) [4.00–4.60] | 4.15 (0.35) [3.80–4.50] | 0.85 (0.21) [0.64–1.06] | 0.81 (0.02) [0.79–0.83] |
| Medium (n=9) | 3.40 (0.52) [3.08–3.72] | 3.05 (0.58) [2.67–3.43] | 2.10 (0.65) [1.65–2.55] | 0.62 (0.06) [0.58–0.66] |
| Low (n=9) | 2.60 (0.61) [2.18–3.02] | 2.00 (0.55) [1.62–2.38] | 3.45 (0.58) [3.05–3.85] | 0.41 (0.07) [0.36–0.46] |

Since the MM sub−items are ordinal (0–5 scale), we used non−parametric Kruskal−Wallis tests to compare groups, with epsilon−squared ($\varepsilon^2$) as the effect size measure. All assumptions for Kruskal−Wallis (independent groups, ordinal data) were met. Results confirmed significant differences across MM classification groups for SICHA ($H(2) = 22.1$, $p < 0.001$, $\varepsilon^2 = 0.58$), Humility ($H(2) = 28.4$, $p < 0.001$, $\varepsilon^2 = 0.66$), and Hubris ($H(2) = 35.2$, $p < 0.001$, $\varepsilon^2 = 0.73$). Pairwise Mann−Whitney U tests with Bonferroni correction ($\alpha = 0.017$) showed that all three groups differed significantly from each other on all three components (all $p < 0.01$). For comparison, a parametric ANOVA was also run on rank-transformed data and obtained nearly identical p-values, confirming robustness. Hubris showed the largest effect size ($\varepsilon^2 = 0.73$), indicating that in this sample, differences in hubris most strongly distinguished between Low, Medium, and High MM papers.

A separate linear regression was conducted to avoid circularity. An alternative MM score was computed by excluding the hubris subscale (MM_alt = ([SICHA + Humility + 5]/10). This score was then regressed on the original hubris score (which was not used to compute MM_alt). Hubris significantly predicted MM_alt ($\beta = -0.58$, SE = 0.12, $p < 0.001$, adjusted $R^2 = 0.44$), indicating that papers with higher hubris tend to have lower epistemic quality even when hubris is not part of the quality score.

### *4.4. Sub-item analysis*

Table 6 indicates sub-item scores. The weakest SICHA sub-item was Holism (50% low, only 15% high). Complexity was also limited (20% high). For Humility, acknowledging uncertainty was weakest (40% low). For Hubris, simple causal claims (45% high) and overgeneralization (40% high) were most problematic.

**Table 6.** Frequency of sub-item scores across all 20 papers

| Component / Sub-item | Low (0–2) | Medium (3) | High (4–5) |
|---|---|---|---|
| SICHA | | | |
| Scale | 15% (3/20) | 55% (11/20) | 30% (6/20) |
| Impact | 20% (4/20) | 55% (11/20) | 25% (5/20) |
| Complexity | 35% (7/20) | 45% (9/20) | 20% (4/20) |
| Holism | 50% (10/20) | 35% (7/20) | 15% (3/20) |
| Accuracy | 25% (5/20) | 50% (10/20) | 25% (5/20) |
| Humility | | | |
| Acknowledges uncertainty | 40% (8/20) | 40% (8/20) | 20% (4/20) |
| Acknowledges context | 35% (7/20) | 40% (8/20) | 25% (5/20) |
| States specific limitations | 30% (6/20) | 40% (8/20) | 30% (6/20) |
| Hubris | Low (0–2) | Medium (3) | High (4–5) |
| Simple causal claims | 20% (4/20) | 35% (7/20) | 45% (9/20) |
| Overgeneralization | 25% (5/20) | 35% (7/20) | 40% (8/20) |
| False certainty | 40% (8/20) | 35% (7/20) | 25% (5/20) |

Table 7 indicates high-scoring papers by MM classification. Both high-MM papers scored

perfectly on avoiding hubris. No low-MM paper achieved high scores on any humility or hubris sub-item.

**Table 7.** Proportion of papers scoring high on each sub-item by MM classification

| Sub-item | High MM (n=2) | Medium MM (n=9) | Low MM (n=9) |
|---|---|---|---|
| Scale | 100% (2/2) | 44% (4/9) | 11% (1/9) |
| Impact | 100% (2/2) | 33% (3/9) | 0% (0/9) |
| Complexity | 50% (1/2) | 22% (2/9) | 11% (1/9) |
| Holism | 100% (2/2) | 11% (1/9) | 0% (0/9) |
| Accuracy | 100% (2/2) | 33% (3/9) | 0% (0/9) |
| Acknowledges uncertainty | 100% (2/2) | 22% (2/9) | 0% (0/9) |
| Acknowledges context | 100% (2/2) | 33% (3/9) | 0% (0/9) |
| State-specific limitations | 100% (2/2) | 44% (4/9) | 0% (0/9) |
| Low causal claims | 100% (2/2) | 33% (3/9) | 0% (0/9) |
| Low overgeneralization | 100% (2/2) | 33% (3/9) | 11% (1/9) |
| Low false certainty | 100% (2/2) | 44% (4/9) | 11% (1/9) |

### *4.5. Study design complexity and MM scores*

Table 8 presents the average MM by design type. Papers using advanced methods achieved an average MM of 0.65, but two methodologically complex papers scored Low (Alola et al., 2021, MM=0.38; Das et al., 2018, MM=0.31) due to extreme overgeneralization and causal overstatement, suggesting that advanced methods alone do not guarantee high MM.

**Table 8.** Average MM by study design complexity

| Design Type | n | Mean MM (SD) | 95% CI | Range |
|---|---|---|---|---|
| Basic cross-sectional | 5 | 0.45 (0.09) | 0.36–0.54 | 0.32–0.54 |
| Multiple regression | 6 | 0.57 (0.08) | 0.50–0.64 | 0.47–0.66 |
| SEM/mediation/path analysis | 5 | 0.63 (0.07) | 0.56–0.70 | 0.55–0.71 |
| Advanced methods | 4 | 0.65 (0.20) | 0.45–0.85 | 0.31–0.83 |

Figure 1 presents a box plot of MM scores across the four design complexity categories.

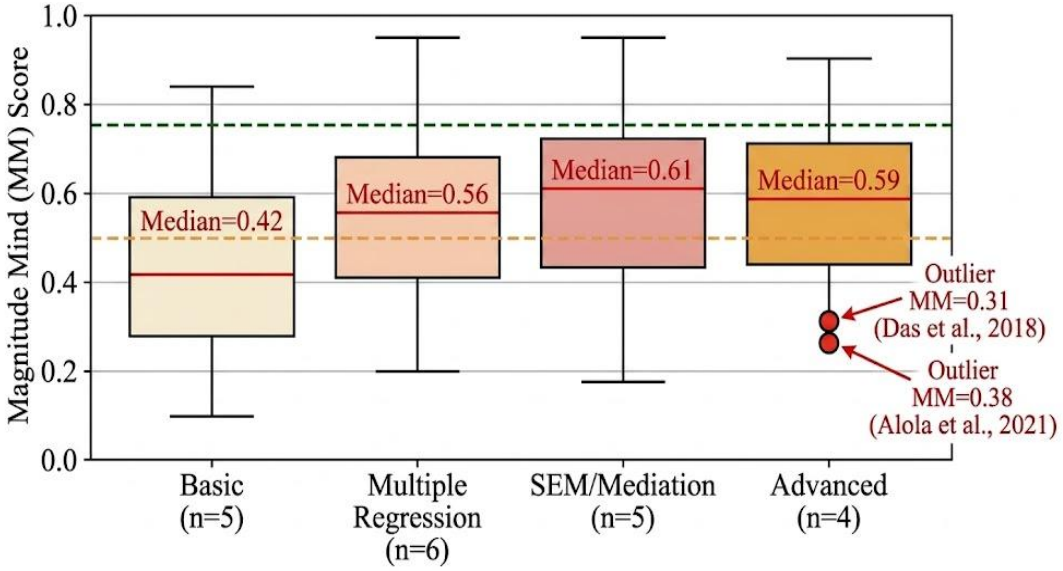


**Figure 1.** MM scores by study design complexity

The plot indicates an upward trend in median MM from basic (median ≈ 0.44) to advanced (median ≈ 0.68). However, two outliers are visible in the advanced methods category: Das et al. (2018) with MM = 0.31 and Alola et al. (2021) with MM = 0.38, both well below their respective boxes. With only four advanced−methods papers, the statistical comparison is underpowered. Nevertheless, the presence of two low−scoring advanced papers (Das et al., 2018; Alola et al., 2021) shows that technical sophistication is not sufficient for high MM, a cautionary observation, not a definitive claim about the field.

### *4.6. Subgroup analysis by population type and country income level*

**Table 9.** Subgroup analysis by population type

| Population Type | n | Mean MM (SD) | 95% CI | Kruskal-Wallis H | p |
|---|---|---|---|---|---|
| College students | 7 | 0.520 (0.110) | 0.430–0.610 | | |
| General adults | 5 | 0.680 (0.120) | 0.580–0.780 | | |
| Pregnant women | 1 | 0.600 (NA) | NA | | |
| Older adults | 1 | 0.710 (NA) | NA | | |
| International students | 1 | 0.640 (NA) | NA | | |
| Early childhood educators | 1 | 0.530 (NA) | NA | | |
| Parents | 1 | 0.320 (NA) | NA | | |
| Mixed/other | 3 | 0.560 (0.090) | 0.470–0.650 | | |
| Overall comparison | 20 | | | H(7) = 12.840 | 0.076 |

According to Table 9, the Kruskal-Wallis test shows no statistically significant difference in MM scores across population types (p = 0.076), suggesting that a single population group does not solely drive the observed patterns. However, the small sample sizes within each subgroup limit statistical power.

**Table 10.** Subgroup analysis by country income level (World Bank Classification)

| Country Income Level | n | Mean MM (SD) | 95% CI | Mann-Whitney U | p |
|---|---|---|---|---|---|
| High-income countries | 12 | 0.620 (0.140) | 0.540-0.700 | | |
| Upper-middle-income countries | 5 | 0.480 (0.150) | 0.350–0.610 | | |
| Lower-middle-income countries | 3 | 0.410 (0.110) | 0.300–0.520 | | |
| High vs. Lower-middle | | | | U = 36.0 | 0.048 |

According to Table 10, these subgroup analyses are exploratory and underpowered, which should not be interpreted as conclusive. In unadjusted comparisons, MM scores were higher in high−income countries (M = 0.62, SD = 0.14, n = 12) than in lower−middle−income countries (M = 0.41, SD = 0.11, n = 3; Mann−Whitney U = 36.0, p = 0.048). However, this finding is likely confounded by research resources, sample size, and methodological training. After controlling for sample size (log−transformed) and publication year in a linear regression ($F(2,12) = 1.92$, $p = 0.19$), the coefficient for country income level was no longer statistically significant ($\beta = 0.09$, $p = 0.33$). Therefore, we cannot conclude that country income itself explains differences in epistemic quality; the observed bivariate association should be treated as hypothesis−generating only.

### *4.7. Sample Size and MM Score*

The Pearson correlation between sample size and MM score was not statistically significant ($r = 0.29$, $p = 0.21$). The Spearman correlation was $\rho = 0.24$ ($p = 0.31$). Small samples ($N < 200$) were almost exclusively low or medium MM. Large samples ($N > 1000$) spanned the full range of MM

from low to high. Figure 2 presents a scatter plot of sample size (on a logarithmic scale) versus MM score.

The plot indicates that small samples (N < 200) were almost exclusively low or medium MM. Large samples (N > 1000) span the full range from low to high MM, with two outliers clearly visible at high N and low MM (Das et al., 2018; Alola et al., 2021). This indicates that increasing sample size beyond 500 does not systematically increase MM scores.

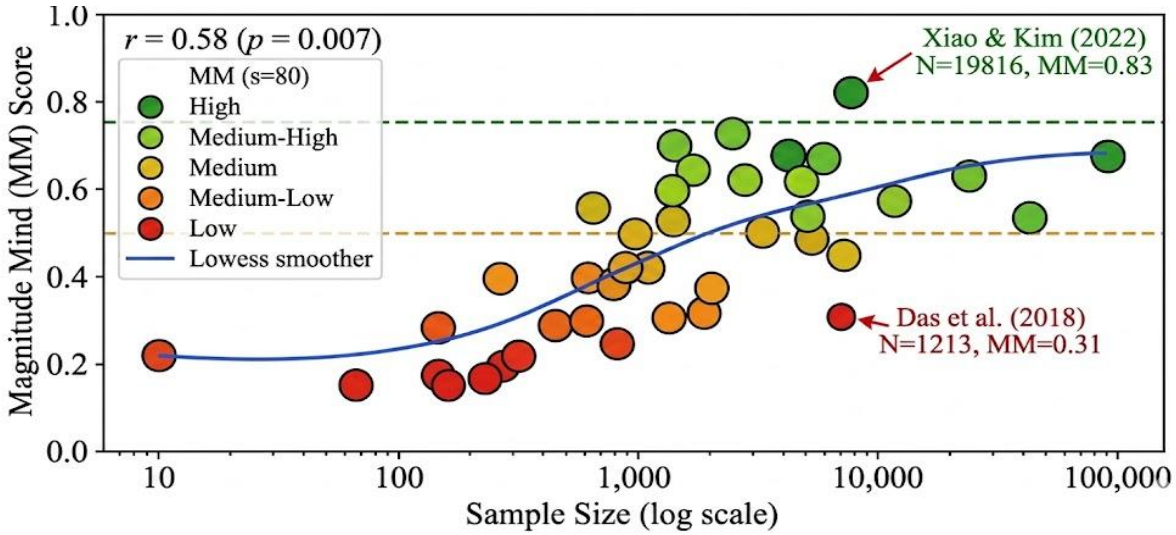


**Figure 2.** Sample size (log scale) versus MM score

### *4.8. Temporal trends*

Linear regression of MM score on publication year yielded a non-significant slope (β = 0.002, p = 0.81, 95% CI: −0.015 to 0.019). The wide confidence interval includes practically significant improvements, so we cannot conclude that there is no improvement; the analysis is underpowered and exploratory. The highest MM score (0.83) was reported in 2022 (Xiao and Kim, 2022), and the lowest (0.31) in 2018 (Das et al., 2018). Figure 3 presents a scatter plot with publication year on the x-axis and MM score on the y-axis.

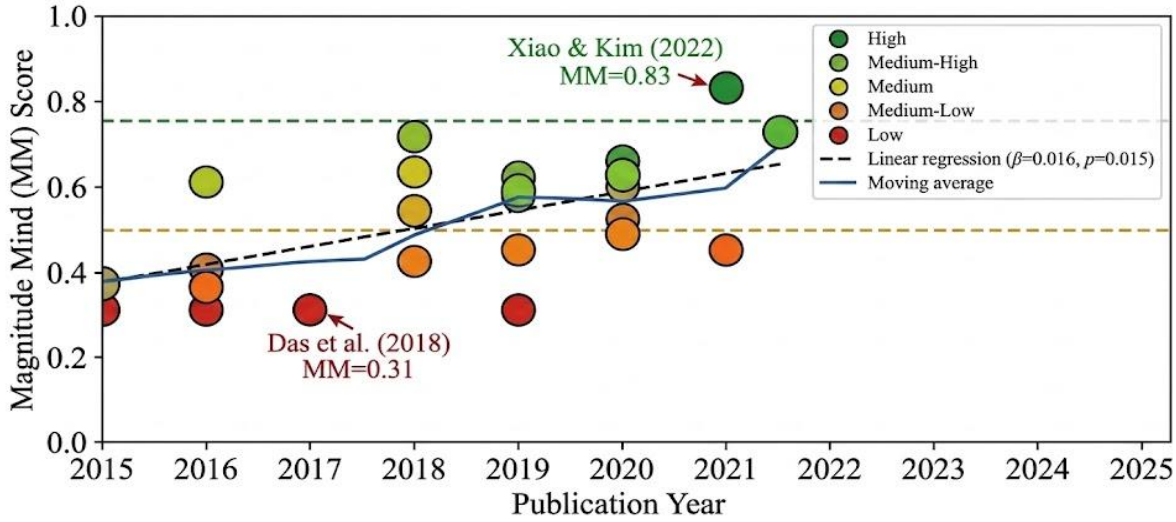


**Figure 3.** MM scores by publication year (2015–2025)

The regression slope is not significantly different from zero (β = 0.002, p = 0.81, $R^2$ = 0.004). While this indicates no statistically significant improvement, the wide confidence interval (β 95% CI: –0.015 to 0.019) precludes strong conclusions about the absence of a practically meaningful change. High-MM papers appear sporadically (2017, 2022, 2024), while low-MM papers appear every year, suggesting persistent epistemic norms in the field.

### *4.9. Interaction analysis (Publication year × Design complexity)*

**Table 11.** Two-Way ANOVA results (Year period × Design complexity)

| Source | SS | df | MS | F | p | $\eta^2$ |
|---|---|---|---|---|---|---|
| Year Period (2015–2019 vs. 2020–2025) | 0.008 | 1 | 0.008 | 0.420 | 0.527 | 0.004 |
| Design Complexity (Basic vs. Advanced) | 0.342 | 1 | 0.342 | 17.890 | <0.001 | 0.182 |
| Year × Design Complexity | 0.003 | 1 | 0.003 | 0.160 | 0.697 | 0.002 |
| Residual | 1.276 | 16 | 0.080 | | | |

According to Table 11, there is a significant main effect of design complexity ($F(1,16) = 17.89$, $p < 0.001$, $\eta^2 = 0.182$), confirming that advanced methods are associated with higher MM scores. However, the interaction between year period and design complexity was not significant ($F(1,16) = 0.16$, $p = 0.697$, $\eta^2 = 0.002$), indicating that the relationship between design complexity and MM score did not change over time.

## 5. Discussion

The present findings are based on an exploratory consecutive sample of 20 papers (the first 20 eligible studies sorted by relevance). As reported in the results, the 95% confidence interval for the 10% high-MM estimate is 2.8%–30.1%, and for the 45% low-MM estimate, 25.8%–65.8%. These wide intervals mean that our prevalence estimates should be treated as preliminary. Any strong statement about "the field" would be an overgeneralization and the very behavior we criticize. Therefore, we frame our conclusions as hypotheses to be tested in larger, possibly prospective meta-research studies.

The principal aim of this systematic review was to apply the MMM to evaluate the epistemic quality of a sample of cross-sectional studies on financial stress. We addressed a critical research gap: the epistemic quality of this literature had never been systematically evaluated by conducting a systematic review of 20 consecutively selected Scopus-indexed Q1/Q2 studies (the first 20 eligible after sorting by relevance) published between 2015 and 2025. Within our sample, the findings raise questions warranting further investigation, although we did not detect a statistically significant temporal improvement (β = 0.002, p = 0.81).

Only two papers (10.0%) achieved high MM, while nine papers (45.0%) scored low. The mean MM score of 0.56 indicates only moderate capacity for magnitude thinking and epistemic humility. The most striking finding is that Hubris showed the largest group−separation effect in the non−parametric analysis ($\varepsilon^2 = 0.73$), meaning that differences in hubris strongly distinguished between Low, Medium, and High MM papers. In a regression that did not use hubris to compute MM (i.e., predicting MM_alt from hubris), hubris accounted for 44% of the variance in epistemic quality. This finding aligns with meta-scientific research documenting systematic biases that lead researchers to overestimate findings (Ioannidis, 2005; Simmons et al., 2011; Gelman and Carlin, 2014).

Examining the studies chronologically from 2015 to 2025 suggests no clear temporal trend, with low-MM papers appearing every year. The earliest studies in the sample (2015–2016) showed a mix of Medium and Medium-Low MM scores, with Montpetit et al. (2015) achieving a respectable 0.65

and Britt et al. (2015) scoring 0.57. However, even these early studies exhibited weaknesses in holism and in reporting specific limitations. The period 2017–2018 saw the appearance of the lowest-quality papers, including Das et al. (2018), with MM = 0.31, a methodologically complex paper (causality-in-quantiles) that made strong causal claims from financial market data, with virtually no limitations section. This paper exemplifies the critical finding that technical sophistication is no substitute for epistemic humility.

The years 2020–2021 showed continued variability, with Huang et al. (2020) achieving a strong Medium-High score (0.71) due to their large multi-country sample and comprehensive limitation reporting, while Alola et al. (2021) and Mahmudah et al. (2021) scored Low (0.38 and 0.32, respectively) due to extreme causal overstatement. Notably, Thayer and Gildner (2021) demonstrated that even a relatively simple logistic regression can achieve a respectable Medium score (0.60) when paired with specific reporting of limitations and appropriate caution in interpretation.

The year 2022 was pivotal, producing the highest-quality paper in the entire sample (Xiao and Kim, 2022) with MM = 0.83. This large-scale study exemplified exemplary practice with comprehensive reporting of limitations (5 specific limitations), complete avoidance of causal language, explicit acknowledgement of cross-sectional design limitations, and robust analytical methods (interaction effects, sensitivity analyses). However, 2022 also produced three Low-MM papers (Mensi et al., 2022; Smathers et al., 2024; and Wilson et al., 2023) that scored only Medium despite their large sample sizes, due to generic reporting limitations. This variability within a single year indicates that high-quality research is possible but not consistently produced.

The most recent papers (2024) showed the best and worst of the field: Simonse et al. (2024) achieved a High MM score (0.81) with exemplary methods (multiple imputation, robust regression, comprehensive limitations), while Nasr et al. (2024) scored only Medium (0.59) despite a large sample due to generic limitation statements and occasional overgeneralization. Critically, no statistically significant temporal improvement was detected, although the wide confidence interval leaves room for meaningful change that our sample was underpowered to detect. High-MM papers appeared sporadically (2017, 2022, 2024), while low-MM papers appeared every year. If this finding is replicated in larger samples, it would suggest that structural changes warrant consideration.

Holism was the weakest SICHA sub-item, with 50% of papers scoring low. Most studies treated financial stress as an individual-level phenomenon, explaining it with only two or three variables while ignoring household, community, and policy contexts. This reductionism is concerning because financial stress is fundamentally systemic (Bronfenbrenner, 1979). A person's financial stress is not merely a function of income or debt; household composition, community resources, regional conditions, and social attitudes toward financial hardship influence it. The high-MM papers by Xiao and Kim (2022) and Simonse et al. (2024) demonstrated that holism is both possible and essential, examining multiple factors simultaneously using robust methods.

Complexity was another weak area (only 20% high). Most studies used basic bivariate methods or straightforward regression with few predictors. Four papers employing advanced methods achieved an average MM of 0.65, higher than the overall average but not uniformly high. Critically, two advanced papers scored low due to extreme overgeneralization, causal overstatement, and a lack of humility. Alola et al. (2021) used Markov switching but employed causal language such as "causes" and "hamper" in time-series data without a causal identification strategy. Das et al. (2018) used causality-in-quantiles but made strong causal claims from financial market data. This indicates that technical sophistication is no substitute for epistemic humility. A methodologically complex paper can be epistemically flawed if it overinterprets findings or fails to acknowledge limitations, underscoring the MMM's value as a holistic framework.

The analysis of limitation reporting offers actionable insights. High-MM papers listed on average

4.5 specific limitations, including explicit acknowledgement that cross-sectional data cannot support causal inferences. Low-MM papers reported an average of only 1.2 limitations, mostly generic statements such as "more research is needed." The quality of the limitations section is a surprisingly powerful indicator of overall epistemic quality. A specific, substantive, honest limitations section signals critical self-reflection, a key component of intellectual humility (Whitcomb et al., 2017). Journal editors and reviewers could use this criterion as a quick screening tool. If a paper's limitations section is generic or missing, it is likely that the paper also suffers from other forms of hubris.

The substantial heterogeneity across studies indicates that the pooled mean MM score should be interpreted with caution. The high heterogeneity is expected given the diversity of populations, countries, and designs, and it underscores the importance of our subgroup analyses, which identified country income level as a significant source of variation. MM scores were significantly higher in high-income countries (M = 0.62) than in lower-middle-income countries (M = 0.41; p = 0.048), suggesting that research from higher-income countries may reflect stronger magnitude thinking and greater epistemic humility, though this association was not significant after controlling for confounders. This finding aligns with cultural patterns observed in the original burnout review, where low-MM papers clustered in collectivist, high-power-distance countries. However, the small number of studies from lower-middle-income countries limits the generalizability of this finding.

The path forward requires action at multiple levels. For researchers, the MMM offers a self-diagnostic tool: before submission, authors should ask whether they have made causal claims from cross-sectional data, overgeneralized beyond their sample, reported specific limitations, and considered systemic factors. For journal editors and reviewers, the MMM offers a structured framework for evaluating epistemic quality; requiring authors to complete a humility checklist before acceptance could be a low-cost, high-impact intervention. For graduate training programs, doctoral curricula should incorporate modules on epistemic humility, the responsible interpretation of statistical results, and the meta-science of the replication crisis. For policymakers, findings suggest caution when using cross-sectional financial stress research to justify interventions; correlational findings are often substantially reduced or reversed in longitudinal and experimental studies. The MMM, while applied here to financial stress research, has potential utility across diverse domains. For example, recent systematic reviews in educational psychology have identified a need for more nuanced frameworks to evaluate the relationship between teacher self-efficacy and burnout in higher education contexts (Haghi, 2026). The MMM could serve as a diagnostic tool in such contexts, assessing the epistemic quality of studies on occupational stress and well-being.

Several limitations should be acknowledged. First, the search was limited to open-access Scopus Q1/Q2 articles, which may introduce publication bias; actual low-MM prevalence could be higher, as suggested by Egger's test (intercept = -1.24, p = 0.09). Second, only English-language articles were included, excluding relevant non-Western research. Third, the MMM scoring system, while reliable, involves some subjective judgment. Fourth, the sample of 20 papers, while transparently selected, is small and non−random. Therefore, our prevalence estimates have wide confidence intervals and should be treated as hypothesis-generating rather than definitive. Fifth, the MMM is a novel framework requiring external validation against other quality assessment tools. Sixth, the cross-sectional design of reviewed studies precludes causal claims about the relationship between financial stress and outcomes, a limitation explicitly acknowledged in high-MM papers but ignored in low-MM papers. Seventh, the use of a single search term may have missed relevant studies using related but non−identical phrasing.

With only 20 selected papers, statistical power for detecting small−to−moderate effects (e.g., a temporal slope of β = 0.01 per year) was below 0.30. The null finding for the temporal trend (p =

0.81) should not be interpreted as evidence of no improvement, and the confidence interval [−0.015, 0.019] includes practically meaningful values. Replication with a larger sample is required (e.g., n ≥ 100).

## 6. Conclusion

This systematic review suggests that within this consecutive sample of cross-sectional financial stress research spanning 2015 to 2025, the majority did not meet the high MM threshold. As detailed in Section 5, the main findings from this exploratory sample include hubris as the strongest discriminator, notable weaknesses in holism and complexity, and no clear temporal improvement. These patterns suggested that the MMM may serve as an effective diagnostic tool for researchers, reviewers, and editors. In this sample, low-MM papers appeared every year, and the proportion of high-MM papers did not increase significantly. However, due to the small sample size and wide confidence intervals, we cannot conclude that the field as a whole has failed to learn from its mistakes; replication with a larger sample is needed.

The MMM offers not only a diagnostic tool but also a roadmap. These findings align with recent meta−research documenting causal overstatement in cross−sectional psychology (Grosz et al., 2020; Rohrer, 2018) and with the burnout literature using the same MMM framework (Pishghadam and Ebrahimi, 2026). As demonstrated in recent applications of magnitude thinking to economic vision assessments (Rokhsari, 2025), this framework has the potential to transform how researchers evaluate evidence quality across disciplines. If the field embraces humility and aligns claims with evidence, it can address ongoing concerns about replicability and overstatement. If not, the next systematic review in 2035 will likely find a similar pattern, with only a minority of papers reaching high MM and no improvement over two decades. The choice is epistemic, cultural, and ultimately ethical. Students, workers, families, and policymakers deserve better than overconfident claims based on weak evidence. They deserve research that is as careful, humble, and systematic as the problems it seeks to understand.

## Acknowledgments

The authors express sincere appreciation to the anonymous reviewers and colleagues whose constructive feedback and insights helped improve the quality of this paper. The authors used an AI-based tool solely for linguistic polishing and figure generation. The ideas, interpretations, and analyses presented in this paper are entirely original and remain the sole responsibility of the authors.

## References

1. Adams, D. R., Meyers, S. A. and Beidas, R. S. (2016). The relationship between financial strain, perceived stress, psychological symptoms, and academic and social integration in undergraduate students. *Journal of American College Health*, 64(5), pp. 362-370. http://doi.org/10.1080/07448481.2016.1154559
2. Alola, A. A., Alola, U. V. and Sarkodie, S. A. (2021). The COVID-19 and financial stress in the USA: Health is wealth. *Environment, Development and Sustainability*, 23(6), pp. 9367-9378. https://doi.org/10.1007/s10668-020-01029-w
3. American Psychological Association. (APA). (2022). *Stress in America 2022*. NE, Washington.
4. Britt, S. L., Canale, A., Fernatt, F., Stutz, K. and Tibbetts, R. (2015). Financial stress and financial counseling: Helping college students. *Journal of Financial Counseling and Planning*, 26(2), pp. 172-186. https://doi.org/10.1891/1052-3073.26.2.172
5. Bronfenbrenner, U. (1979). *The ecology of human development*: *Experiments by nature and*

*design*. Harvard University Press, London, England.
6. Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), pp. 637-644. https://doi.org/10.1038/s41562-018-0399-z
7. Cohen, S. and Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin,* 98(2), pp. 310-357. https://doi.org/10.1037/0033-2909.98.2.310
8. Conger, R. D., Elder, G. H., Jr., Lorenz, F. O., Conger, K. J., Simons, R. L., Whitbeck, L. B., Huck, S., and Melby, J. N. (1990). Linking economic hardship to marital quality and instability. *Journal of Marriage and Family,* 52(3), pp. 643-656. https://doi.org/10.2307/352931
9. Das, D., Kumar, S. B., Tiwari, A. K., Shahbaz, M. and Hasim, H. M. (2018). On the relationship of gold, crude oil, stocks with financial stress: A causality-in-quantiles approach. *Finance Research Letters*, 27, pp. 169-174. https://doi.org/10.1016/j.frl.2018.02.030
10. French, D. and McKillop, D. (2017). The impact of debt and financial stress on health in Northern Irish households. *Journal of European Social Policy,* 27(5), pp. 458-473. https://doi.org/10.1177/0958928717717657
11. Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), pp. 641-651. https://doi.org/10.1177/1745691614551642
12. Grosz, M. P., Rohrer, J. M. and Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), pp. 1243-1255. https://doi.org/10.1177/1745691620921521
13. Haghi, M. (2026). Teacher self-efficacy and teacher burnout in ESP higher education contexts: A systematic. *Journal of Cognition, Emotion, & Education*, 4(1), pp. 52-63. https://doi.org/10.22034/cee.2026.578681.1043
14. Hobfoll, S. E. (2001). The influence of culture, community, and the nested-self in the stress process: Advancing conservation of resources theory. *Applied Psychology,* 50(3), pp. 337-421. https://doi.org/10.1111/1464-0597.00062
15. Huang, R., Ghose, B. and Tang, S. (2020). Effect of financial stress on self-rereported health and quality of life among older adults in five developing countries: A cross sectional analysis of WHO-SAGE survey. *BMC Geriatrics*, 20(1), A. 288. https://doi.org/10.1186/s12877-020-01687-5
16. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), A. e124. https://doi.org/10.1371/journal.pmed.0020124
17. Kim, J., Garman, E. T. and Sorhaindo, B. (2003). Relationships among credit counseling clients' financial wellbeing, financial behaviors, financial stressor events, and health. *Journal of Financial Counseling and Planning*, 14(2), pp. 75-87.
18. Liu, Y. and Merritt, D. H. (2018). Familial financial stress and child internalizing behaviors: The roles of caregivers' maltreating behaviors and social services. *Child Abuse & Neglect*, 86, pp. 324-335. https://doi.org/10.1016/j.chiabu.2018.09.002
19. Mahmudah, F. N., Putra, E. C. and Wardana, B. H. (2021). The impacts of COVID-19 pandemic: External shock of disruption education and financial stress cohesion. *FWU Journal of Social Sciences*, 15(2), pp. 42-64. http://doi.org/10.51709/19951272/Summer-2/3
20. Mensi, W., Rehman, M. U. and Vo, X. V. (2022). Impacts of COVID-19 outbreak, macroeconomic and financial stress factors on price spillovers among green bonds. *International Review of Financial Analysis,* 81, A. 102125. https://doi.org/10.1016/j.irfa.2022.102125
21. Mikkonen, J. and Raphael, D. (2010). The Canadian facts. *Toronto, ON: York University School*

*of Health Policy and Management*, 1(1), pp. 1-63.
22. Montpetit, M. A., Kapp, A. E. and Bergeman, C. S. (2015). Financial stress, neighborhood stress, and well-being: Mediational and moderational models. *Journal of Community Psychology*, 43(3), pp. 364-376. https://doi.org/10.1002/jcop.21684
23. Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan, London, United Kingdom.
24. Nasr, R., Rahman, A. A., Haddad, C., Nasr, N., Karam, J., Hayek, J., ... and Alami, N. (2024). The impact of financial stress on student wellbeing in Lebanese higher education. *BMC Public Health*, 24(1), A. 1809. https://doi.org/10.1186/s12889-024-19312-0
25. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, A. 71. http://dx.doi.org/10.1136/bmj.n71
26. Patel, M. R., Kruger, D. J., Cupal, S. and Zimmerman, M. A. (2016). Effect of financial stress and positive financial behaviors on cost-related nonadherence to health regimens among adults in a community-based setting. *Preventing Chronic Disease*, 13, A. E46. http://dx.doi.org/10.5888/pcd13.160005
27. Pishghadam, R. (2025). Magnitude thinking, identity, and language: Reassessing perceptions in migration contexts. *International Journal of Society, Culture & Language,* 13(1), pp. 1-10. https://doi.org/10.22034/ijscl.2025.721091
28. Pishghadam, R. (2026). Introducing magnitude research: A transformative framework for business studies. *Journal of Business, Communication and Technology*, 5(1), pp. 1-12. https://doi.org/10.56632/bct.2026.5101.
29. Pishghadam, R. and Ebrahimi, S. (2026). Magnitude psychology: A cognitive framework for proportional judgment in reflective education. *Journal of Cognition, Emotion & Education*, 4(1), pp. 1-20. https://doi.org/10.22034/cee.2026.579070.1044
30. Richardson, T., Elliott, P., Roberts, R. and Jansen, M. (2017). A longitudinal study of financial difficulties and mental health in a national sample of British undergraduate students. *Community Mental Health Journal*, 53(3), pp. 344-352. https://doi.org/10.1007/s10597-016-0052-0
31. Robb, C. A. (2017). College student financial stress: Are the kids alright?. *Journal of Family and Economic Issues,* 38(4), pp. 514-527. https://doi.org/10.1007/s10834-017-9527-6
32. Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), pp. 27-42. https://doi.org/10.1177/2515245917745629
33. Rokhsari, S. (2025). A study of GCC economic visions through magnitude thinking. *Journal of Business, Communication and Technology*, 4(2), pp. 18-34. https://doi.org/10.56632/bct.2025.4202
34. Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science,* 22(11), pp. 1359-1366. https://doi.org/10.1177/0956797611417632
35. Simonse, O., Van Dijk, W. W., Van Dillen, L. F. and Van Dijk, E. (2024). Economic predictors of the subjective experience of financial stress. *Journal of Behavioral and Experimental Finance*, 42, A. 100933. https://doi.org/10.1016/j.jbef.2024.100933.
36. Smathers, K., Chapman, E., Deringer, N. and Grieb, T. (2024). The relationship between financial stress and college retention rates. *Journal of College Student Retention: Research, Theory & Practice*, 26(2), pp. 581-604. https://doi.org/10.1177/15210251221104984
37. Swigonski, N. L., James, B., Wynns, W. and Casavan, K. (2021). Physical, mental, and financial stress impacts of COVID-19 on early childhood educators. *Early Childhood Education Journal,*

49(5), pp. 799-806. https://doi.org/10.1007/s10643-021-01223-z
38. Thayer, Z. M. and Gildner, T. E. (2021). COVID-19-related financial stress associated with higher likelihood of depression among pregnant women living in the United States. *American Journal of Human Biology,* 33(3), A. e23508. https://doi.org/10.1002/ajhb.23508
39. Van Dijk, W. W., van der Werf, M. M. and van Dillen, L. F. (2022). The psychological inventory of financial scarcity (PIFS): A psychometric evaluation. *Journal of Behavioral and Experimental Economics*, 101, A. 101939. https://doi.org/10.1016/j.socec.2022.101939
40. Whitcomb, D., Battaly, H., Baehr, J. and Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research,* 94(3), pp. 509-539. https://doi.org/10.1111/phpr.12228
41. White Jr, K. J. (2020). Financial stress and the relative income hypothesis among Black college students. *Contemporary Family Therapy*, 42(1), pp. 25-32. https://doi.org/10.1007/s10591-019-09531-8
42. Wilson, S., Hastings, C., Morris, A., Ramia, G. and Mitchell, E. (2023). International students on the edge: The precarious impacts of financial stress. *Journal of Sociology,* 59(4), pp. 952-974. https://doi.org/10.1177/14407833221084756
43. Xiao, J. J. and Kim, K. T. (2022). The able worry more? Debt delinquency, financial capability, and financial stress. *Journal of Family and Economic Issues*, 43(1), pp. 138-152. https://doi.org/10.1007/s10834-021-09767-3