Incentive effects on decisions under risk and over time
A Meta-Analysis of Experimental Studies*

Yuchi Li[1] and Ferdinand M. Vieider[1]

[1]*RISLαβ, Department of Economics, Ghent University*

5 June 2026

Abstract

Real monetary incentives are a core principle of experimental economics, yet evidence on whether they materially affect individual decisions under risk and over time remains mixed. We provide a quantitative reassessment by analyzing 624 standardized effect sizes from 69 papers that experimentally vary whether choices are hypothetical or incentivized. We estimate the underlying incentive effect using an outlier-robust Bayesian hierarchical measurement-error model, and supplement it with multiple publication-bias diagnostics embedded within a Bayesian model-averaging framework based on leave-one-out cross-validation. Across all approaches, the estimated incentive effect is extremely small. The posterior mean lies well within the region of negligible effects (Cohen's $d \approx 0.05$). Although true effect sizes are heterogeneous, this variation is only weakly related to study characteristics. Significant moderators include design features such as within- versus between-subjects implementation and decisions involving mixed gain-loss outcomes. Overall, real incentives do not materially alter behaviour in standard individual decision tasks used to measure risk-taking and delay-discounting.

1 Motivation

The use of real monetary incentives has long been a defining principle of experimental economics. Classic contributions argued that financially salient and dominant incentives are essential for eliciting true preferences in the laboratory, ensur-

*This research was supported by the Research Foundation—Flanders under the project "Causal Determinants of Preferences" (G008021N). We are indebted to Michael Birnbaum and Peter Wakker for helpful comments and discussions. We did not register a pre-analysis plan, since the current manuscript contains a meta-analysis of existing papers. All errors remain our own.

ing that participants' decisions reflect the economic tradeoffs under study (Smith, 1982; Plott, 1986). Although this position has shaped decades of experimental practice, the debate over the practical importance—and even the necessity—of real incentives has never been fully resolved. There is widespread agreement that real payments are indispensable in some domains, such as eliciting willingness to pay for socially desirable outcomes (Carson and Groves, 2007). Yet for individual decision-making tasks involving risk or intertemporal tradeoffs, the evidence is more mixed. Early influential studies reported sizable differences between hypothetical and incentivized choices—such as Holt and Laury (2002), who reported a systematic increase in risk aversion in incentivized tasks—whereas more recent high-stakes and large-sample experiments investigating risk-taking and delay-discounting often find negligible or no effects (e.g. Brañas-Garza, Estepa-Mohedano, Jorrat, Orozco and Rascón-Ramírez, 2021; Brañas-Garza, Jorrat, Espín and Sánchez, 2023). As a consequence, the decision of whether to incentivize subjects frequently remains guided more by intuition and convention than by a systematic assessment of empirical evidence.

In this paper, we provide a comprehensive quantitative assessment of the effect of incentive provision on individual decision-making in tasks involving risk and delays. We assemble the universe of experimental studies that vary incentives between real and hypothetical conditions and extract 624 effect-size estimates from 69 papers. By encoding all contrasts in a common metric (Cohen's d), we systematically characterize both the magnitude and the direction of incentive effects across diverse designs and decision contexts. We then analyze these effect sizes using state-of-the-art Bayesian hierarchical meta-analytic methods, which allow us to separate true effects from sampling noise, assess the presence of publication bias, and quantify heterogeneity across studies and experimental features.

Why meta-analysis. Although real monetary incentives are traditionally viewed as essential for eliciting economically meaningful choices, the theoretical rationale for strong incentive effects in individual decision tasks is far from unequivocal. Incentives may suppress experimenter-demand effects or reduce careless responding,

but they need not eliminate systematic violations of expected utility or time-consistent discounting. Classic work on preference reversals, for instance, showed that behavioural regularities persisted or even intensified when salient incentives were introduced (Grether and Plott, 1979), and more recent work documents that many well-known biases remain stable across hypothetical and incentivized settings, even with high stakes (Enke, Gneezy, Hall, Martin, Nelidov, Offerman and Van De Ven, 2023; Gneezy et al., 2024). Thus, theory offers no clear prediction about when or whether incentives should matter in individual decision-making tasks.

Empirically, the existing evidence is similarly inconsistent. Influential early studies such as Holt and Laury (2002) reported substantial differences between hypothetical and incentivized choices on risk-taking propensities, but these findings derive from relatively small samples and highly specific elicitation methods. More recent high-stakes or large-sample studies frequently find much smaller or null effects. Because individual studies differ widely in design, stakes, implementation, and measurement, it is difficult to infer whether such discrepancies reflect true heterogeneity, sampling variation, or selective reporting. A cumulative and principled assessment of the experimental evidence is therefore required.

Relation to previous studies and contribution. Previous meta-analyses have examined related questions, including incentive effects on time discounting (Matousek, Havranek and Irsova, 2022), present bias (Imai, Rutter and Camerer, 2021; Cheung, Tymula and Wang, 2023), and loss aversion (Brown, Imai, Vieider and Camerer, 2024). The studies by Matousek et al. (2022) and Brown et al. (2024), in particular, report little evidence that incentives systematically affect behaviour. Our analysis differs in that we focus exclusively on studies that experimentally vary incentives while holding the underlying choice problem constant. Rather than treating incentive provision as a study characteristic, we encode directly the treatment effects generated by these experimental manipulations. As a result, the estimates synthesized in our meta-analysis inherit a causal interpre-

tation from the randomized comparisons on which they are based. This distinction matters: as we show below, our conclusions regarding incentive effects on risk-taking in mixed-outcome gambles differ from those reported by Brown et al. (2024).

We analyze these effect sizes using a robust Bayesian hierarchical meta-analytic framework that separates true effects from sampling noise, accommodates heavy-tailed heterogeneity, explicitly models publication bias, and incorporates study characteristics through meta-regression. To test for publication bias, we estimate a broad set of models correcting for publication bias and selection effects, and use Bayesian model averaging to stress-test the robustness of our predictions to alternative assumptions about selective reporting. To do so while retaining full modelling flexibility—including heavy-tailed distributions, cross-classification of effect sizes, and a heterogeneous set of selection models—we implement a prediction-optimized Bayesian model-averaging procedure based on leave-one-out cross-validation (Vehtari, Gelman and Gabry, 2017; Yao, Vehtari, Simpson and Gelman, 2018). In contrast to Bayes-factor-based model averaging approaches such as RoBMA (Maier, Matzke, Rouder, Wagenmakers and Ly, 2022), our procedure assigns weights according to predictive performance and remains applicable in settings where Bayes-factor weighting can become unstable. We discuss the underlying technical considerations in the methods section.

Key findings. Our key finding is that the true incentive effect on individual decisions under risk or over time is, on average, too small to be of practical relevance. This holds not only for the pooled average across all studies, but also for most sub-categories, e.g. when analysis is applied separately to risk-taking and delay-discounting, or to gains versus losses. We also find few moderators of these effects in meta-regression, and the ones we do find—such as for risk-taking in mixed gain-loss gambles, and for within- versus between-subject experimental incentive variation—raise important interpretational issues. Finally—and even though this analysis is based on a much smaller set of observations—we also do not find much evidence that real incentives reduce variability in behaviour.

Because the largest reported incentive effects almost exclusively come from small, noisy studies, a natural concern is that these findings may be inflated by selective reporting rather than reflecting genuine behavioural responses to incentives. We therefore examine small-study patterns using a broad set of methods. In addition to standard tests, we deploy a battery of publication-bias models that explicitly adjust the estimated mean effect for potential publication bias or selection effects. Even though the conclusion about whether selective reporting is present differs between approaches, the bias-corrected mean effect remains extremely small and lies well within the negligible region. Thus, while the methods differ in their implied degree of publication bias, they agree that correcting for it does not reveal any substantively meaningful effect of real versus hypothetical incentives.

True heterogeneity. Although the mean effect is negligible, the estimated heavy-tailed distribution of true effect sizes indicates that incentive effects vary meaningfully across studies, with a nontrivial proportion of studies exhibiting small but real deviations in either direction. We find no differences in incentive effects between decisions in the risk and time domain. By contrast, incentive effects are more pronounced in the loss domain and especially in mixed gain-loss tasks than for gains. Meta-regression reveals that real incentives make subjects considerably more risk *seeking* for mixed gain-loss lotteries. This pattern is consistent with house-money or endowment-integration mechanisms inherent in standard implementations of monetary losses. Other study characteristics—such as whether all subjects are paid, whether all decisions are incentivized, whether the study was conducted online or in the field, or whether it was published in economics—explain little additional variation.

Taken together, these results lead to a strikingly robust conclusion: *real monetary incentives have, on average, no meaningful impact on individual decision-making under risk or over time.* The small pockets of heterogeneity that do emerge are readily explained by features of incentive implementation rather than by genuine motivational effects of incentives. In particular, the deviations we observe in mixed gain-loss choices arise in domains where incentives are almost always

implemented through loss-from-endowment procedures, which are known to induce house-money effects. These patterns therefore reflect the psychology of loss implementation, not heightened sensitivity to monetary incentives.

Our results are particularly striking when viewed alongside recent evidence from other domains. A notable example is the meta-analysis of Cala, Havranek, Irsova, Luskova, Matousek and Novak (2026), which examines the effect of financial incentives on performance across a broad range of economics experiments. Despite focusing on a different class of outcomes, Cala et al. (2026) similarly find that incentive effects are generally small after correcting for publication bias, with somewhat stronger effects under loss framing. Taken together, their findings and ours suggest that the behavioural impact of monetary incentives may be considerably smaller than traditionally assumed, and that the most robust departures from this pattern arise in environments involving losses.

2 Methods

2.1 Literature Search and Study Selection

We conducted a comprehensive search for empirical studies comparing decisions made under *real incentives* with decisions made under *hypothetical outcomes* in tasks involving risk, uncertainty, or intertemporal choice. We carried out the primary search in April 2026 using Web of Science (All Databases), without restrictions on publication year or document type.

Our search terms captured contrasts between hypothetical and real rewards as well as decision contexts involving risk or delay. We screened reference lists of all identified studies, and we supplemented the search using Peter Wakker's annotated bibliography, which contains a dedicated category for variation in incentives. We furthermore circulated our list of studies on the ESA and JDM-society mailing

lists to elicit any studies we might have missed.[1] This process yielded 602 initial records and 238 additional records from backward citation searches, bibliographic sources and society feedback.

We included studies if they (i) compared hypothetical and real incentives using behavioral measures, and (ii) held constant the ranges of reward magnitudes, probabilities, or delays across conditions. These criteria ensure that incentive effects are not confounded with known context effects such as the magnitude effect in temporal discounting or stake effects across outcome ranges. We excluded seventy-five studies because they violated this design requirement. The full dataset contains 73 papers, including 25 temporal discounting studies and 55 risk-taking studies (some papers include both risk and delay tasks, and are thus counted in both categories). Online Appendix A provides full search terms, details on inclusion and exclusion criteria. Online Appendix L lists all included papers.

2.2 Coding of Effects and Study Characteristics

For each study, we coded a measure of the effect of incentives on choice behaviour. A first challenge arose from the wide variation in reporting standards: many papers did not focus explicitly on incentive effects or reported them only indirectly. When papers reported multiple effect sizes—for example, because they included several experimental tests, reported both nonparametric and structural estimates, or estimated multiple behavioural parameters from the same structural model—we included all eligible effects (we will explicitly account for their statistical dependence in the hierarchical structure of our model). After excluding five effects for which the direction of the effect could not be established, we were left with 69 papers containing 86 distinct experiments, which between them contribute a total of 624 effect sizes. These constitute the primary unit of analysis.

[1] While we did write up an initial conceptual note detailing the specific issues likely to arise in our meta-analysis and how to deal with them, we ultimately did not pre-register this note because it remained necessarily vague. In meta-analysis, many methodological choices flow from data features, as is exemplified by our analysis of publication bias below. In the absence of precise procedural guidelines, we thus believe that striving for open and transparent reporting and including robustness analyses to various modelling assumptions is the best option at our disposal.

To compare effects across heterogeneous reporting formats, we converted all incentive contrasts to *Cohen's d* (Cohen, 1988). When papers reported group means and standard deviations, we computed d using the pooled standard deviation. When papers reported inferential statistics (e.g., t, F, or z statistics, or regression coefficients), we converted these to d using standard transformations. Separate formulas were used for between-subject and within-subject designs to account for the corresponding correlation structures in choice behaviour. Online Appendices C and D provide the full formulas for computing effect size d and its standard error. To further allay concerns that may arise when applying Cohen's d to choice proportions, we additionally calculated Cohen's h in these cases. Online Appendix E compares the result for this subset of studies, and finds no significant differences between the two measures.

A second challenge arose from the diversity of choice architectures. A substantial majority of effect sizes were derived from *nonparametric* behavioural measures. These include proportions of patient versus impatient choices in intertemporal tasks, proportions of risky versus safe choices under risk, indifference points, and Area Under the Curve (AUC) measures. In total, 458 of the 624 effect sizes fall into this category. For intertemporal choice, we coded these measures so that larger values of Cohen's d indicate *greater impatience* under real incentives. Analogously, for risky decisions we coded nonparametric contrasts so that larger values correspond to *increased risk aversion*.

The remaining effect sizes were derived from *parametric* estimations. These include parameters capturing constructs such as utility curvature or loss aversion, as well as discounting parameters in intertemporal choice. We only included parametric measures when their directional interpretation in terms of risk aversion or impatience was unambiguous. For example, utility curvature parameters consistently map onto risk aversion; proportional discounting parameters such as k in $1/(1+kt)$, where t is the time delay, likewise provide a monotonic proxy for impatience. Our dataset contains quasi-hyperbolic (β–δ) discounting estimates. Quasi-hyperbolic estimates were included because both β and δ^{-1} can be inter-

preted as proxies for impatience in a behaviourally consistent manner. We did not include fully hyperbolic (or time sensitivity) parameters, nor probability sensitivity parameters, since they do not map monotonically onto impatience or risk aversion and would therefore not allow for consistent coding.

In addition to the key effect sizes, we coded major design features of each study, including the experimental setting (laboratory, field, online), the subject population (students vs. general population), reward type (monetary vs. non-monetary), decision domain (gains, losses, mixed), and the incentive scheme (e.g., paying a subset of subjects, paying one randomly selected choice, paying all choices). Online Appendix F provides the full list of characteristics and operational definitions.

Finally, because parametric and nonparametric effect sizes may differ systematically in scale, noise properties, and behavioural interpretation, we conduct extensive robustness checks and include parameter type as a moderator in our meta-regression analyses. These analyses confirm that our main results are not driven by differences between nonparametric and parametric measures.

3 Aggregate results

We present our findings in several stages, starting from an aggregate analysis. We first describe nonparametric patterns in the data. We then estimate the meta-analytic mean using a hierarchical Bayesian measurement error model and examine the posterior inferences based on that model.

3.1 Incentivized vs hypothetical studies: raw effects

Descriptive results. We begin by presenting descriptive evidence on the distribution of effect sizes across all decision types and outcome domains. Panel A of Figure 1 plots the density of the *absolute values* of all 624 encoded effect sizes d_i. It also plots the distribution of the 458 effect sizes based on non-parametric measures for comparison. The distribution is heavily concentrated near zero: fully 52% of

all effect sizes are smaller than 0.2. Thus, more than half of the reported effects do not even reach Cohen's threshold for a "small" effect. Both the mode and median effect sizes are therefore best characterized as negligible. An additional 35% of effect sizes fall into Cohen's "small" category, while medium-sized and large effects are rare, at 9% and 4% respectively. The distribution of nonparametric effects closely resembles the overall distribution.

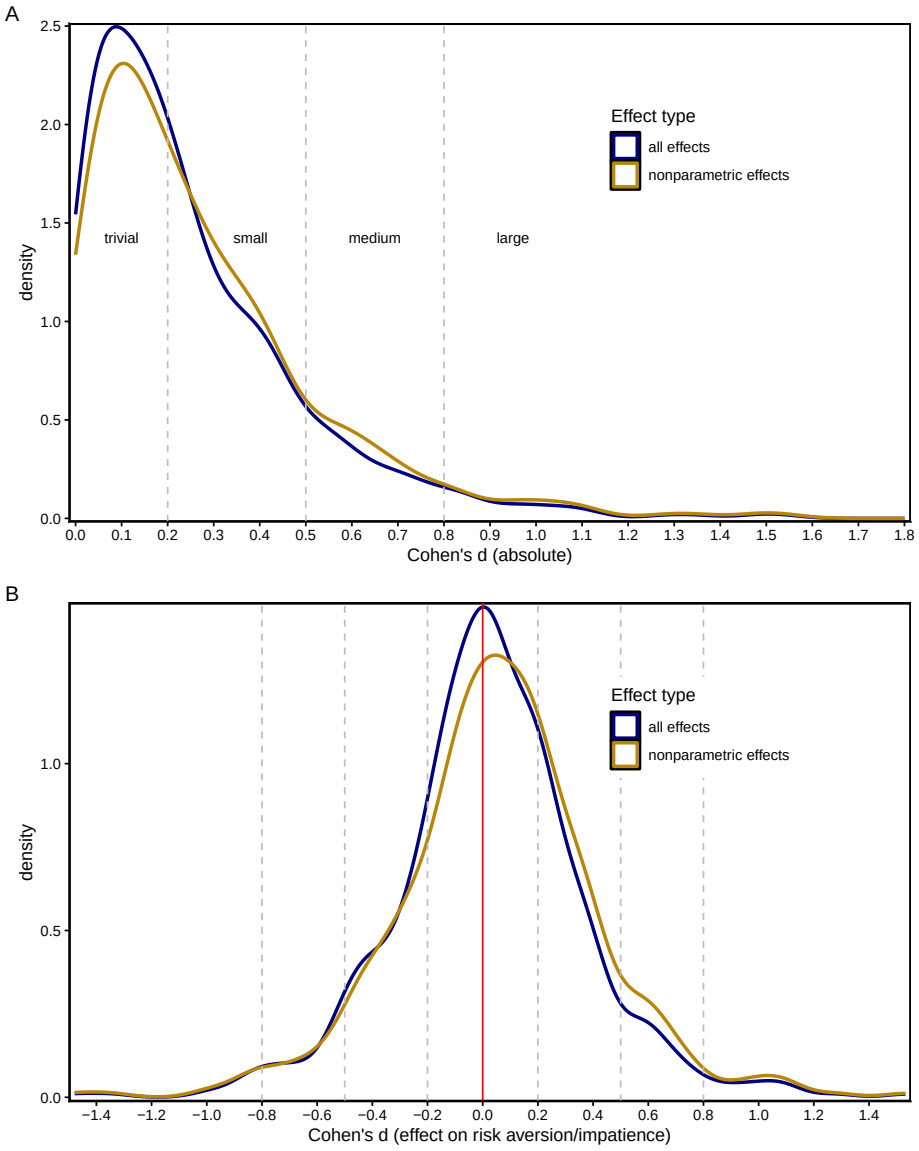

Figure 1: Probability density of Cohen's d

Distribution of 624 Cohen's d effect sizes across studies. Panel A shows the distribution of absolute effect sizes. Panel B shows the distribution preserving the sign of the effect, with positive values corresponding to increased risk aversion or impatience under real incentives, and negative values corresponding to increased risk seeking or patience.

Because Panel A represents absolute effect sizes, it does not capture the *direction* of the incentive effect. Beyond whether incentives have any effect on behaviour, an important question is whether the reported effects exhibit a consistent directional pattern. To address this, we coded all effect sizes so that positive values indicate greater risk aversion (or impatience) under real incentives, whereas negative values indicate greater risk seeking (or patience).

Panel B of Figure 1 plots the distribution of signed effect sizes on the negative–positive continuum (separately using the full data and the nonparametric measures only). The distribution is strikingly symmetric around zero: the mode (= 0), median (= 0.021), and mean (= 0.024) all lie extremely close to zero. Moreover, larger positive effects (greater risk aversion or impatience under real incentives) are almost perfectly counterbalanced by larger negative effects (greater risk seeking or patience under real incentives). In short, the descriptive evidence reveals *no coherent directional pattern* in the literature. Incentive effects pointing in one direction are nearly exactly offset by effects pointing in the opposite direction.

3.2 Hierarchical Bayesian Model

To analyze the incentive effects statistically, we develop a Bayesian Hierarchical Measurement-Error Model (BHMEM). A hierarchical (random-effects) specification allows us to model genuine cross-study variation in true effect sizes, which we expect to be present based on the large number of studies and the diversity of behavioural tasks included in this meta-analysis. The Bayesian framework provides additional flexibility to extend the model in response to substantive questions that arise from the data.

Hierarchical Bayesian Model with Experiment Cross-Classification. Our basic unit of analysis is Cohen's d, representing the observational estimate of the incentive effect in study i. Each observed effect size d_i is measured with sampling uncertainty, for which we use the reported sampling variance se_i^2. Following

standard meta-analytic practice, we assume the measurement-error model

$$d_i \sim \mathcal{N}\Big(\widehat{d}_i,\ se_i^2\Big)\,, \tag{1}$$

where $\widehat{d}_i$ denotes the *true* (latent) effect size underlying study i. The hierarchical model then specifies how these latent true effects vary across studies. To accommodate potential outliers and true heterogeneity across studies, we model the distribution of the true effects using a Student–t specification:

$$\widehat{d}_i \sim \text{Student-}t(\nu,\ \mu + \gamma_x,\ \sigma)\,, \tag{2}$$

where μ is the population-level mean effect, σ is the between-study scale parameter (with σ^2 representing the variance in true effect sizes across studies), and where $\nu \geq 2$ denotes the degrees of freedom. Estimating ν lets the data determine the appropriate tail behaviour: small values of ν yield heavy tails that downweight outlying effect sizes, while for large values of ν the distribution is approximately normal. This makes the model robust to outliers without imposing strong assumptions about their presence.

In addition, we cross-classify effects to account for statistical dependency:

$$\gamma_x \sim \mathcal{N}(0,\ \tau_x^2). \tag{3}$$

The term γ_x introduces an experiment-level random effect, ensuring that all effect sizes originating from the same experiment share a common shift relative to the overall mean. This induces the appropriate correlation among within-experiment estimates and prevents single experiments for which many outcomes are reported from disproportionately influencing the meta-analysis.

Under this specification, each observed effect size d_i combines two sources of variability: (i) sampling variance se_i^2, arising from measurement error in the individual study, and (ii) hierarchical variance components governed by σ and τ_x^2, capturing genuine heterogeneity in the underlying true effects across studies and experi-

ments. The hierarchical model thus separates study-level noise from substantive cross-study variation and shrinks noisy estimates toward the overall mean μ.

Robustness to correlations between effects and errors. Calculating standardized effect sizes as Cohen's d risks introducing spurious correlations between the effect sizes themselves and their standard errors. To assess the robustness of our inferences, we therefore repeat all analyses in the paper using the MAIVE instrumental-variable approach proposed by Irsova, Bom, Havranek and Rachinger (2025). MAIVE replaces the reported squared standard error, SE_i^2, with the component predicted by sample size, $1/N_i$, which is not mechanically linked to the estimated effect size. Specifically, we estimate $SE_i^2 = \widehat{\psi}_0 + \widehat{\psi}_1 \frac{1}{N_i} + \nu_i$ and construct the instrumented standard error as $SE_i^{IV} = \sqrt{\widehat{\psi}_0 + \widehat{\psi}_1 \frac{1}{N_i}}$. We then re-estimate our main meta-analytic specifications using this instrumented precision measure. Online Appendix I reproduces the entire analysis in this paper using this approach and leaves all substantive conclusions unchanged.

Bayesian posterior inferences. We now deploy our BHMEM to study the distribution of true effect sizes. We estimate the model in Stan (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li and Riddell, 2017) using mildly regularizing hyperpriors Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2014), chosen to be at least one order of magnitude wider than the plausible ranges suggested by the data. The results reported here are not sensitive to reasonable variations in these hyperpriors. We assessed convergence by verifying the absence of divergent transitions and ensuring that all $\widehat{R}$ statistics are very close to 1, with $\widehat{R} \leq 1.01$ accepted as an indication of satisfactory mixing. Online Appendix G reports full details, including the Stan code used; Vieider (2024) provides a tutorial introduction to Bayesian hierarchical modeling in Stan.

The estimated degree-of-freedom parameter of the Student–t distribution is $\nu = 2.147$ (95% CrI [2.004, 2.519]), confirming substantial tail heaviness and thus validating the Student–t specification as an outlier-robust choice. The estimated meta-analytic mean is $\mu = 0.045$. Its 95% credible interval (CrI) includes zero,

$[-0.002,\ 0.093]$, and falls entirely within the range of negligible effect sizes. Panel A of Figure 2 compares the raw effect sizes d_i to the posterior distribution of true effect sizes $\widehat{d_i}$. The posterior distribution is substantially narrower, with two-thirds of all $\widehat{d_i}$ falling in the negligible-effect interval $[-0.2, 0.2]$. This illustrates meta-analytic shrinkage: each d_i is pulled toward the meta-analytic mean in proportion to its standard error. The extent of shrinkage suggests that studies reporting larger effects tend to be relatively noisy—a point to which we will return below.

Meta-analytic shrinkage affects not only the estimated effect sizes but also the precision with which they are estimated. Because the posterior standard deviations sd_i of the true effects (equivalent to a standard error in frequentist statistics) incorporate both sampling variability and shrinkage toward the meta-analytic mean, they are typically smaller than the raw standard errors se_i. The implications for statistical significance are therefore ambiguous ex ante: shrinkage pulls effects toward zero, but it also reduces the uncertainty surrounding the estimate.

To assess statistical significance in our Bayesian framework, we define a *region of practical equivalence* (RPE) around the null hypothesis of no effect. Following Cohen's conventions, it is natural to take the interval $[-0.2, 0.2]$ to represent negligible effects. We classify a study as "positive" if at least 95% of its posterior mass lies above 0.2, and as "negative" if at least 95% lies below -0.2. Studies with at least 95% of their posterior mass within the RPE are classified as unambiguously negligible.[2] All remaining studies are classified as "uncertain", inasmuch as they do not provide sufficient information for unambiguous classification.

Panel B of Figure 2 plots the posterior mean of the true effect size $\widehat{d_i}$ against its standard deviation sd_i, and colour-codes the points depending on their classification. Overall, 3.7% of studies show a positive effect and 3.5% a negative effect. The fact that clearly positive and clearly negative effects occur in both directions suggests the presence of moderate genuine heterogeneity in true effects—a topic we examine at some length below. By contrast, 16.3% of studies are unambigu-

[2]The choice of a 95% posterior probability threshold follows statistical convention; other thresholds would lead to qualitatively similar conclusions.

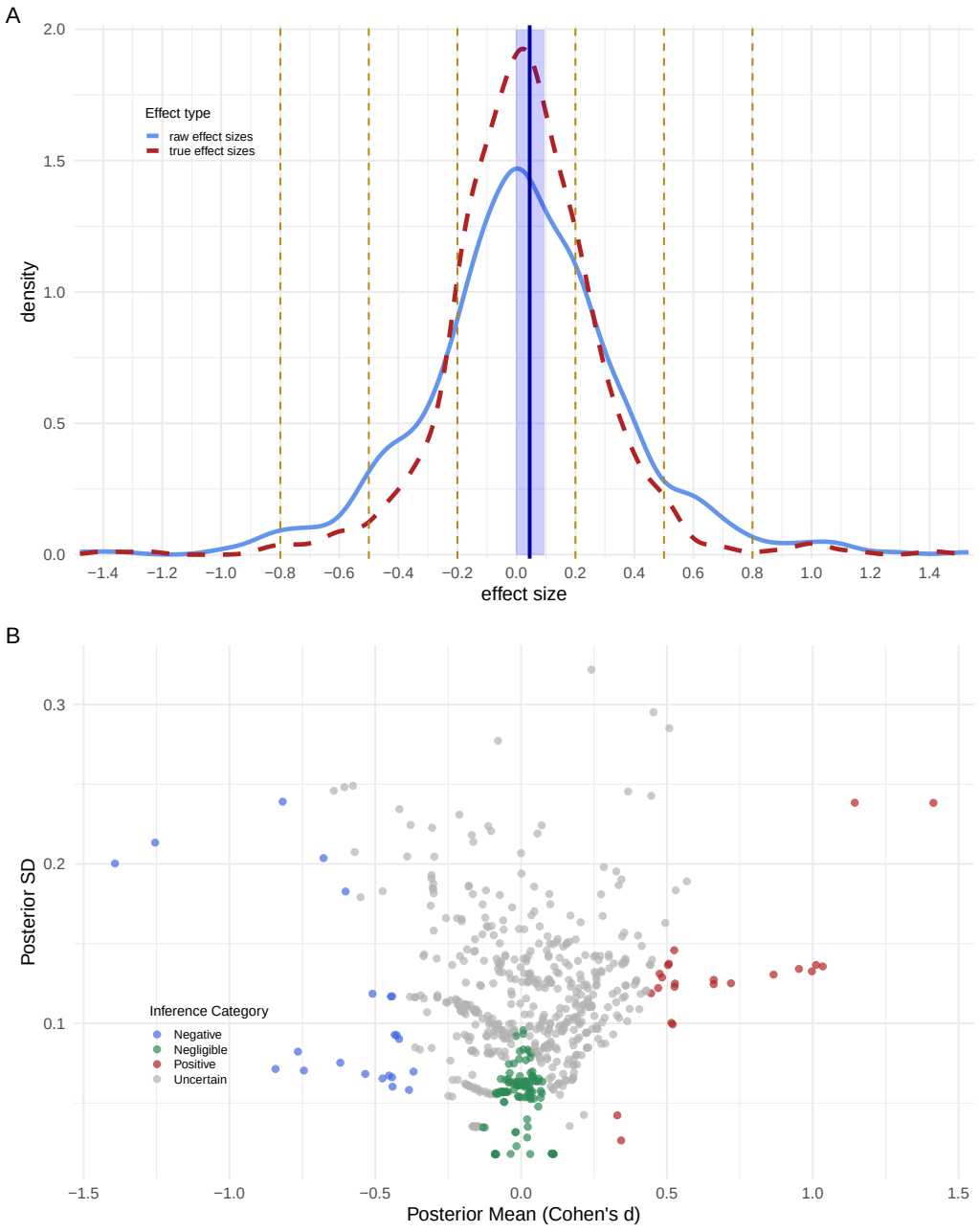

Figure 2: Posterior inferences on effect sizes

Posterior inferences from the BHMEM. Panel A compares the distribution of raw effect sizes d_i with the posterior distribution of true effect sizes $\widehat{d}_i$. Panel B plots the posterior means $\widehat{d}_i$ (x-axis) against their posterior standard deviations (y-axis), which correspond to standard errors in frequentist terminology.

ously negligible. These provide *positive evidence of absence* of incentive effects, not merely absence of evidence: for these studies, a practically null effect is genuinely likely. Such cases are more than twice as common as positive and negative effects *combined*. Finally, 76.4% of studies are too imprecise or too small to yield

a clear conclusion, reflecting low power or unfavorable signal-to-noise ratios.

4 Is there publication bias?

The meta-analytic results above indicate that larger effects—whether positive or negative—tend to be associated with greater sampling noise. This may simply reflect sampling variation in small studies, but it may also be symptomatic of publication bias. We therefore begin with a set of nonparametric diagnostics.

4.1 Nonparametric examination of small study effects

Panel A of Figure 3 presents a funnel plot of the raw effect sizes d_i against $-\ln(se_i)$, a measure of the precision of the effect. More precise studies thus appear at the top of the graph. The figure shows a clear pattern: the most extreme effect sizes (which are both positive and negative) occur almost exclusively in imprecise studies, whereas the most precise studies tend to cluster around zero, with only a few small positive exceptions. In the absence of small-study effects, the estimated effect sizes should not systematically vary with their standard errors; precise and imprecise studies would be centered on the same underlying value.

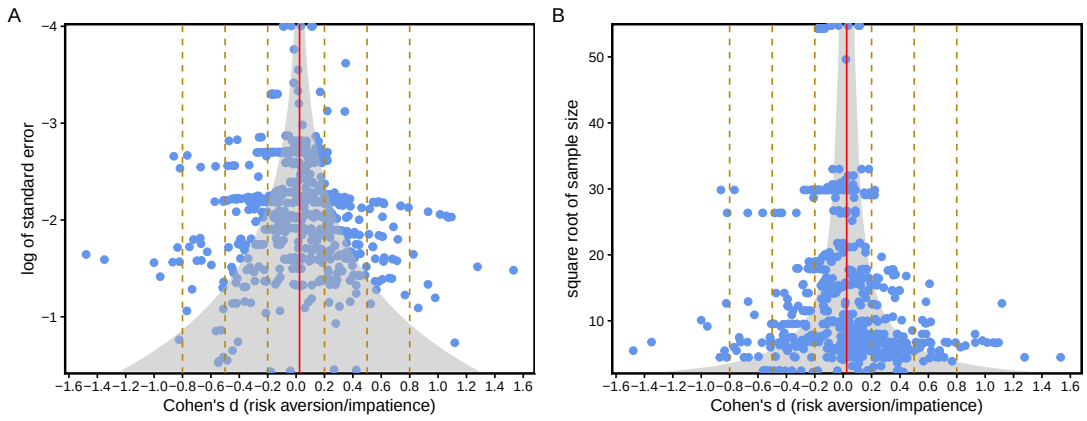

Figure 3: Funnel plot of Cohen's d against its log-standard error

The figure plots the raw effect sizes d_i against $-ln(se_i)$ (panel A) and against $\sqrt{N}$ (panel B). The gray area in panel A indicates a zone containing non-significant results. The gray area in panel B provides a similar measure, given by $\frac{1.5}{\sqrt{N}}$. The scaling factor of 1.5 is used because it approximates the average *standard deviation* in the sample.

Panel B of Figure 3 plots effect sizes against $\sqrt{N}$ to display the same relationship in terms of sample size. The conclusions remain unchanged: small studies generate nearly all large positive and large negative effects, whereas larger studies (with a few exceptions) converge toward negligible effects. The correlations between effect size and precision are sizable: $|d_i|$ is negatively correlated with $\sqrt{N}$ ($\rho = -0.348$, $p < 0.001$), with similar patterns for positive effects ($\rho = -0.360$, $p < 0.001$) and negative effects ($\rho = -0.340$, $p < 0.001$). Signed effects also show a negative relationship with precision ($\rho = -0.220$, $p < 0.001$). These findings indicate pronounced small-study effects. Although such effects do not necessarily imply publication bias, publication bias is a common mechanism capable of producing these patterns.

A standard diagnostic for funnel-plot asymmetry is Egger's regression, which regresses the standardized effect size d_i/se_i on study precision $1/se_i$. Under the null of no small-study effects, the intercept should be zero. Applied to absolute effect sizes, Egger's test yields a strongly positive intercept ($\beta_0 = 1.261$, 95% CrI $[1.052, 1.470]$; slope $\beta_1 = 0.068$, 95% CrI $[0.052, 0.084]$), indicating that small, imprecise studies tend to report disproportionately large deviations from zero. In Egger's framework, this is the classical pattern consistent with publication bias. When applied to signed effect sizes, however, the pattern disappears: the intercept is small and uncertain ($\beta_0 = 0.277$, 95% CrI $[-0.028, 0.596]$), and the slope is near zero ($\beta_1 = -0.021$, 95% CrI $[-0.046, 0.002]$). This divergence is informative: it implies that small studies tend to report extreme effects, but not systematically in the positive or negative direction. In other words, the small-study pattern we observe is about magnitude, not sign.

Such symmetric exaggeration is compatible with publication bias (journals preferentially publishing "large" effects in either direction), but it is also compatible with genuine heterogeneity combined with sampling noise. Because Egger's test relies on assumptions that are violated in our setting—normality of true effects, homogeneity across studies, and statistical independence of effect sizes—the contradictory signals between absolute and signed versions cannot be taken as defini-

tive evidence of bias. Instead, they point toward the need for explicit parametric models of selection, which we turn to next.

4.2 Formal tests and adjustments for publication bias

The effects documented above are indicative of small-study effects, in the sense that smaller and less precise studies are more likely to show significant effects in either direction. We next examine whether the distribution of effects is indicative of publication bias—the tendency of null results to be less likely to be written up by authors or published by journals. To do this, we field a battery of commonly used methods that can adjust and correct meta-analytic estimates for publication bias and selection effects. Here, we briefly list the methods and describe their key characteristics; Online Appendix H.1 provides a more extensive discussion including technical details. The tests we use are

- **PET–PEESE:** The Precision-Effect Test (PET) and the Precision-Effect Estimate with Standard Error (PEESE) are regression-based tools designed to detect and correct for publication bias by exploiting the empirical relationship between reported effect sizes and their standard errors (Stanley, 2008; Stanley and Doucouliagos, 2014). Both approaches regress reported effect sizes on a measure of their precision: the standard error in PET and its square in PEESE.

- **Vevea & Hedges selection model:** the Vevea and Hedges (1995) model explicitly models the probability of a study being selected for publication. The approach combines two components: (i) an effect-size model, analogous to our Bayesian Hierarchical Measurement Error Model (BHMEM), that characterizes the distribution of study outcomes in the absence of selective publication, and (ii) a selection model that assigns relative probabilities to studies based on the p-value associated with their effect estimate. This formulation yields effect-size estimates that adjust for selective reporting and allows formal inference on the presence of publication bias. We estimate both *unidirectional* (V&H-UD) and *bidirectional* (V&H-BD) versions of the

model. In the unidirectional specification, p-values are based on $|z_i|$, imposing symmetric selection weights for positive and negative effects. The bidirectional specification computes p-values from the signed test statistic, allowing selection probabilities to differ between positive and negative effects.

- **Andrews & Kasy selection model:** The Andrews and Kasy (2019) approach explicitly models both the distribution of true effects and the selection mechanism governing which results are observed in the published sample. Unlike the Vevea & Hedges model, which estimates relative selection weights across p-value regions, the A&K model directly parameterizes the probability that a result enters the observed sample. We estimate two versions: one with smooth global patterns (QI), and one allowing for flexible local nonlinearities (NS).

For completeness and comparability with the existing literature, we estimate all of these models in two different settings. First, we use implement them in standard fixed-effects settings. Second, we implement Bayesian hierarchical versions of all models by embedding their structure inside our baseline random-effects BHMEM. Online Appendix H.1 provides the technical details, and Online Appendix H.2 the statistical code.

Selection patterns implied by the individual publication-bias models. Each of the approaches discussed above captures a different aspect of publication selection. PET–PEESE detects *small-study effects* through the association between effect sizes and their standard errors; the Vevea & Hedges model identifies *discrete jumps* in publication probability across p-value intervals; and the Andrews & Kasy model estimates a smooth *selection function* describing the absolute probability that a study with a given test statistic enters the published literature. Figure 4 shows that the Vevea–Hedges and Andrews–Kasy models recover markedly different selection mechanisms. The V&H model, which imposes stepwise changes at conventional p-value thresholds, produces the expected pat-

tern: a sharp increase in publication probability for statistically significant results. In the case of the bidirectional specification, the increase in selection probability is less pronounced for significantly *negative* results—which are nonetheless much more likely than nonsignificant effects to be published.

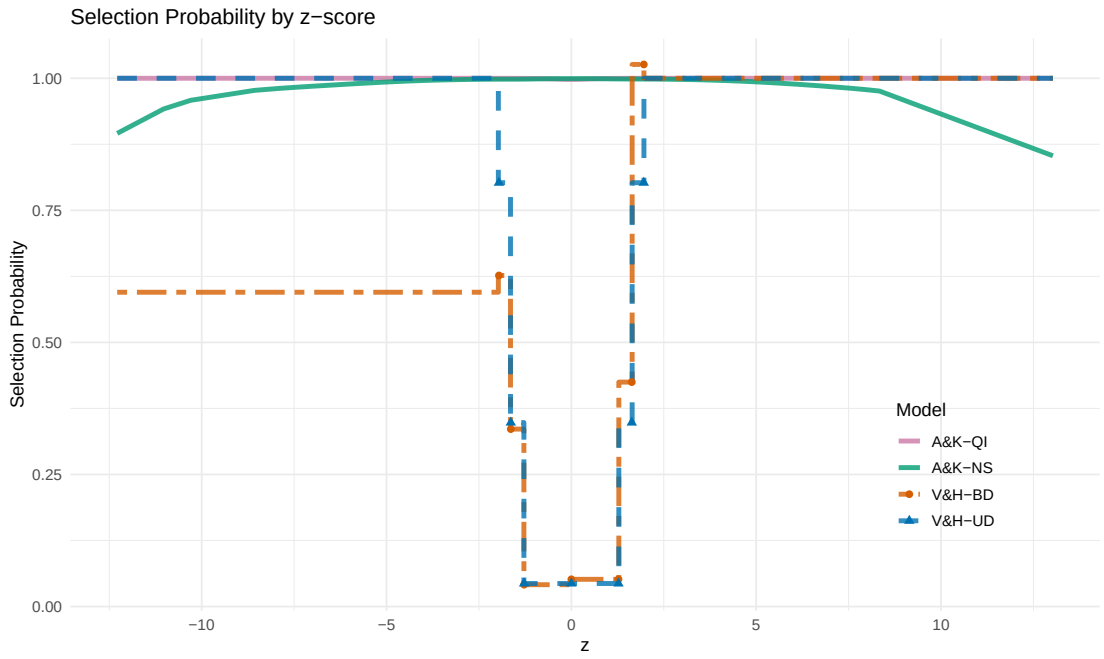

Figure 4: Posterior mean selection probabilities for the Vevea & Hedges and Andrews & Kasy models under random-effects specifications.

In contrast, the A&K model places no structural restrictions on how selection varies with the test statistic. In our dataset, the estimated selection functions are relatively flat and do not exhibit a strong increase around the 5% significance threshold; in some regions the publication probability even declines for extremely large z-values (possibly due to few effect sizes in those regions). This divergence from the V&H pattern reflects the much greater flexibility of the A&K specification, as well as the substantial heterogeneity of our sample: many studies with non-significant or modest effects appear to have been published, while extreme effects may not receive disproportionately more weight in the published literature. Rather than contradicting V&H, the A&K model therefore captures a different, smoother dimension of selection that need not align with stepwise threshold effects.

The PET–PEESE diagnostics paint a more nuanced picture. The fixed-effect PET model yields a positive and statistically significant slope ($\lambda_{\text{PET,FE}} = 0.28$, $p < 0.001$), suggesting that smaller and noisier studies tend to report larger effects. However, once between-study heterogeneity is accounted for, the slope becomes statistically indistinguishable from zero ($\lambda_{\text{PET,RE}} = -0.07$, $p = 0.374$). A similar pattern emerges for the PEESE specification: the fixed-effect model detects a significant positive slope ($\lambda_{\text{PEESE,FE}} = 1.15$, $p < 0.001$), whereas the random-effects model does not ($\lambda_{\text{PEESE,RE}} = -0.10$, $p = 0.412$). These differences imply that the apparent small-study effects in the raw data are closely tied to genuine heterogeneity across studies rather than to selective publication alone.

Importantly, the bias-adjusted intercepts (interpreted as the effect size for an infinitely precise study) remain close to zero under the random-effects specifications ($\mu_{\text{PET,RE}} = 0.055$, $\mu_{\text{PEESE,RE}} = 0.047$; see Table 1 for statistical information), both lying well within the range of negligible effects. Thus, while PET–PEESE detects small-study patterns under a fixed-effect formulation, the Bayesian hierarchical versions of these models do not provide strong evidence of systematic publication bias once study-level heterogeneity is incorporated.[3]

4.3 Results from Robust Bayesian Model Averaging

To synthesise the evidence across all publication-bias models, we implement a Prediction–Optimised Bayesian Model Averaging (PoBMA) framework. PoBMA evaluates a broad family of meta-analytic specifications—including the baseline measurement-error model (BHMEM), PET–PEESE, Vevea & Hedges selection models, and the continuous-selection models of Andrews & Kasy—and combines them into a single posterior distribution of the underlying effect.

[3] A potential concern could be that, in the hierarchical PET–PEESE models, the heavy-tailed random-effects distribution might absorb patterns that would otherwise be attributed to publication bias, thereby driving the PET–PEESE slope toward zero. This is not the case: the PET–PEESE slope is identified from the *within-study* relationship between $\widehat{d}_i$ and se_i, whereas the Student–t random-effects distribution captures *between-study* heterogeneity in latent true effects. These components are orthogonal in the likelihood, so heterogeneity cannot generate or eliminate a dependence of effect sizes on their standard errors.

The PoBMA is based on stacking weights derived from leave-one-out cross-validation (LOO; Vehtari et al., 2017; Yao et al., 2018). This differs from approaches implemented in existing model-averaging software such as `RoBMA` (Maier et al., 2022), which uses Bayes factors to determine the weights assigned to different models. The key distinction is that PoBMA prioritizes out-of-sample predictive performance rather than marginal-likelihood-based model evidence. As a result, model weights need not follow the ranking of individual models: they are chosen to maximize the predictive performance of the combined model. More importantly for the present application, LOO-based stacking remains stable under the modelling choices required by our data. Our framework combines heterogeneous families of publication-bias models while allowing for heavy-tailed heterogeneity and cross-classified dependence structures. In this setting, Bayes-factor calculations become numerically unstable and do not provide a reliable basis for model averaging.[4] Stacking allows us to retain these modelling features and still perform coherent model averaging across the full set of candidate specifications.

Each model is estimated under both fixed-effect and random-effects specifications, yielding a total of 14 candidates. This framework allows the data to determine (i) which publication-bias corrections receive empirical support, and (ii) the resulting bias-adjusted mean effect size μ. Table 1 reports the LOO differences, stacking weights, and posterior means for μ for all models. Several patterns emerge clearly.

First, the first four models have similar predictive power, and all receive relatively large weights. Notably, more complex models such as PET, PEESE, and the A&K specifications do not perform meaningfully better than our simple BHMEM, which serves as their common baseline specification.[5] Whereas these models neither improve upon nor perform substantially worse than the BHMEM, the Vevea

[4]This instability arises because the models being averaged differ substantially in parameterization and dimensionality, while also incorporating highly flexible distributional assumptions.

[5]Stacking weights need not mirror the ordering of models by Δelpd. Their purpose is to maximise the predictive performance of the *combined* model, meaning that a specification with weaker standalone predictive accuracy may nevertheless receive a positive weight if it contributes complementary predictive variation.

Model	**Δelpd**	**weight**	μ
Random PEESE	0.0 (0.0)	0.261	0.047 (-0.007, 0.102)
Random PET	-1.1 (3.5)	0.239	0.055 (-0.017, 0.133)
Random MEM	-2.9 (4.1)	0.146	0.046 (0.000, 0.094)
Random A&K-NS	-3.5 (3.9)	0.259	0.045 (-0.002, 0.092)
Random A&K-QI	-4.8 (3.7)	0	0.045 (-0.002, 0.093)
Random V&H-BD	-8.6 (5.1)	0.095	0.029 (-0.022, 0.080)
Random V&H-UD	-65.0 (9.8)	0	0.033 (-0.005, 0.075)
Fixed V&H-UD	-1559.6 (182.3)	0	0.000 (0.000, 0.000)
Fixed PEESE	-1560.5 (185.7)	0	-0.012 (-0.019,-0.005)
Fixed PET	-1564.0 (188.5)	0	-0.022 (-0.031,-0.012)
Fixed A&K-NS	-1567.4 (184.0)	0	-0.005 (-0.011, 0.001)
Fixed A&K-QI	-1567.7 (184.1)	0	-0.005 (-0.011, 0.001)
Fixed MEM	-1568.0 (184.1)	0	-0.005 (-0.011, 0.002)
Fixed V&H-BD	-1568.8 (185.0)	0	-0.009 (-0.016,-0.003)

Table 1: Leave-one-out cross-validation (LOO) results and stacking weights for the 14 model specifications included in the Prediction-optimized Model Averaging (PoBMA) framework. The 'Model' column lists each specification, indicating fixed- or random-effects assumptions and the type of publication bias correction applied: hierarchical Bayesin measurement error model (MEM), Precision-Effect Test (PET), Precision-Effect Estimate with Standard Error (PEESE), Vevea & Hedges selection models under unidirectional (V&H-UD) or bidirectional (V&H-BD) specifications, and Andrews & Kasy selection models using natural spline (A&K-NS) or quadratic interpolation (A&K-QI). Δelpd denotes the difference in expected log predictive density relative to the best-fitting model (standard error in parentheses), 'weight' indicates the model stacking weight, and μ gives the posterior mean of the bias-corrected effect size (with 95% credible intervals in parentheses).

& Hedges stepwise-selection models display poorer predictive performance (with the unidirectional specification nonetheless contributing a small positive weight). By contrast, all fixed-effect specifications receive effectively zero weight, reflecting the substantial between-study heterogeneity in our data and the incompatibility of fixed-effect assumptions with the empirical structure of incentive-effect estimates.

Second, across all random-effects models, the bias-adjusted population mean remains small. Posterior means range from approximately 0.029 (V&H-BD) to 0.055

(PET), and all associated 95% credible intervals lie entirely within the "negligible" range of $[-0.2, 0.2]$. No model yields credible evidence for a substantively meaningful effect of real versus hypothetical incentives. We compute the overall model-averaged estimate of the underlying effect size by combining the individual models using their stacking weights. The resulting posterior mean is

$$\mu_{\text{PoBMA}} = 0.047 \quad (95\% \text{ CrI } [0.020,\ 0.074]),$$

which again falls squarely within the negligible region. PoBMA therefore confirms that, after accounting for publication bias and heterogeneity using a wide range of correction methods, the meta-analytic effect of real incentives on choice behaviour remains very small.

Taken together, the PoBMA results reinforce the main conclusion from the individual publication-bias models: although the raw data exhibit small-study patterns, once we account for heterogeneity and alternative mechanisms of selective reporting, the underlying effect of incentive provision on decision behaviour is very small and likely negligible.

5 Variability in true effect sizes

We next examine whether the residual variability in true effect sizes—after accounting for sampling error and shrinkage—is systematically related to characteristics of the underlying studies. In particular, we start by presenting disaggregated results by decision type, decision domain, and along various task characteristics. We then move on to meta-regression analysis of the effects.

5.1 Meta-analytic means and variation by category

While the aggregate analysis above provides a useful overall summary, it may obscure meaningful differences across domains and task types. If incentive ef-

fects vary systematically—for example, between risk-taking and delay-discounting tasks—aggregate estimates may conceal such heterogeneity, potentially attenuating effects that operate in opposite directions. To address this concern, we now report meta-analytic means at a more disaggregated level.

Subgroup characteristics	Raw Effects			PoBMA Estimates			Obs.
	Mean	Med.	SD	Mean	Med.	SD	*N*
Topic Domain							
Temporal discounting	-0.049	-0.002	0.290	-0.030	-0.007	0.185	119
Risk taking	0.042	0.028	0.353	0.036	0.028	0.278	505
Source of Data							
Lab experiment	0.056	0.068	0.383	0.053	0.061	0.290	450
Field experiment	-0.043	-0.072	0.106	-0.040	-0.068	0.094	54
Online experiment	-0.074	-0.077	0.204	-0.070	-0.066	0.168	117
Treatment Design							
Within-subjects	0.107	0.130	0.365	0.101	0.109	0.293	246
Between-subjects	-0.029	-0.032	0.318	-0.027	-0.027	0.230	378
Incentive Realization Scheme							
All subjects paid	0.047	0.042	0.357	0.044	0.041	0.274	506
Random subjects paid	-0.072	-0.073	0.257	-0.066	-0.075	0.190	118
All decisions realized	0.090	0.126	0.399	0.082	0.080	0.286	181
Random decisions realized	-0.003	0.000	0.314	-0.001	0.005	0.250	443
Payoff Domain							
Gain	0.018	0.015	0.309	0.016	0.019	0.234	443
Loss	0.014	0.014	0.278	0.017	0.059	0.234	87
Mixed	0.080	0.108	0.480	0.076	0.094	0.366	111
Estimation Method							
Parametric	-0.058	-0.035	0.253	-0.056	-0.057	0.186	166
Non-parametric	0.054	0.053	0.366	0.052	0.048	0.282	458
Publication Status							
Unpublished	0.063	0.042	0.296	0.031	0.040	0.177	46
Published	0.021	0.018	0.347	0.023	0.021	0.270	578
Published in econ Top5	0.038	-0.023	0.383	0.033	-0.010	0.340	102
Published in economics	-0.012	-0.019	0.353	-0.008	-0.006	0.295	345
Published in other field	0.070	0.069	0.326	0.063	0.061	0.213	279

Table 2: Summary of effect sizes by subgroup characteristics. Raw effects correspond to the original Cohen's d, while PoBMA estimates refer to the posterior estimates of true effect sizes.

Table 2 reports meta-analytic summary statistics by subgroup (Online Appendix J.1 provides a more extensive disaggregation that leaves the main conclusions unchanged). Across all categories, the PoBMA estimates remain small. Mean effect

sizes range from -0.070 to 0.101, while median effect sizes range from -0.075 to 0.109, well within the conventional negligible-effect interval of $[-0.2, 0.2]$. Thus, the conclusion that incentive effects are negligible on average does not arise solely from pooling across heterogeneous domains, task types, or implementation procedures.[6] Although subgroup means occasionally differ in sign, all remain close to zero and none suggest substantively meaningful effects. We next examine whether the remaining differences across domains and study characteristics are statistically meaningful.

5.2 Meta-regression

The disaggregated means above give a first indication that differences across decisions types, domains, and tasks in incentive effects are likely to be small. To assess this issue more systematically—and to do so while controlling for several aspects of the tasks, domains, and subject pool while doing so, we next extend the Bayesian hierarchical measurement-error model by replacing the population mean μ in Eq. (2) with a meta-regression term. Let X denote an $N \times K$ matrix of study-level predictors and let $\boldsymbol{\alpha}$ be the corresponding K-dimensional vector of regression coefficients. The model becomes:

$$\widehat{d}_i \sim \text{Student-}t(\nu,\ X_i\boldsymbol{\alpha} + \gamma_x,\ \sigma)\,,$$

where X_i is the row of predictors for study i, γ_x is the experiment-level random effect, and σ captures the between-study standard deviation of the true effects after accounting for the covariates. The goal of this meta-regression is to explain true heterogeneity in $\widehat{d}_i$ after adjusting for the influence of sampling error.

Before turning to the substantive moderators, it is important to clarify the choice of model used to analyse heterogeneity. Although the PoBMA framework provides

[6] As a rough benchmark for economic magnitude, we examined the subset of 120 estimates originally reported as proportions or frequencies. Across these studies, incentivized subjects exhibited an average increase of 5.35 percentage points (median: 3.19 percentage points) in outcomes associated with greater risk aversion or impatience relative to hypothetical conditions. Even in these more directly interpretable units, the estimated incentive effects remain modest.

our preferred estimate of the *overall* effect size by averaging across publication-bias corrections and model structures, it is not suited for meta-regression. The publication-bias models included in PoBMA impose substantially different likelihoods and regression structures (e.g. PET and PEESE introduce precision terms, Vevea–Hedges applies p-value-based weighting, and—to complicate things further—fixed-effect variants do not allow for heterogeneity at all), making moderator effects non-comparable across models. By contrast, the Bayesian hierarchical measurement-error model (BHMEM) provides a unified and coherent platform for studying systematic variation in the latent true effect sizes $\widehat{d}_i$. It explicitly separates sampling error from between-study variability, accommodates experiment-level clustering, and allows covariates to be incorporated in a consistent manner. Importantly, the BHMEM is amongst the predictively most accurate models in the LOO comparison (Table 1), and amongst the models with equal predictive performance it is the most parsimonious. For these reasons, all inferences about heterogeneity across study characteristics are based on the BHMEM rather than on the model-averaged PoBMA estimates.

All regression results reported below come from a single meta-regression that includes the following predictors: a dummy for decisions in the time domain (relative to risk tasks); a within-subjects dummy (relative to a between-subjects design); dummies for loss and mixed outcome domains (relative to gains); a dummy indicating whether the effect is based on a parametric estimate rather than a non-parametric measure; dummies for field and online experiments (relative to laboratory studies); the probability with which a participant is selected for payment in between-subject randomization schemes (binarized to paying all subjects versus paying only some); and a dummy indicating whether all decisions in a study are incentivized. We also control for whether a study is published and whether it appears in an economics journal. The full regression tables, as well as several robustness analyses including additional controls (such as geographical location) and continuous version of binarized variables are reported in Online Appendix J.2.

5.3 Domain differences: risk, time, mixed, and losses

We begin by examining differences across decision domains—most prominently, between risk and time.[7] Panel A of Figure 5 plots kernel density estimates of the raw effect sizes for risk and time tasks. At a descriptive level, the two distributions are remarkably similar: both are centered close to zero and show a broadly symmetric shape around that point. There is no visually apparent shift suggesting that incentive effects systematically differ between the two domains.

While the nonparametric distributions provide no indication of domain-level differences, visual comparisons alone cannot account for sampling error, between-paper heterogeneity, or correlations with other study characteristics. We therefore turn to the meta-regression analysis, which formally tests whether the underlying, bias-adjusted effect sizes differ between risk and time preferences once these factors are taken into account. Panel C plots the posterior difference in true effects between time and risk domains, together with its 95% credible interval. The interval spans zero comfortably, indicating that—after adjusting for noise and study-level covariates—there is no meaningful difference in incentive effects between risk and intertemporal choice tasks (p-value = 0.198).

Panel B shows the distribution of raw effect sizes separately for gains, losses, and mixed-outcome choices. Effects for gains are perfectly centered on zero, indicating no detectable incentive effect in this domain. In contrast, mixed gain–loss choices exhibit a much broader distribution that does not reveal an immediate directional pattern. To clarify these patterns, Panel C reports the corresponding meta-regression estimates. The credible interval for losses includes zero, indicating that the effect is not statistically meaningful (p-value = 0.676). For mixed gain–loss choices, however, the estimated effect is significantly negative (p-value = 0.002): incentives push choices in the direction of greater *risk seeking* (i.e., reduced risk aversion).

These results raise important interpretational considerations. Together with the

[7] The dataset includes a single effect size from an ambiguity task (i.e. choices under unknown probabilities). We classify this as part of the "risk" domain for present purposes.

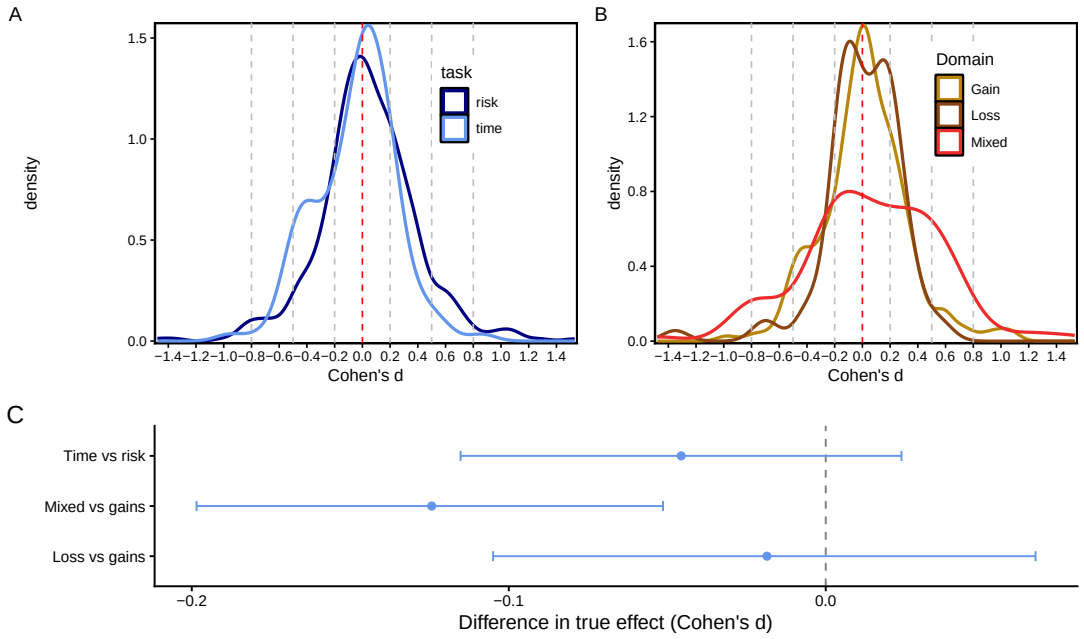

Figure 5: Cohen's d by decision domain

null effect for gains, the findings suggest that the observed domain differences may reflect features of the *incentivization mechanism itself*, rather than intrinsic differences in preferences under real versus hypothetical payment. Nearly all incentivized studies in our dataset implement losses by deducting them from an initial endowment. This creates the possibility of *house-money*–type integration, whereby subjects mentally combine the experimental endowment with the subsequent losses (Thaler and Johnson, 1990). The qualitative pattern we observe is exactly what such integration predicts. For mixed gambles, integration with an endowment can diminish the effective loss component; for loss-averse individuals, this should *increase* risk taking, precisely the direction found in our data. Direct evidence for such mechanisms has been documented experimentally by Jelschen and Schmidt (2023), who show how unconditional endowments can be mentally assimilated and thereby alter risk-taking behavior.

It is important to emphasize how these findings differ from those in prior meta-analyses examining loss aversion. Brown et al. (2024) report no systematic difference between incentivized and hypothetical conditions. The key distinction is

methodological: these earlier studies rely on *between-study* comparisons, where incentive conditions are not randomized and can be confounded with other design features (e.g., stake size, elicitation method, sample composition) that often covary with payment schemes. In contrast, the effects estimated in our meta-analysis are based on *within-study* random variation in incentives. They therefore admit a causal interpretation: the differences we observe reflect the consequences of incentivization itself, rather than uncontrolled differences across studies.

5.4 Design differences: treatment and incentives

Panel A of Figure 6 separates effect sizes according to whether the experiment used a between- or within-subjects manipulation. This distinction is theoretically important: within-subjects designs are more susceptible to contrast effects (a given change appears more pronounced when experienced side-by-side) and to experimenter-demand effects (subjects may infer what the experimenter "wants" from observing treatment variation). See, for example, Greenwald (1975) for an early and influential discussion.

This matters for interpreting the broader literature. The canonical evidence often cited as proof that incentives "matter" in individual decision tasks—notably Holt and Laury (2002)—relies on a within-subjects manipulation of payoff salience. As several commentators have pointed out, such designs confound incentive effects with contrast- and demand-induced shifts in choice patterns (e.g. Read, 2005). Our data reveal precisely this pattern: within-subject manipulations tend to yield noticeably larger effect sizes than between-subject designs. The top bar in Panel C confirms that this difference is statistically significant in our meta-regression (p-value $= 0.038$).

Panel B displays density estimates conditioned on whether *all* subjects were paid versus whether only a subset were paid. In most experiments employing partial payment, the selection probability is ≤ 0.2; in such cases a dummy indicator is more appropriate than a continuous measure. Our qualitative conclusions remain unchanged, however, when we use the continuous probability as a covariate in

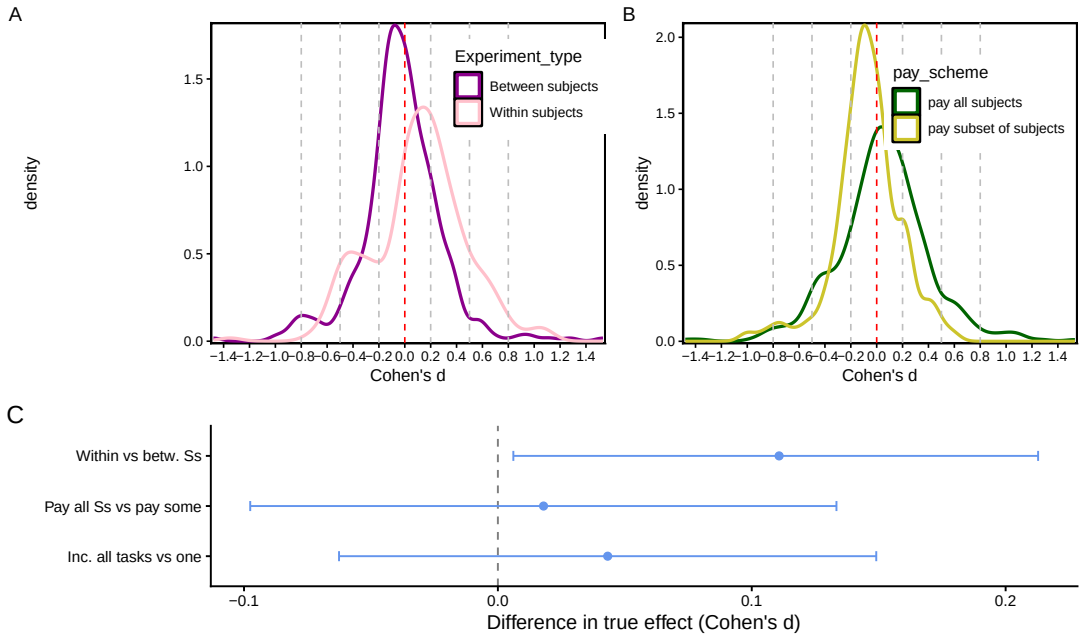

Figure 6: Cohen's d by treatment design

the meta-regression. Paying all subjects produces a distribution that is tightly centered and symmetric around 0. When only a subset of subjects is paid, the distribution appears slightly shifted to the left. This pattern is not supported by the meta-regression estimates in Panel C, which indicate no statistically meaningful effect of the payment probability on incentive effects (p-value = 0.753).

Panel C further includes a dummy for whether *all decisions* made by a subject were incentivized.[8] Here, too, the meta-regression detects no systematic impact on effect sizes (p-value = 0.422). In summary, although within-subject designs tend to amplify incentive effects, the specific payment scheme—paying all subjects, paying a subset, or paying all decisions—does not appear to meaningfully influence the magnitude of incentive effects. Taken together, these findings indicate that, however subjects are paid, incentive schemes have no detectable effect on individual choice behaviour—at least for the types of tasks represented in our dataset.

[8]This design feature is not simply a subset of the "pay-all-subjects" category: some studies pay all decisions even when only a subset of subjects is paid.

5.5 Other measures of heterogeneity

Most other controls we included—field or online experiments versus lab experiments, publication status, etc—yield no significant effect. The full regressions—as well as robustness regressions with additional controls, and different quantification of the incentive variables—can be found in Online Appendix J.2.

Taken together, these results show that although some study characteristics—most notably within–subjects designs and mixed gain–loss tasks—exhibit statistically detectable shifts in estimated incentive effects, these differences are substantively small. More importantly, meta-regression explains essentially none of the between-study heterogeneity in true effect sizes. Consequently, the residual heterogeneity appears to reflect idiosyncratic study-level variation rather than systematic differences in design, domain, or incentive implementation. In sum, once sampling error and selective reporting are accounted for, incentive provision produces no consistent or meaningful change in individual choice behaviour across the diverse experimental tasks included in our dataset.

6 Incentive effects on choice variability

In keeping with the bulk of the literature, our analysis thus far has focused on incentive effects on mean behaviour, such as risk-taking propensities and patience. An interesting question, however, concerns whether incentives may affect choice variability. To answer this question, we also coded all available measures that could serve as proxies for inconsistency or stochasticity in choice, such as first order stochastic dominance violations, transitivity violations and preference reversals, as well as measures of randomness, such as response variances and error terms estimated in structural models. This procedure yielded 230 observations from 29 out of the 73 papers (with the remaining papers not reporting measures of choice variability). We also encoded response times as a proxy for attention. We coded these measures so that positive values indicate reduced inconsistency

or stochasticity under real incentives, whereas negative values indicate increased inconsistency or stochasticity.

Raw effects. Figure 7 plots the density of both the absolute (Panel A) and signed (Panel B) values of the 230 effect sizes. Overall, 66.5% of the effect sizes fall into the negligible region. An additional 26.5% fall into Cohen's "small" category, while medium-sized and large effects are rare, accounting for 3.0% and 3.9%, respectively. For the signed effects, the mode (= 0), median (= 0.045), and mean (= 0.080) all fall into the negligible region.

Meta-analytic mean. We next estimate the mean aggregate effect using the BHMEM and following the same procedures as above. The estimated degrees of freedom of the Student–t distribution are $\nu = 2.374$ (95% CrI [2.008, 3.499]), again confirming substantial tail heaviness and thereby supporting the Student–t specification as an outlier-robust choice. The estimated meta-analytic mean is $\mu = 0.071$, with a 95% credible interval (CrI) of [0.019, 0.128]. Although the posterior mean is credibly positive, the credible interval falls entirely within the range of negligible effect sizes, suggesting that any aggregate effect of incentives on choice variability is too small to be of practical importance.

Panel A of Figure 8 compares the raw effect sizes, d_i, with the posterior distribution of true effect sizes, $\widehat{d}_i$. The posterior distribution is substantially narrower, with 88.7% of all $\widehat{d}_i$ falling within the negligible-effect interval $[-0.2, 0.2]$. Panel B of Figure 8 further plots the posterior true effect size, $\widehat{d}_i$, against its posterior standard deviation, sd_i, with points colour-coded by classification. Overall, only 3.0% of studies have posteriors falling above the cutoff of 0.2, while none show a negative effect. By contrast, 51.7% of studies are unambiguously negligible, thus providing *positive evidence of absence*: for these studies, a practically null effect is genuinely likely. Such cases are more than ten times as common as positive and negative effects combined. Finally, 45.2% of studies are too imprecise or too small to yield a clear conclusion, reflecting low power or unfavorable signal-to-noise ratios.

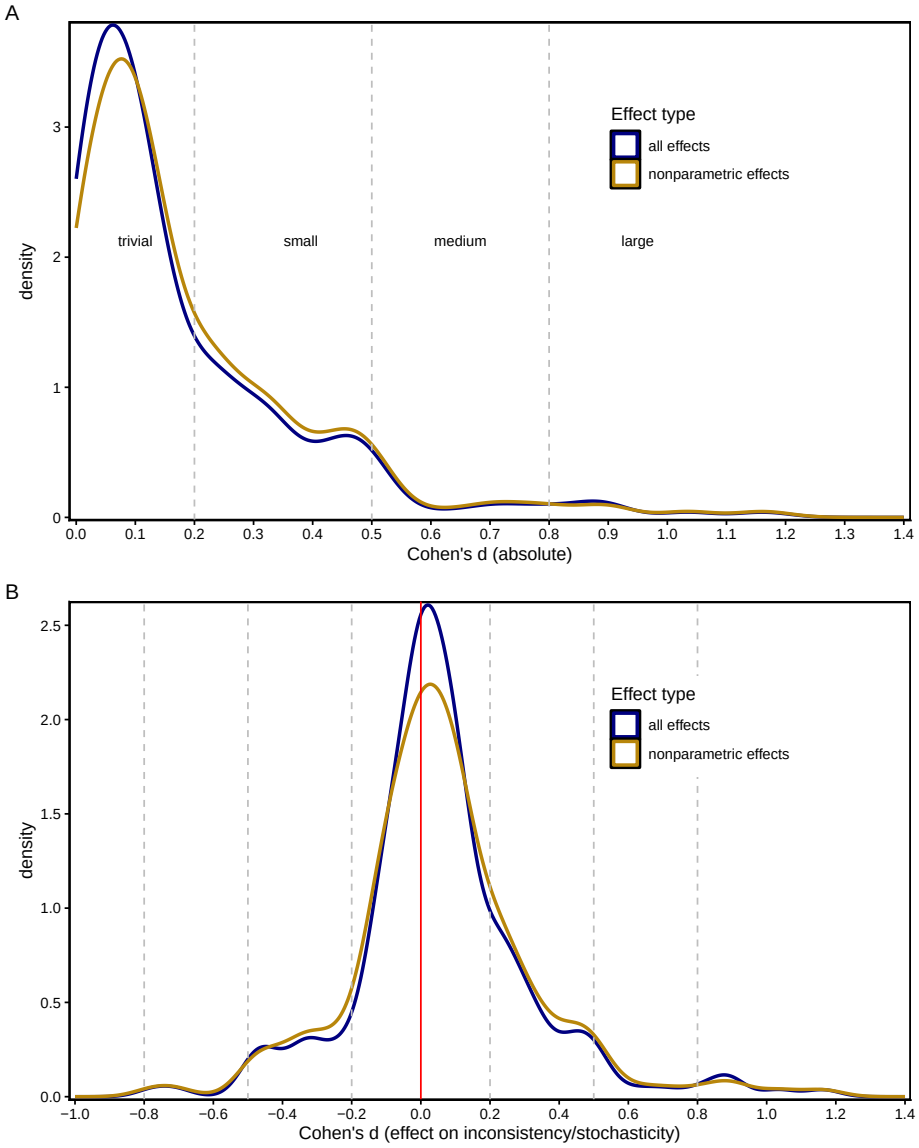

Figure 7: Probability density of Cohen's d

Distribution of 230 Cohen's d effect sizes across studies. Panel A shows the distribution of absolute effect sizes. Panel B shows the distribution preserving the sign of the effect, with positive values corresponding to reduced inconsistency or stochasticity under real incentives, and negative values corresponding to increased inconsistency or stochasticity.

In Online Appendix K, we furthermore report an analysis of publication bias, as well as results disaggregated by category. The publication-bias analysis does not alter the substantive conclusions: all models indicate negligible effects consistent with those discussed above. Disaggregation per category shows that these results are not an artefact of aggregation, but persist when examining specific sub-categories of decision domains and task types in isolation.

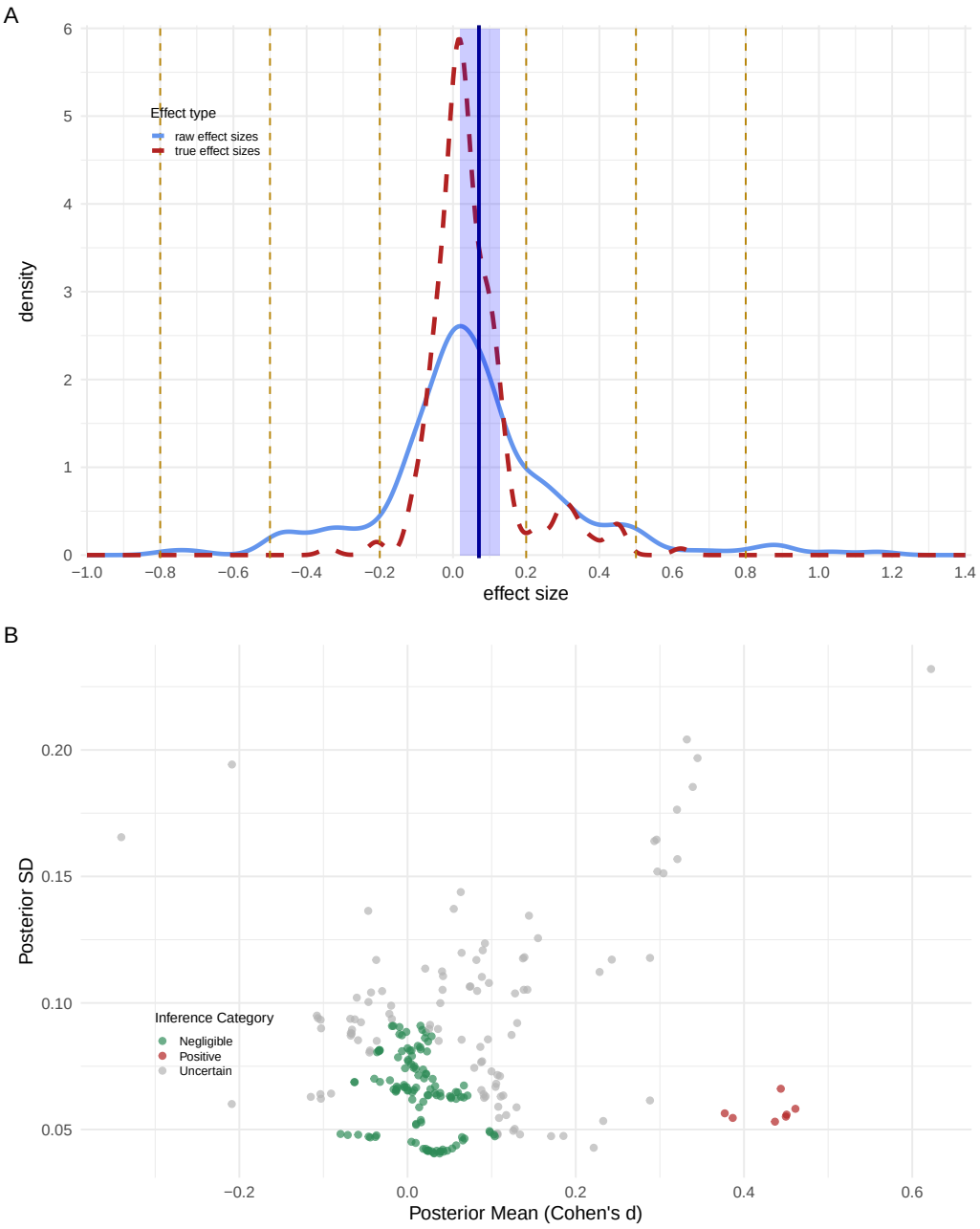

Figure 8: Posterior inferences on effect sizes

Posterior inferences from the BHMEM. Panel A compares the distribution of raw effect sizes d_i with the posterior distribution of true effect sizes $\widehat{d}_i$. Panel B plots the posterior means $\widehat{d}_i$ (x-axis) against their posterior standard deviations (y-axis), which correspond to standard errors in frequentist terminology.

7 Conclusion

This study offers a systematic, causally identified evaluation of whether real monetary incentives materially change behaviour in canonical individual decision-making tasks. Pooling 624 effect sizes from studies that randomly vary incentive

provision, and analyzing them with an outlier-robust Bayesian hierarchical framework, we show that the average effect of real incentives on decisions under risk and over time is negligible. This conclusion holds across all estimation strategies, including models that adjust for small-study patterns and publication bias. While incentive effects do vary across studies, the magnitude of this heterogeneity is limited and explained only weakly by observable design features.

Interpretation of our results. A natural question concerns why we observe virtually no incentive effects in the data. One traditional view holds that individuals possess direct access to their preferences, and that reporting these preferences requires little cognitive effort. If so, there is little reason to expect hypothetical choices to diverge systematically from incentivized ones: instructing subjects to "answer as if the choices were for real" may already suffice for stable preference revelation (e.g. Tversky and Kahneman, 1992). Several economic studies similarly find that incentive provision often fails to eliminate well-known "biases" (Grether and Plott, 1979; Enke et al., 2023).

Moreover, real incentives can sometimes introduce complications rather than resolve them. Complex payment schemes, loss implementation rules, or unfamiliar randomization procedures may impose additional cognitive load, increase task misunderstanding, or generate experimenter-demand concerns (e.g. Camerer and Hogarth, 1999; Hertwig and Ortmann, 2001). In such cases, incentives may *add noise* rather than improve preference elicitation. This possibility is consistent with our finding that the only clear departures from the overall null effect occur in mixed gain–loss decisions—domains where incentives are typically implemented via loss-from-endowment mechanisms known to generate house-money effects and other framing distortions. Thus, these deviations likely reflect artefacts of implementation rather than genuine incentive responsiveness.

Noisy cognition. A very different interpretation comes from recent research arguing that many patterns of choice under risk and over time may arise not from stable preferences, but from systematic *cognitive frictions*. In these models, behaviour

is shaped by noisy number perception, imprecise mental representations, or probabilistic computation rather than by the optimization of a well-defined utility function (Khaw, Li and Woodford, 2021). Related experimental evidence demonstrates that seemingly innocuous manipulations of numerical format—such as displaying outcomes in different numerical units—can produce systematic changes in observed behaviour (e.g. Garagnani and Vieider, 2025; Oprea and Vieider, 2026).

Within this noisy-cognition framework, we would expect incentives to influence choices only to the extent that they increase attention and thereby reduce processing noise. This prediction stands in sharp contrast to the "easy-access-to-preferences" account discussed above, under which hypothetical and incentivized choices should be similar precisely because preferences are readily retrieved. Our findings speak directly to this *incentive-based attention* mechanism: we see no systematic shift in mean decisions when incentives are introduced. Thus, while our results are entirely consistent with the idea that cognitive frictions shape behaviour, they suggest that standard monetary incentives do not, on their own, attenuate those frictions in the kinds of tasks studied here.

We caution against overgeneralizing this conclusion. Our analysis focuses on individual decisions under risk and over time, and does not speak directly to domains involving strategic interaction, real-world consequences, or costly actions outside the laboratory. Recent evidence suggests that incentive effects may also be limited in many performance tasks, but the importance of incentives likely remains context dependent. For the study of individual decision making under risk and over time, however, our results call for a reassessment of standard experimental practices and a reconsideration of when incentive provision is truly necessary, and when it may alter the very behaviours researchers seek to measure.

References

Andrews, Isaiah, and Maximilian Kasy (2019) 'Identification of and correction for publication bias.' *American Economic Review* 109(8), 2766–2794

Brañas-Garza, Pablo, Diego Jorrat, Antonio M Espín, and Angel Sánchez (2023) 'Paid and hypothetical time preferences are the same: Lab, field and online evidence.' *Experimental Economics* 26(2), 412–434

Brañas-Garza, Pablo, Lorenzo Estepa-Mohedano, Diego Jorrat, Victor Orozco, and Ericka Rascón-Ramírez (2021) 'To pay or not to pay: Measuring risk preferences in lab and field.' *Judgment and Decision Making* 16(5), 1290–1313

Brown, Alexander L., Taisuke Imai, Ferdinand M. Vieider, and Colin F. Camerer (2024) 'Meta-analysis of empirical estimates of loss aversion.' *Journal of Economic Literature* 62(3), 485–616

Cala, P, T Havranek, Z Irsova, M Luskova, J Matousek, and J Novak (2026) 'Financial incentives and performance: a meta-analysis of experiments in economics.' *Journal of Political Economy: Microeconomics*

Camerer, Colin F, and Robin M Hogarth (1999) 'The effects of financial incentives in experiments: A review and capital-labor-production framework.' *Journal of Risk and Uncertainty* 19, 7–42

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017) 'Stan: A probabilistic programming language.' *Journal of Statistical Software* 76(1), 1–32

Carson, Richard T, and Theodore Groves (2007) 'Incentive and informational properties of preference questions.' *Environmental and resource economics* 37, 181–210

Cheung, Stephen L., Agnieszka Tymula, and Xueting Wang (2023) 'Quasi-hyperbolic present bias: A meta-analysis.' *SSRN Electronic Journal*

Cohen, Jacob (1988) *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. (Hillsdale, NJ: Lawrence Erlbaum Associates)

Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo

Offerman, and Jeroen Van De Ven (2023) 'Cognitive biases: Mistakes or missing stakes?' *Review of Economics and Statistics* 105(4), 818–832

Garagnani, Michele, and Ferdinand M. Vieider (2025) 'Economic consequences of numerical adaptation.' *Psychological Science* 36(6), 407–420

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin (2014) *Bayesian data analysis,* vol. 3 (CRC press Boca Raton, FL)

Gneezy, Uri, Yoram Halevy, Brian Hall, Theo Offerman, and Jeroen van de Ven (2024) 'How real is hypothetical? a high-stakes test of the allais paradox.' Technical Report

Greenwald, Anthony G. (1975) 'Consequences of prejudice against the null hypothesis.' *Psychological Bulletin* 82(1), 1–20

Grether, David M, and Charles R Plott (1979) 'Economic theory of choice and the preference reversal phenomenon.' *The American Economic Review* 69(4), 623–638

Hertwig, Ralph, and Andreas Ortmann (2001) 'Experimental practices in economics: A methodological challenge for psychologists?' *Behavioral and brain sciences* 24(3), 383–403

Holt, Charles A., and Susan K. Laury (2002) 'Risk Aversion and Incentive Effects.' *American Economic Review* 92(5), 1644–1655

Imai, Taisuke, Tom A Rutter, and Colin F Camerer (2021) 'Meta-analysis of present-bias estimation using convex time budgets.' *The Economic Journal* 131(636), 1788–1814

Irsova, Zuzana, Pedro RD Bom, Tomas Havranek, and Heiko Rachinger (2025) 'Spurious precision in meta-analysis of observational research.' *Nature Communications* 16(1), 8454

Jelschen, Hauke, and Ulrich Schmidt (2023) 'Windfall gains and house money: The effects of endowment history and prior outcomes on risky decision–making.' *Journal of Risk and Uncertainty* 66(3), 215–232

Khaw, Mel Win, Ziang Li, and Michael Woodford (2021) 'Cognitive imprecision

and small-stakes risk aversion.' *The Review of Economic Studies* 88(4), 1979–2013

Maier, Maximilian, Dora Matzke, Jeffrey N. Rouder, Eric-Jan Wagenmakers, and Alexander Ly (2022) 'Robust bayesian meta-analysis: Addressing publication bias with model averaging.' *Psychological Methods* 27(5), 790–808

Matousek, Jindrich, Tomas Havranek, and Zuzana Irsova (2022) 'Individual discount rates: a meta-analysis of experimental evidence.' *Experimental Economics* 25(1), 318–358

Oprea, Ryan, and Ferdinand M. Vieider (2026) 'Information processing and the anomalies of risk and time.' *Working Paper*

Plott, Charles R (1986) 'Rational choice in experimental markets.' *Journal of Business* pp. S301–S327

Read, Daniel (2005) 'Monetary incentives, what are they good for?' *Journal of Economic Methodology* 12(June), 265–276

Smith, Vernon L (1982) 'Microeconomic systems as an experimental science.' *The American economic review* 72(5), 923–955

Stanley, Tom D (2008) 'Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection.' *Oxford Bulletin of Economics and statistics* 70(1), 103–127

Stanley, Tom D, and Hristos Doucouliagos (2014) 'Meta-regression approximations to reduce publication selection bias.' *Research Synthesis Methods* 5(1), 60–78

Thaler, Richard H, and Eric J Johnson (1990) 'Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice.' *Management science* 36(6), 643–660

Tversky, Amos, and Daniel Kahneman (1992) 'Advances in Prospect Theory: Cumulative Representation of Uncertainty.' *Journal of Risk and Uncertainty* 5, 297–323

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017) 'Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.' *Statistics and Computing* 27(5), 1413–1432

Vevea, Jack L, and Larry V Hedges (1995) 'A general linear model for estimating effect size in the presence of publication bias.' *Psychometrika* 60(3), 419–435

Vieider, Ferdinand M. (2024) 'Bayesian estimation of decision models.' Technical Report, RISL$\alpha\beta$

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018) 'Using stacking to average Bayesian predictive distributions.' *Bayesian Analysis* 13(3), 917–1007

ONLINE APPENDIX

A Search Strategy and Additional Details

A.1 Full Search Terms

The primary search was conducted in April 2026 using Web of Science (All Databases), across research areas including Psychology, Behavioral Sciences, and Business Economics. The following Boolean search string was used:

```
((hypothetical OR fictive) AND (real OR actual) AND (incentive
OR reward OR payoff))
```

This search yielded 602 records. An additional 103 papers were identified through backward citation searches, and 106 papers through Peter Wakker's annotated bibliography (search term: "real incentives/hypothetical choice"). We then shared the preliminary list of papers we had screened for inclusion on the most commonly used mailing lists (ESA and JDM Society) to solicit additional articles and unpublished results we might have missed, which yielded a further 29 papers.

A.2 Inclusion and Exclusion Details

Studies were included if:

1. They compared behavior under hypothetical and real incentives.
2. The incentive manipulation occurred within tasks involving risk or intertemporal choice.
3. The ranges of magnitudes, probabilities, or delays were held constant or directly comparable across incentive conditions.

Studies were excluded if real and hypothetical conditions differed in:

- reward magnitudes,
- probability ranges,
- time delays,
- commodity type,
- participant recruitment.

These exclusions helped prevent confounding influences related to factors such as magnitude, delay, commodity, and demographic characteristics. In total, 75 studies were excluded on these grounds, leaving a full dataset of 73 papers covering 90 distinct experiments and a total of 629 effect sizes. These observations can be used to infer the absolute incentive effect.

However, since the main interest of this paper focuses on the signed effects, five effect sizes with unclear direction were excluded. The resulting sample therefore

consists of 624 signed effect sizes drawn from 69 papers comprising 86 experiments, which constitute the primary dataset for the analysis in this paper.

B Outcome Coding Details

B.1 Temporal Discounting

We coded the following outcomes:

- Proportion of choices for smaller-sooner versus larger-later rewards.
- Indifference points and Area Under the Curve (AUC) measures.
- Estimated discounting parameters (exponential, hyperbolic, quasi-hyperbolic).

B.2 Risk Taking

We coded the following outcomes:

- Proportion of risky versus safe choices.
- Certainty equivalents and/or AUC of the utility function.
- Prospect Theory parameter estimates:
  - utility curvature,
  - probability weighting,
  - loss aversion.
- Balloon Analogue Risk Task (BART) measures.

B.3 Choice Variability

We coded the following proxies:

- Response inconsistency:
  - Independence violations.
  - Dominance violations.
  - Transitivity violations.
  - Monotonicity violations.
  - Preference reversals.
  - Nonsystematic response patterns.
- Response stochasticity:
  - Estimated error parameters in prospect-theory models.
  - Response or residual variances.
  - Extreme responses.

In addition, we also encoded 27 extra effects on response times as a measure of attention.

C Effect Size Computation: Full Formulas

For each study, we computed Cohen's d using the information reported (test statistics, summary moments, or regression output). Throughout, N_1 and N_2 denote the sample sizes in the real and hypothetical conditions, respectively. For within-subject designs, N denotes the number of paired observations (participants providing both responses), and ρ denotes the within-subject correlation when available. We report two variants of the effect size: d_0 (assuming $\rho = 0$ when correlation is unavailable) and, where applicable, $d_{0.5}$ (an alternative assuming $\rho = 0.5$).

C.1 Between-subject designs

Directly reported d: If Cohen's d was reported, we used it directly:

$$d_0 = d_{0.5} = d.$$

t statistic: For an independent-samples t test:

$$d_0 = d_{0.5} = |t| \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}.$$

F statistic: For a two-group comparison reported as an ANOVA F statistic:

$$d_0 = d_{0.5} = \sqrt{F} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}.$$

Mann–Whitney test using Z: When a standardized Z statistic was reported, we first computed

$$r = \frac{|Z|}{\sqrt{N_1 + N_2}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1 - r^2}}.$$

Mann–Whitney test using U: When the Mann–Whitney U statistic was reported, we converted U to Z via

$$Z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}, \qquad r = \frac{|Z|}{\sqrt{N_1 + N_2}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1 - r^2}}.$$

χ^2 statistic: When a χ^2 statistic was reported:

$$r = \frac{\sqrt{\chi^2}}{\sqrt{N_1 + N_2}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1 - r^2}}.$$

Means and standard deviations: When group means and SDs were available, we computed the pooled SD

$$s_p = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}},$$

and then

$$d_0 = d_{0.5} = \frac{|\bar{X}_1 - \bar{X}_2|}{s_p}.$$

If SDs were not reported, we recovered them from other summary statistics. When a standard error was reported, we used $s = SE\sqrt{N}$. When a 95% confidence interval for a mean was reported as $[CI^l, CI^h]$, we computed

$$\bar{X} = \frac{CI^l + CI^h}{2}, \qquad SE = \frac{CI^h - CI^l}{2z}, \quad z = 1.96.$$

Regression-based effect sizes: When effect sizes were derived from regression output, we first translated the regression coefficients into an implied contrast between the real and hypothetical conditions, denoted by Δ (e.g., a difference between two condition-specific estimates, or a linear combination of coefficients when interactions were present). We then standardized this contrast by an appropriate scale parameter S constructed from reported standard deviations (or closely related quantities) and, when necessary, a two-sample scaling factor.

Specifically, for specifications that directly yielded condition-specific levels (e.g., separate intercepts or mean-equivalent coefficients), we treated the two coefficients as $\hat{\mu}_1$ and $\hat{\mu}_2$ and computed

$$d = \frac{|\Delta|}{S}, \qquad \Delta = \hat{\mu}_1 - \hat{\mu}_2,$$

where S was the pooled SD (or a pooled SD analogue) built from the reported within-condition SDs.

For specifications that reported the contrast as a single regression coefficient, we set Δ equal to that coefficient and computed a standardized mean-difference equivalent using an externally provided SD estimate and the standard two-sample scaling:

$$d = \frac{|\Delta|}{S}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}.$$

When interaction terms were present and the extraction provided multiple dispersion components per condition, we constructed two variants of the scale parameter S: one that combines dispersion components assuming zero covariance (yielding d_0), and an alternative that imposes a nonzero covariance structure consistent with $\rho = 0.5$ (yielding $d_{0.5}$). In all regression-based cases, d was defined as the absolute standardized real–hypothetical contrast, and we set $d_0 = d_{0.5}$ whenever the dispersion construction did not depend on the covariance assumption.

C.2 Within-subject designs

Let N denote the number of paired observations, and we define N_Σ as the sum of the reported group sizes when needed for rank-based conversions.

Directly reported d: If Cohen's d was reported, we used it directly:

$$d_0 = d_{0.5} = d.$$

t statistic: If the within-subject correlation ρ was not available, we computed

$$d_0 = |t|\sqrt{\frac{2}{N}}, \qquad d_{0.5} = |t|\sqrt{\frac{1}{N}}.$$

If ρ was available, we used

$$d_0 = d_{0.5} = |t|\sqrt{\frac{2(1-\rho)}{N}}.$$

F statistic: Analogously, for a within-subject F statistic:

$$d_0 = \sqrt{F}\sqrt{\frac{2}{N}}, \qquad d_{0.5} = \sqrt{F}\sqrt{\frac{1}{N}} \quad (\rho \text{ unavailable}),$$

and when ρ was available,

$$d_0 = d_{0.5} = \sqrt{F}\sqrt{\frac{2(1-\rho)}{N}}.$$

Wilcoxon signed-rank using Z: When a standardized Z statistic was reported:

$$r = \frac{|Z|}{\sqrt{N_\Sigma}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1-r^2}}.$$

Wilcoxon signed-rank using V: When the Wilcoxon signed-rank statistic V was reported, we converted it to Z via

$$Z = \frac{V - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}, \qquad r = \frac{|Z|}{\sqrt{N_\Sigma}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1-r^2}}.$$

χ^2 statistic: When a χ^2 statistic was reported:

$$r = \frac{\sqrt{\chi^2}}{\sqrt{N_\Sigma}}, \qquad d_0 = d_{0.5} = \frac{2r}{\sqrt{1-r^2}}.$$

Means and standard deviations: Let $\Delta = |\bar{X}_R - \bar{X}_H|$. If ρ was not available, we used

$$S_{\text{within},0} = \frac{\sqrt{s_R^2 + s_H^2}}{\sqrt{2}}, \qquad S_{\text{within},0.5} = \sqrt{s_R^2 + s_H^2 - s_R s_H},$$

and computed

$$d_0 = \frac{\Delta}{S_{\text{within},0}}, \qquad d_{0.5} = \frac{\Delta}{S_{\text{within},0.5}}.$$

If ρ was available, we first computed the SD of the difference

$$S_\Delta = \sqrt{s_R^2 + s_H^2 - 2\rho s_R s_H},$$

and then set

$$S_{\text{within}} = \frac{S_\Delta}{\sqrt{2(1-\rho)}}, \qquad d_0 = d_{0.5} = \frac{\Delta}{S_{\text{within}}}.$$

D Standard Error of Cohen's d

D.1 Between-subject designs

For between-subject designs, the standard error of d was computed as

$$se(d) = \sqrt{\frac{N_1 + N_2}{N_1 N_2} + \frac{d^2}{2(N_1 + N_2)}}.$$

D.2 Within-subject designs

When ρ was not available, we used

$$se(d_0) = \sqrt{\frac{2}{N} + \frac{d_0^2}{N}}, \qquad se(d_{0.5}) = \sqrt{\frac{1}{N} + \frac{d_{0.5}^2}{2N}}.$$

When ρ was available, we applied the correlation adjustment

$$se(d_0) = se(d_{0.5}) = \sqrt{\left(\frac{2}{N} + \frac{d^2}{N}\right)(1-\rho)}.$$

E Cohen's h versus Cohen's d

We conducted an additional robustness check for the subset of effects based on proportion or frequency outcomes. Among the 624 encoded effects, 120 report outcomes originally measured as proportions or frequencies. For this subset, we additionally computed effect sizes using:

$$\text{Cohen's } h = \left|2\arcsin\left(\sqrt{p_T}\right) - 2\arcsin\left(\sqrt{p_C}\right)\right|.$$

Figure 9 plots the density of the absolute effect sizes (Panel A) and signed effect sizes (Panel B) for both Cohen's d and Cohen's h for these 120 observations. The distribution of Cohen's h closely resembles that of Cohen's d.

- **Absolute values:** The mean is 0.295 for Cohen's h (0.292 for Cohen's d), and the median is 0.258 for Cohen's h (0.243 for Cohen's d). The difference is not statistically significant according to a Wilcoxon signed-rank test (p-value = 0.634).
- **Signed values:** The mean is 0.113 for Cohen's h (0.113 for Cohen's d), and the median is 0.079 for Cohen's h (0.079 for Cohen's d). Again, the difference is not statistically significant (Wilcoxon signed-rank test, p-value = 0.275).

These results indicate that using Cohen's d instead of Cohen's h for proportion outcomes does not mechanically attenuate effect sizes toward zero in our data.

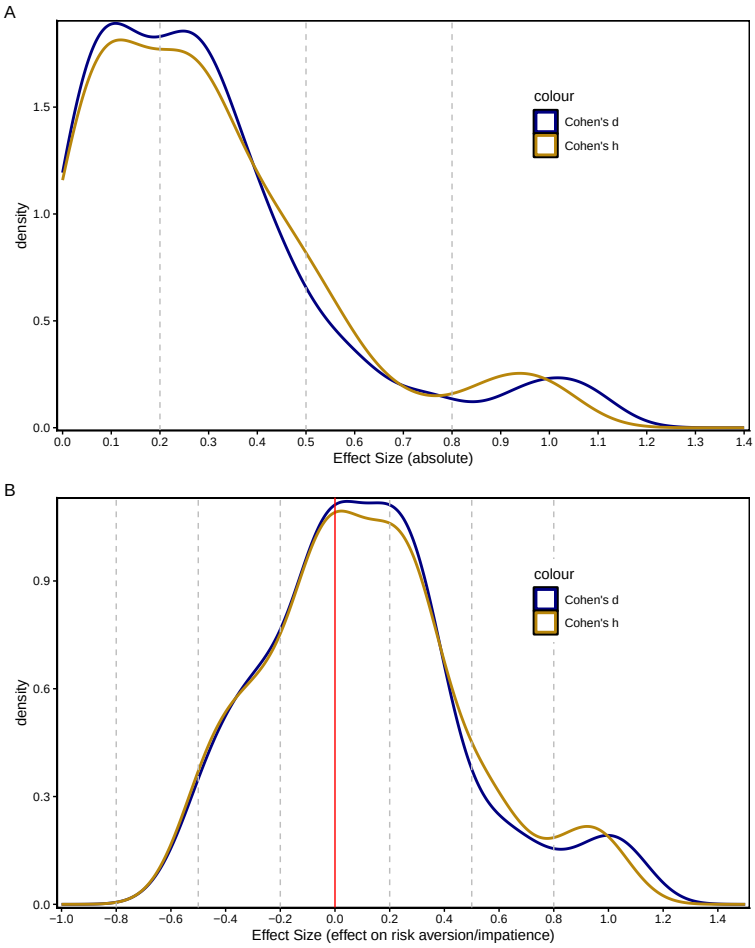

Figure 9: Probability density of Cohen's d versus Cohen's h

Distribution of 120 effect sizes across studies computed using Cohen's d and Cohen's h, respectively. Panel A shows the distribution of absolute effect sizes. Panel B shows the distribution preserving the sign of the effect, with positive values corresponding to increased risk aversion or impatience under real incentives, and negative values corresponding to increased risk seeking or patience.

F Coded Study Characteristics

Table 3 lists all variables extracted from each study.

Table 3: Full List of Coded Study Characteristics

Variable	Description
Source of Data	
`source_lab_exp`	=1 if laboratory experiment
`source_class_exp`	=1 if classroom experiment
`source_field_exp`	=1 if field experiment
`source_online_exp`	=1 if online experiment
`source_other_exp`	Other type of experiment
Treatment Design	

Continued from previous page

Variable	Description
`within_subjects`	= 1 if the design is within-subjects
`between_subjects`	= 1 if the design is between-subjects
Location of Experiment	
`loc_country`	Country
`loc_continent`	Continent
Subject Pool	
`subject_uni`	=1 if university students/staff
`subject_general`	=1 if general population
`subject_other`	Other specific group
Elicitation Task	
`choice_list`	=1 if choice list
`choice_binary`	=1 if binary choice
`choice_bid`	=1 if bidding
`choice_BART`	=1 if balloon analogue risk task
`choice_iterated`	=1 if iterative adjustment procedure
`choice_other`	Other type of task
Chances of Realization	
`prob_subject`	Probability that a subject is selected for payment
`prob_decision`	Probability that a decision is realized for payment
`prob_overall`	Overall probability of a real payment
Payoff Domain	
`domain_gain`	=1 if all outcomes positive
`domain_loss`	=1 if all outcomes negative
`domain_mixed`	=1 if positive and negative outcomes mixed within trials
`domain_gl`	=1 if positive and negative outcomes appear across trials
`endowment`	Endowment provided to cover losses
Reward Type	
`reward_money`	=1 if monetary rewards
`reward_health`	=1 if health-related goods
`reward_other`	Other type of outcomes
Stimuli Magnitude	
`reward_low`	Smallest reward amount
`reward_high`	Largest reward amount
`prob_low`	Lowest probability
`prob_high`	Highest probability
`delay_low`	Shortest delay
`delay_high`	Longest delay
Publication Status	
`published`	=1 if published in a peer-reviewed journal
`published_econ`	=1 if published in an economics journal
`published_econ_top5`	=1 if published in a Top-5 economics journal
`published_other_field`	Published in another field journal

G Details for Hierarchical Bayesian Model

G.1 Hyperpriors on parameters

The hyperpriors for the parameters in BHMEM are specified as

$$\mu \sim \mathcal{N}(0, 5),$$

$$\nu \sim \text{Exponential}(0.5),$$

$$\sigma \sim \text{Exponential}(1),$$

$$\tau_x \sim \text{Exponential}(1).$$

G.2 Stan code

```
data{
    int<lower=1> N;
    vector[N] cd;
    vector[N] se;
    int<lower=1> K;
    matrix[N,K] x;
    int<lower=1> P;
    array[N] int pid;
}
parameters{
    vector[N] eps;
    vector[K] beta;
    vector[P] mup;
    real<lower=2> df;
    real<lower=0> sigma;
    real<lower=0> tau;
}
transformed parameters {
    vector[N] dhat = x * beta + mup[pid] + eps;
}
model{
    sigma ~ exponential( 1 );
    tau ~ exponential( 1 );
    beta ~ normal( 0 , 5 );
    df ~ exponential( 0.5 );
    // residuals distribution:
    eps ~ student_t(df, 0, sigma);
    // distribution of paper-level residuals
    mup ~ normal( 0 , tau );
    // measurement error model
    cd ~ normal(dhat, se);
}
generated quantities {
  vector[N] log_lik;
  for (i in 1:N)
    log_lik[i] = normal_lpdf(cd[i] | dhat[i], se[i]);
}
```

H Publications Models

H.1 Technical details

H.1.1 PET–PEESE model

The Precision-Effect Test (PET) and the Precision-Effect Estimate with Standard Error (PEESE) are regression-based tools designed to detect and correct for publication bias by exploiting the empirical relationship between reported effect sizes and their standard errors (Stanley, 2008; Stanley and Doucouliagos, 2014). In their classical form, both methods are implemented as *fixed-effect* meta-regressions: if publication bias is present, smaller and less precise studies tend to report larger effects, generating a systematic association between estimated effects and their standard errors.

For completeness and comparability with the existing literature, we estimate the standard fixed-effect PET and PEESE regressions. In addition, we implement Bayesian hierarchical versions of both models by embedding the PET–PEESE structure inside our baseline random-effects BHMEM. These hierarchical extensions provide several advantages: they allow publication bias to be assessed while simultaneously accounting for (i) genuine between-study heterogeneity in true effects, (ii) statistical dependence among effect sizes reported in the same paper, and (iii) non-normality of the distribution of true effects, since the underlying BHMEM uses a Student-t specification. Together, these features make the hierarchical PET–PEESE models considerably more flexible and better suited to our data than their classical fixed-effect counterparts.

In the hierarchical formulation, the mean effect is allowed to depend on study precision as follows:

$$\widehat{d}_i \sim \text{Student-}t(\nu,\ \mu + \gamma_x + \lambda\, se_i,\ \sigma)\,,$$

where μ is the bias-adjusted mean effect, γ_x is an experiment-level random effect, and λ captures the dependence of effect sizes on study precision (PET). Under publication bias, λ is expected to differ from zero.[9]

The corresponding PEESE specification replaces the standard error with its square:

$$\widehat{d}_i \sim \text{Student-}t\big(\nu,\ \mu + \gamma_x + \lambda\, se_i^2,\ \sigma\big)\,.$$

[9] In the classical PET–PEESE formulation (Stanley, 2008; Stanley and Doucouliagos, 2014), the PET regression is applied to the standardized effect size d_i/se_i and regresses it on $1/se_i$. This standardization is required under the original *fixed-effect* assumptions, which treat sampling error as the sole source of variation across studies. In our Bayesian hierarchical specification, sampling variance is already modeled explicitly through the likelihood $d_i \sim \mathcal{N}(\widehat{d}_i, se_i^2)$, and between-study heterogeneity is captured by the random-effects distribution. Consequently, the PET–PEESE regression is formulated directly at the level of the latent true effect sizes $\widehat{d}_i$. This avoids double-counting sampling variance and allows PET–PEESE to operate consistently within a random-effects framework.

PEESE typically provides a less biased estimate of μ when a genuine nonzero effect exists, whereas PET is more reliable when the true effect is close to zero.

Across both PET and PEESE, the intercept μ corresponds to the predicted effect size for an infinitely precise study ($se_i \to 0$) and thus serves as the publication bias-adjusted estimate of the underlying population effect.

Note: PET-PEESE has the advantages of being simple, intuitive, and easy to implement, while also providing a framework for diagnosing and adjusting for small-study effects. However, its limitations should be acknowledged. The approach relies on specific linear model assumptions: PET assumes a linear relationship between effect sizes and standard errors, whereas PEESE assumes a linear relationship between effect sizes and squared standard errors. These assumptions may not hold in empirical data, and model misspecification can introduce additional bias. Moreover, PET-PEESE is not suitable for all forms of publication bias, as it primarily assumes that bias manifests through small-study effects. More complex selection mechanisms, such as preferential publication of statistically significant results, directional reporting, or p-hacking, may not be adequately corrected by this method.

H.1.2 Vevea & Hedges selection model

Other than PET–PEESE, the Vevea and Hedges (1995) model explicitly models the probability of a study being selected for publication. The approach combines two components: (i) an effect-size model, analogous to our Bayesian Hierarchical Measurement Error Model (BHMEM), that characterizes the distribution of study outcomes in the absence of selective publication, and (ii) a selection model that assigns relative probabilities to studies based on the p-value associated with their effect estimate. This formulation yields effect-size estimates that adjust for selective reporting and allows formal inference on the presence of publication bias.[10]

Let p_i denote the one-tailed p-value of study i, and let $w(p_i)$ denote the probability that a study with p-value p_i is observed. The weight function is typically specified as piecewise constant across K ordered intervals of p-values. Let the endpoints of the jth interval be a_{j-1} and a_j, with $a_0 = 0$ and $a_K = 1$. If p_i falls in the jth interval, the associated selection weight is

$$w(p_i) = \omega_j, \qquad \text{if } p_i \in (a_{j-1}, a_j].$$

Because only the relative publication probabilities are identified, the model requires a normalization. Following standard practice, the interval containing the *most statistically significant* results (i.e., $(0, a_1]$) is normalized to

$$\omega_1 = 1.$$

[10]The original Vevea & Hedges model is formulated under a fixed-effect meta-analytic framework. Our implementation includes both the classical fixed-effect version and a Bayesian hierarchical extension embedded within the BHMEM, allowing the model to accommodate the substantial between-study heterogeneity present in our dataset.

All other weights ω_j are therefore interpreted *relative* to this baseline. For example, $\omega_3 = 0.4$ implies that studies with p-values in interval 3 are published with 40% of the probability of studies in the most significant interval, whereas values $\omega_j > 1$ indicate intervals with a higher publication probability than the baseline.

The selection model can alternatively be expressed in terms of the corresponding test statistic $z_i = d_i/se_i$. Defining $b_j = -\Phi^{-1}(a_j)$, selection weights can be written as

$$w(z_i) = \begin{cases} \omega_1, & \text{if } b_1 < z_i \leq \infty, \\ \omega_j, & \text{if } b_j < z_i \leq b_{j-1}, \\ \omega_K, & \text{if } -\infty < z_i \leq b_{K-1}, \end{cases}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse standard normal cumulative distribution function. Given these weights, the likelihood contribution of an effect size d_i is

$$f(d_i \mid \cdot) = \frac{w(z_i)\,\phi(d_i \mid \widehat{d}_i, se_i^2)}{\sum_{j=1}^{K} \omega_j B_{ij}(\widehat{z}_i)},$$

where $\phi(\cdot)$ is the normal density, $\widehat{d}_i$ is the latent true effect under the effect-size model, $\widehat{z}_i = \widehat{d}_i/se_i$, and $B_{ij}(\widehat{z}_i)$ is the probability that a normal random variable with mean $\widehat{z}_i$ and unit variance falls in the jth selection interval:

$$B_{ij} = \begin{cases} 1 - \Phi(b_1 - \widehat{z}_i), & j = 1, \\ \Phi(b_{j-1} - \widehat{z}_i) - \Phi(b_j - \widehat{z}_i), & 1 < j < K, \\ \Phi(b_{K-1} - \widehat{z}_i), & j = K. \end{cases}$$

We estimate both *unidirectional* (V&H-UD) and *bidirectional* (V&H-BD) versions of the model. In the unidirectional specification, p-values are based on $|z_i|$, imposing symmetric selection weights for positive and negative effects. The bidirectional specification computes p-values from the signed statistic, allowing asymmetric selection depending on the effect's direction.

Following common practice, we partition the p-value distribution using thresholds at 0.025, 0.05, and 0.10, corresponding to conventional significance levels in empirical research. These intervals determine the selection weights ω_j, which quantify how much more (or less) likely studies in each significance band are to appear in the published sample.

Note: The Vevea & Hedges selection model offers the advantage of going beyond the detection of small-study effects by directly modeling the publication selection mechanism. Specifically, it assumes that the probability of a study entering the literature depends on its p-value interval, allowing studies with different levels of statistical significance to have different publication probabilities. Through a weight function, the model explicitly represents the relative selection probabilities across p-value intervals, making its structure transparent and interpretable. However, the method relies on strong assumptions about the selection function,

particularly the definition of p-value cutoffs. The resulting estimates can be sensitive to researcher-specified intervals which are partly subjective. The model is also highly dependent on the quality of p-value information derived from effect sizes and standard errors; incomplete reporting, unclear effect directions, inaccurate standard errors, rounding, selective reporting, or imprecise transformations can affect the estimates. In addition, because the selection function is discrete, studies near significance thresholds, such as $p = .049$ or $p = .051$, may exert disproportionate influence on the estimated selection weights.

H.1.3 Andrews & Kasy selection model

The Andrews and Kasy (2019) approach provides a general framework for identifying and correcting publication bias by explicitly modelling both the distribution of true effects and the selection mechanism governing which results are observed in the published sample.[11] Unlike the Vevea & Hedges model, which specifies relative selection weights across discrete p-value intervals, the A&K model directly parameterizes the *absolute* probability that a study with a given test statistic is published.

Let $z_i = d_i/se_i$ denote the test statistic for study i. The central object in the Andrews–Kasy framework is a selection function $p(z_i)$ describing the probability that a study with test statistic z_i is published (or written up). The observed distribution of effect sizes is therefore a reweighted version of the latent (unpublished) distribution:

$$f(d_i \mid \cdot) \;\propto\; p(z_i)\,\phi\Big(d_i \mid \widehat{d}_i,\; se_i^2\Big)\,,$$

where $\phi(\cdot)$ denotes the normal density and $\widehat{d}_i$ is the latent true effect implied by the effect-size model. Because $p(z_i)$ is modelled on the logit scale, the estimated publication probabilities are constrained to lie in the unit interval $(0, 1)$. Unlike the Vevea–Hedges framework, which identifies only *relative* odds of publication across p-value intervals, the Andrews–Kasy model targets absolute publication probabilities.

Identification of the selection function relies on the fact that studies in a meta-analysis typically differ in their sampling variances se_i^2. This variation means that two studies with similar underlying effects can nonetheless produce different test statistics $z_i = d_i/se_i$, purely because their standard errors differ. The resulting heteroskedasticity in z_i provides the key source of identifying variation that allows the publication probabilities $p(z_i)$ to be recovered; see Andrews and Kasy (2019) for details.

To model the selection function flexibly, Andrews and Kasy (2019) propose para-

[11] The original Andrews & Kasy framework is derived under assumptions that parallel a fixed-effect meta-analysis, with heterogeneity incorporated through a nonparametric distribution of true effects rather than an explicit random-effects structure. In our implementation, we estimate both the classical version and a Bayesian hierarchical extension embedded in the BHMEM to account for between-study heterogeneity.

metric and semiparametric basis expansions. We consider two widely used specifications:

- **Quadratic interpolation (A&K–QI).** A parsimonious specification in which
$$p(z_i) = \text{logit}^{-1}\big(\omega_0 + \omega_1 z_i + \omega_2 z_i^2\big),$$
allowing smooth nonlinear variation in publication probability across the range of z-values.

- **Natural spline interpolation (A&K–NS).** A more flexible specification in which
$$p(z_i) = \text{logit}^{-1}\big(\boldsymbol{\omega}^\top \boldsymbol{b}(z_i)\big),$$
where $\boldsymbol{b}(z)$ is a natural spline basis with knots placed at the conventional significance thresholds (-1.960, -1.282, 1.282, 1.960). This formulation permits highly flexible, data-driven modelling of selection patterns without imposing strong shape restrictions.[12]

Taken together, these two specifications span a wide range of plausible selection mechanisms, from smooth global patterns (QI) to flexible local nonlinearities (NS). In our implementation, both selection models are embedded within the Bayesian hierarchical measurement-error framework described above, allowing publication bias to be assessed while simultaneously accounting for between-paper heterogeneity, statistical dependence, and measurement error in a unified structure.

Note: The Andrews & Kasy selection model aims to identify continuous, absolute publication probabilities by modeling publication likelihood as a function of study results, rather than relying on a limited set of fixed p-value cutoffs with discrete relative weights. However, the approach still relies on strong structural assumptions about the specific form of the selection function. In addition, its results may be less accessible to audiences, as the continuous selection function is generally less intuitive than discrete p-value-based weighting schemes.

H.2 Stan code

H.2.1 PET–PEESE model

```
data{
    int<lower=1> N;  // number of observations
    vector[N] cd;  // effect sizes
    vector[N] se;  // standard errors
    int<lower=1> P;  // number of studies
    array[N] int pid; // index of studies
    int<lower=1> K;  // number of covariates
    matrix[N,K] x;  // covariates matrix
}

```

[12] Using a natural cubic spline basis with five degrees of freedom produces nearly identical conclusions; we adopt the significance-knot specification because of its interpretability.

```
parameters{
    vector[K] beta;
    vector[P] mup;
    vector[N] eps;
    real<lower=0> tau;
    real<lower=0> sigma;
    real<lower=2> df;
}

transformed parameters {
    vector[N] dhat = x * beta + mup[pid] + eps;
}

model{
    beta ~ normal( 0 , 5 );
    tau ~ exponential( 1 );
    sigma ~ exponential( 1 );
    df ~ exponential( 0.5 );

    // residuals distribution:
    eps ~ student_t(df, 0, sigma);

    // distribution of paper-level residuals
    mup ~ normal( 0 , tau );

    // measurement error model
    target += normal_lpdf(cd | dhat, se);
}

generated quantities {
    vector[N] log_lik;

    for (i in 1:N)
      log_lik[i] = normal_lpdf(cd[i] | dhat[i], se[i]);
}
```

H.2.2 Vevea & Hedges selection model

Unidirectional:

```
data{
    int<lower=1> N;  // number of observations
    vector[N] cd;  // effect sizes
    vector[N] se;  // standard errors
    int<lower=1> P;  // number of studies
    array[N] int pid; // index of studies
    int<lower=1> K;  // number of covariates
    matrix[N,K] x;  // covariates matrix
}

parameters {
    vector[K] beta;
    vector[P] mup;
    vector[N] eps;
```

```
    real<lower=0> tau;
    real<lower=0> sigma;
    real<lower=2> df;
    vector<lower=0>[3] alpha;  // selection weights for bins 2-4
}

transformed parameters {
    vector[N] dhat = x * beta + mup[pid] + eps;
    vector[4] w;
    w[1] = 1.0;  // fixed for p < 0.05
    for (j in 1:3)
      w[j+1] = alpha[j];
}

model {
    vector[N] log_lik_sel;

    beta ~ normal( 0 , 5 );
    tau ~ exponential( 1 );
    sigma ~ exponential( 1 );
    df ~ exponential( 0.5 );
    alpha ~ exponential(1);

    // residuals distribution:
    eps ~ student_t(df, 0, sigma);

    // distribution of paper-level residuals
    mup ~ normal( 0 , tau );

    // likelihood (with selection correction)
    for (i in 1:N) {
      real z = abs(cd[i] / se[i]);
      real log_f = normal_lpdf(cd[i] | dhat[i], se[i]);
      real mu_z = abs(dhat[i] / se[i]);

      vector[4] p;
      p[1] = 2 * (1 - normal_cdf(1.960 | mu_z, 1));  // p < 0.05
      p[2] = 2 * (normal_cdf(1.960 | mu_z, 1) - normal_cdf(1.645 |
          mu_z, 1));  // 0.05 < p < 0.10
      p[3] = 2 * (normal_cdf(1.645 | mu_z, 1) - normal_cdf(1.282 |
          mu_z, 1));  // 0.10 < p < 0.20
      p[4] = 2 * (normal_cdf(1.282 | mu_z, 1) - normal_cdf(0 |
         mu_z, 1));  // p > 0.20

      real norm_const = dot_product(w, p);

      int bin;
      if (z >= 1.960) {
        bin = 1;
      } else if (z >= 1.645) {
        bin = 2;
      } else if (z >= 1.282) {
        bin = 3;
      } else {
        bin = 4;
```

```
      }
      log_lik_sel[i] = log_f + log(w[bin]) - log(norm_const);
    }
    target += sum(log_lik_sel);
}

generated quantities {
    vector[N] log_lik;
    vector[4] sel_prob;
    vector[4] bin_width = [0.05, 0.05, 0.1, 0.8]';

    sel_prob[1] = 1.0;
    for (j in 2:4)
      sel_prob[j] = (w[j] / bin_width[j]) / (w[1] / bin_width[1]);

    for (i in 1:N)
      log_lik[i] = normal_lpdf(cd[i] | dhat[i], se[i]);
}
```

Bidirectional:

```
data{
    int<lower=1> N;  // number of observations
    vector[N] cd;  // effect sizes
    vector[N] se;  // standard errors
    int<lower=1> P;  // number of studies
    array[N] int pid; // index of studies
    int<lower=1> K;  // number of covariates
    matrix[N,K] x;  // covariates matrix
}

parameters {
    vector[K] beta;
    vector[P] mup;
    vector[N] eps;
    real<lower=0> tau;
    real<lower=0> sigma;
    real<lower=2> df;
    vector<lower=0>[7] alpha;  // selection weights for bins 2-8
}

transformed parameters {
    vector[N] dhat = x * beta + mup[pid] + eps;
    vector[8] w;
    w[1] = 1.0;  // fixed for positive p < 0.05
    for (j in 1:7)
      w[j+1] = alpha[j];
}

model {
    vector[N] log_lik_sel;

    beta ~ normal( 0 , 5 );
    tau ~ exponential( 1 );
    sigma ~ exponential( 1 );
```

```
    df ~ exponential( 0.5 );
    alpha ~ exponential(1);

    // residuals distribution:
    eps ~ student_t(df, 0, sigma);

    // distribution of paper-level residuals
    mup ~ normal( 0 , tau );

    // likelihood (with selection correction)
    for (i in 1:N) {
      real z = cd[i] / se[i];
      real log_f = normal_lpdf(cd[i] | dhat[i], se[i]);
      real mu_z = dhat[i] / se[i];

      vector[8] p;
      p[1] = 1 - normal_cdf(1.960 | mu_z, 1);  // p < 0.05, pos.
      p[2] = normal_cdf(1.960 | mu_z, 1) - normal_cdf(1.645 | mu_z
         , 1);  // 0.05 < p < 0.10, pos.
      p[3] = normal_cdf(1.645 | mu_z, 1) - normal_cdf(1.282 | mu_z
         , 1);  // 0.10 < p < 0.20, pos.
      p[4] = normal_cdf(1.282 | mu_z, 1) - normal_cdf(0 | mu_z, 1)
         ;  // p > 0.20, pos.
      p[5] = normal_cdf(0 | mu_z, 1) - normal_cdf(-1.282 | mu_z,
         1);  // p > 0.20, neg.
      p[6] = normal_cdf(-1.282 | mu_z, 1) - normal_cdf(-1.645 |
         mu_z, 1);  // 0.10 < p < 0.20, neg.
      p[7] = normal_cdf(-1.645 | mu_z, 1) - normal_cdf(-1.960 |
         mu_z, 1);  // 0.05 < p < 0.10, neg.
      p[8] = normal_cdf(-1.960 | mu_z, 1);  // p < 0.05, neg.

      real norm_const = dot_product(w, p);

      int bin;
      if (z >= 1.960) {
        bin = 1;
      } else if (z >= 1.645) {
        bin = 2;
      } else if (z >= 1.282) {
        bin = 3;
      } else if (z >= 0) {
        bin = 4;
      } else if (z > -1.282) {
        bin = 5;
      } else if (z > -1.645) {
        bin = 6;
      } else if (z > -1.960) {
        bin = 7;
      } else {
        bin = 8;
      }
      log_lik_sel[i] = log_f + log(w[bin]) - log(norm_const);
    }
    target += sum(log_lik_sel);
}
```

```
84
85 generated quantities {
86     vector[N] log_lik;
87     vector[8] sel_prob;
88     vector[8] bin_width = [0.025, 0.025, 0.05, 0.4, 0.4, 0.05,
          0.025, 0.025]';
89
90     sel_prob[1] = 1.0;
91     for (j in 2:8)
92       sel_prob[j] = (w[j] / bin_width[j]) / (w[1] / bin_width[1]);
93
94     for (i in 1:N)
95       log_lik[i] = normal_lpdf(cd[i] | dhat[i], se[i]);
96 }
```

H.2.3 Andrews & Kasy selection model

```
1  data{
2      int<lower=1> N;  // number of observations
3      vector[N] cd;  // effect sizes
4      vector[N] se;  // standard errors
5      int<lower=1> P;  // number of studies
6      array[N] int pid; // index of studies
7      int<lower=1> K;  // number of covariates
8      matrix[N,K] x;  // covariates matrix
9      int<lower=1> J;  // number of spline basis functions
10     matrix[N,J] B;  // spline basis matrix
11 }
12
13 parameters {
14     vector[K] beta;
15     vector[P] mup;
16     vector[N] eps;
17     real<lower=0> tau;
18     real<lower=0> sigma;
19     real<lower=2> df;
20     vector[J] alpha;  // spline weights for selection
21 }
22
23 transformed parameters {
24     vector[N] dhat = x * beta + mup[pid] + eps;
25     vector[N] sel_prob = inv_logit( B * alpha );
26 }
27
28 model {
29     beta ~ normal( 0 , 5 );
30     tau ~ exponential( 1 );
31     sigma ~ exponential( 1 );
32     df ~ exponential( 0.5 );
33     alpha ~ normal( 0 , 5 );
34
35     // residuals distribution:
36     eps ~ student_t(df, 0, sigma);
37
```

```
    // distribution of paper-level residuals
    mup ~ normal( 0 , tau );

    // likelihood (with selection correction)
    target += normal_lpdf(cd | dhat, se) + sum(log(sel_prob));
}

generated quantities {
    vector[N] log_lik;

    for (i in 1:N)
      log_lik[i] = normal_lpdf(cd[i] | dhat[i], se[i]);
}
```

I MAIVE Analyses

MAIVE addresses the potential endogeneity of reported precision in meta-regression. MAIVE replaces the reported squared standard error, SE_i^2, with the component predicted by sample size, $1/N_i$, which is less directly manipulable. Specifically, it relies on the following regression:

$$SE_i^2 = \widehat{\psi}_0 + \widehat{\psi}_1 \frac{1}{N_i} + \nu_i.$$

We then construct the instrumented standard error as:

$$SE_i^{IV} = \sqrt{\widehat{\psi}_0 + \widehat{\psi}_1 \frac{1}{N_i}}.$$

In the second stage, we re-estimate our main meta-analytic specifications using this instrumented precision measure. This includes the baseline measurement-error model (MEM), PET–PEESE, the Vevea–Hedges selection models, and the continuous-selection models of Andrews–Kasy.

The resulting estimates are very similar to those reported in the main analysis, indicating that our conclusions are not driven by dependence between standardized effect sizes and their reported standard errors. The MAIVE adjustment suggests that the originally reported standard errors are likely somewhat downward-biased and excessively heterogeneous relative to the precision implied by sample size. Specifically, the mean standard error rises from 0.150 for the originally reported SE to 0.164 for the instrumented SE, and the median increases from 0.129 to 0.148. At the same time, the standard deviation declines from 0.093 to 0.064.

In MAIVE framework, reported precision is considered spurious when it exceeds the precision implied by a correctly specified model with appropriate functional form and error-term properties. Such overstatement can undermine subsequent analyses that rely on reported standard errors, including inverse-variance weighting and publication-bias correction methods. Therefore, the increase in the instrumented standard errors is consistent with the this framework and indicates a

correction for potentially spuriously precise estimates. While the lower dispersion implies that idiosyncratic variation in reported standard errors—possibly arising from specification choices, reporting practices, or model-dependent standard-error calculations—is partially filtered out.

Adjusted PET–PEESE:

The pattern does not change substantially under this adjusted procedure—the slope is significantly positive only under the fixed-effect specification, whereas it becomes statistically indistinguishable from zero under the random-effects specification.

$\lambda_{\text{PET,FE}} = 1.32$ $(p < 0.001)$; $\lambda_{\text{PEESE,FE}} = 2.56$ $(p < 0.001)$.

$\lambda_{\text{PET,RE}} = 0.40$ $(p = 0.154)$; $\lambda_{\text{PEESE,RE}} = 0.27$ $(p = 0.371)$.

Adjusted Vevea-Hedges and Andrews-Kasy models:

The selection mechanisms displayed in Figure 10 remain similar to those in the unadjusted specifications.

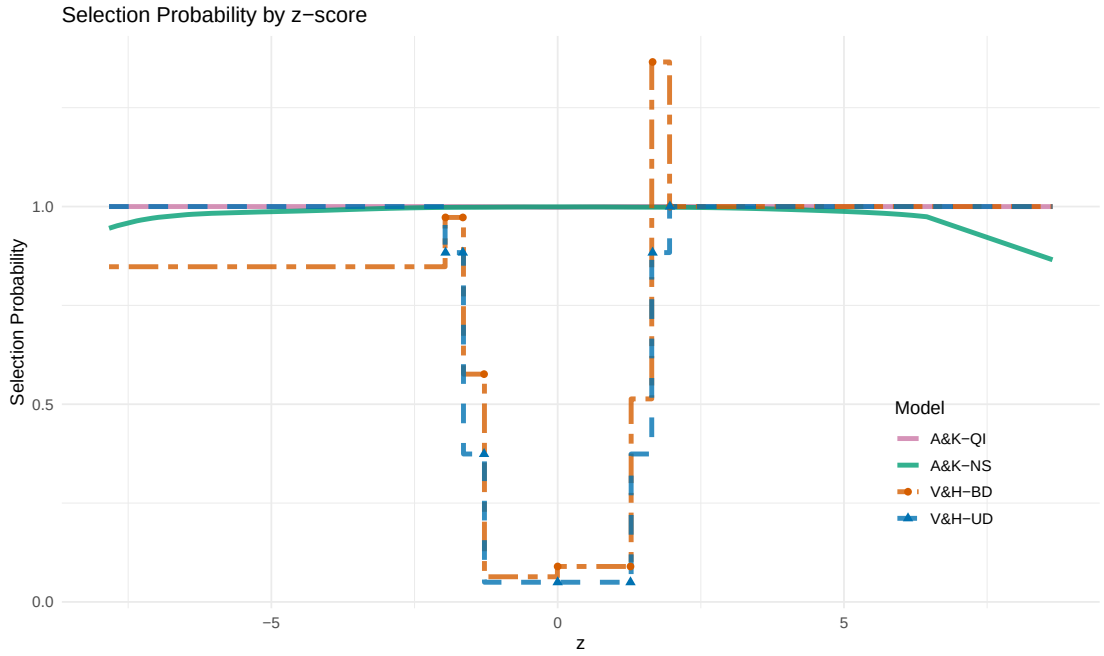

Figure 10: Posterior mean selection probabilities for the Vevea & Hedges and Andrews & Kasy models under random-effects specifications.

Adjusted PoBMA:

The LOO differences, stacking weights, and posterior means of μ for all instrumented variants under PoBMA are reported in Table 4. The bias-corrected population means for the instrumented variants do not substantively differ from those

of their uninstrumented counterparts across all random-effects models, with the exception of PET, which now implies a negative incentive effect. However, PET receives only negligible stacking weight (0.003), and its 95% credible interval still lies entirely within the negligible range of $[-0.2, 0.2]$.

The resulting model-averaged posterior estimate of the underlying effect size under PoBMA is

$$\mu_{\text{PoBMA,IV}} = 0.037 \quad (95\% \text{ CrI } [0.002,\ 0.073]),$$

which again falls squarely within the negligible-effect region and is even smaller than its uninstrumented counterpart ($\mu_{\text{PoBMA}} = 0.047$).

Model	**Δelpd**	**weight**	μ
Random A&K-NS	0.0 (0.0)	0.392	0.041 (-0.009, 0.090)
Random PEESE	-0.9 (2.8)	0.403	0.035 (-0.035, 0.108)
Random A&K-QI	-2.2 (2.7)	0.002	0.042 (-0.009, 0.092)
Random MEM	-3.6 (3.7)	0.058	0.043 (-0.006, 0.094)
Random PET	-4.1 (3.3)	0.003	-0.022 (-0.155, 0.116)
Random V&H-BD	-6.2 (3.8)	0	0.035 (-0.023, 0.091)
Random V&H-UD	-58.6 (11.6)	0.142	0.033 (-0.008, 0.074)
Fixed PET	-747.9 (91.1)	0	-0.186 (-0.229,-0.143)
Fixed PEESE	-763.8 (91.9)	0	-0.056 (-0.075,-0.037)
Fixed V&H-UD	-781.5 (93.5)	0	0.000 (-0.001, 0.001)
Fixed A&K-QI	-783.9 (93.5)	0	-0.003 (-0.014, 0.008)
Fixed MEM	-783.9 (93.5)	0	-0.003 (-0.014, 0.008)
Fixed A&K-NS	-784.0 (93.5)	0	-0.003 (-0.014, 0.009)
Fixed V&H-BD	-812.1 (93.4)	0	-0.040 (-0.063,-0.017)

Table 4: Leave-one-out cross-validation (LOO) results and stacking weights for the 14 model specifications included in the Prediction-optimized Model Averaging (PoBMA) framework. The 'Model' column lists each specification, indicating fixed- or random-effects assumptions and the type of publication bias correction applied: hierarchical Bayesin measurement error model (MEM), Precision-Effect Test (PET), Precision-Effect Estimate with Standard Error (PEESE), Vevea & Hedges selection models under unidirectional (V&H-UD) or bidirectional (V&H-BD) specifications, and Andrews & Kasy selection models using natural spline (A&K-NS) or quadratic interpolation (A&K-QI). Δelpd denotes the difference in expected log predictive density relative to the best-fitting model (standard error in parentheses), 'weight' indicates the model stacking weight, and μ gives the posterior mean of the bias-corrected effect size (with 95% credible intervals in parentheses).

J Heterogeneity Analysis

J.1 Subgroup statistics summery

Subgroup characteristics	Raw Effects			PoBMA Estimates			Obs.
	Mean	Med.	SD	Mean	Med.	SD	*N*
Topic Domain							
Temporal discounting	-0.049	-0.002	0.290	-0.030	-0.007	0.185	119
Risk taking	0.042	0.028	0.353	0.036	0.028	0.278	505
Source of Data							
Lab experiment	0.056	0.068	0.383	0.053	0.061	0.290	450
Field experiment	-0.043	-0.072	0.106	-0.040	-0.068	0.094	54
Online experiment	-0.074	-0.077	0.204	-0.070	-0.066	0.168	117
Subject Pool							
University students/staff	0.052	0.062	0.380	0.048	0.062	0.282	335
Non-university participants	-0.036	-0.033	0.227	-0.032	-0.036	0.167	249
Location of Experiment							
North America	0.027	0.042	0.374	0.037	0.039	0.262	209
Europe	-0.033	-0.017	0.291	-0.038	-0.020	0.228	231
Africa	0.020	0.028	0.194	0.020	0.025	0.122	46
Asia	0.120	0.113	0.336	0.101	0.107	0.258	90
Oceania	0.116	0.000	0.486	0.112	0.045	0.430	48
Treatment Design							
Within-subjects	0.107	0.130	0.365	0.101	0.109	0.293	246
Between-subjects	-0.029	-0.032	0.318	-0.027	-0.027	0.230	378
Elicitation Task							
Binary choice	0.032	0.040	0.313	0.039	0.041	0.236	344
Choice list	-0.019	-0.033	0.297	-0.032	-0.031	0.233	126
Bidding	-0.021	0.010	0.515	-0.010	-0.005	0.375	57
Other type of task	0.080	0.108	0.372	0.059	0.093	0.305	97
Trial Procedure							
Static procedure	0.036	0.022	0.360	0.030	0.024	0.278	498
Iterative adjustment procedure	-0.022	0.002	0.263	-0.004	0.011	0.195	126
Incentive Realization Scheme							
All subjects paid	0.047	0.042	0.357	0.044	0.041	0.274	506
Random subjects paid	-0.072	-0.073	0.257	-0.066	-0.075	0.190	118
All decisions realized	0.090	0.126	0.399	0.082	0.080	0.286	181
Random decisions realized	-0.003	0.000	0.314	-0.001	0.005	0.250	443
Payoff Domain							
Gain	0.018	0.015	0.309	0.016	0.019	0.234	443
Loss	0.014	0.014	0.278	0.017	0.059	0.234	87
Mixed	0.080	0.108	0.480	0.076	0.094	0.366	111
Reward Type							
Monetary	0.024	0.014	0.347	0.022	0.020	0.269	586
Non-monetary	0.034	0.044	0.290	0.039	0.039	0.154	38
Estimation Method							
Parametric	-0.058	-0.035	0.253	-0.056	-0.057	0.186	166
Non-parametric	0.054	0.053	0.366	0.052	0.048	0.282	458
Publication Status							
Not published in a peer-reviewed journal	0.063	0.042	0.296	0.031	0.040	0.177	46
Published in a peer-reviewed journal	0.021	0.018	0.347	0.023	0.021	0.270	578
Published in a Top-5 economics journal	0.038	-0.023	0.383	0.033	-0.010	0.340	102
Published in an economics journal	-0.012	-0.019	0.353	-0.008	-0.006	0.295	345
Published in another field journal	0.070	0.069	0.326	0.063	0.061	0.213	279

Table 5: Summary of effect sizes by subgroup characteristics. Raw effects correspond to the original Cohen's d, while PoBMA estimates refer to the posterior estimates of true effect sizes.

J.2 Meta regression result

Moderators	(1)	(2)	(3)	(4)
Time (vs Risk)	−0.046	−0.045	−0.040	−0.039
	(0.036)	(0.035)	(0.036)	(0.036)
Within (vs Between)	**0.111**	**0.111**	0.102	0.102
	(0.053)	(0.052)	(0.055)	(0.056)
Loss (vs Gain)	−0.019	−0.017	−0.024	−0.022
	(0.044)	(0.045)	(0.046)	(0.045)
Mixed (vs Gain)	**-0.124**	**-0.125**	**-0.136**	**-0.136**
	(0.038)	(0.038)	(0.039)	(0.038)
Param (vs Nonpar)	**-0.089**	**-0.089**	**-0.088**	**-0.088**
	(0.030)	(0.031)	(0.030)	(0.031)
Field (vs Lab)	−0.046	−0.044	−0.110	−0.111
	(0.093)	(0.093)	(0.133)	(0.134)
Online (vs Lab)	−0.047	−0.047	−0.040	−0.041
	(0.073)	(0.074)	(0.080)	(0.080)
Pay all Ss	0.018		0.014	
	(0.058)		(0.061)	
Prob. Ss Paid		0.020		0.018
		(0.064)		(0.066)
Inc. all Tasks	0.043	0.044	0.017	0.018
	(0.054)	(0.053)	(0.058)	(0.059)
Published	0.031	0.029	0.004	0.004
	(0.115)	(0.115)	(0.138)	(0.136)
EconJ	−0.019	−0.018	−0.006	−0.004
	(0.057)	(0.057)	(0.062)	(0.063)
Africa (vs N.America)			0.045	0.046
			(0.152)	(0.151)
Asia (vs N.America)			0.102	0.104
			(0.091)	(0.091)
Europe (vs N.America)			−0.011	−0.010
			(0.069)	(0.071)
Oceania (vs N.America)			−0.102	−0.104
			(0.172)	(0.172)
Constant	0.017	0.016	0.046	0.038
	(0.124)	(0.128)	(0.143)	(0.145)

Table 6: Effects significant at the 5% level are highlighted in bold, and standard errors are reported in parentheses.

K Analyses on Choice Variability

Publication bias

The correlations between effect size and precision are sizable: $|d_i|$ is negatively correlated with $\sqrt{N}$ ($\rho = -0.366$, $p < 0.001$), with similar patterns for positive effects ($\rho = -0.312$, $p < 0.001$) and negative effects ($\rho = -0.503$, $p < 0.001$). These findings indicate pronounced small-study effects. Although such patterns do not necessarily imply publication bias, publication bias is one plausible mechanism that can generate them.

A standard diagnostic for funnel-plot asymmetry is Egger's regression, which regresses the standardized effect size, d_i/se_i, on study precision, $1/se_i$. Under the null hypothesis of no small-study effects, the intercept should be zero. Applied to absolute effect sizes, Egger's test yields a strongly positive intercept ($\beta_0 = 0.618$, 95% CrI $[0.265, 0.967]$; slope $\beta_1 = 0.071$, 95% CrI $[0.034, 0.108]$), indicating that small, imprecise studies tend to report disproportionately large deviations from zero. In Egger's framework, this is the classical pattern consistent with publication bias.

When applied to signed effect sizes, however, this pattern disappears. The intercept is small and uncertain ($\beta_0 = -0.149$, 95% CrI $[-0.618, 0.291]$), while the slope is close to zero ($\beta_1 = 0.081$, 95% CrI $[0.035, 0.129]$). This divergence is informative: it suggests that small studies tend to report more extreme effects, but not systematically in either the positive or negative direction. In other words, the small-study pattern we observe concerns magnitude rather than sign. Such symmetric exaggeration is compatible with publication bias, for example if journals preferentially publish large effects in either direction, but it is also compatible with genuine heterogeneity combined with sampling noise.

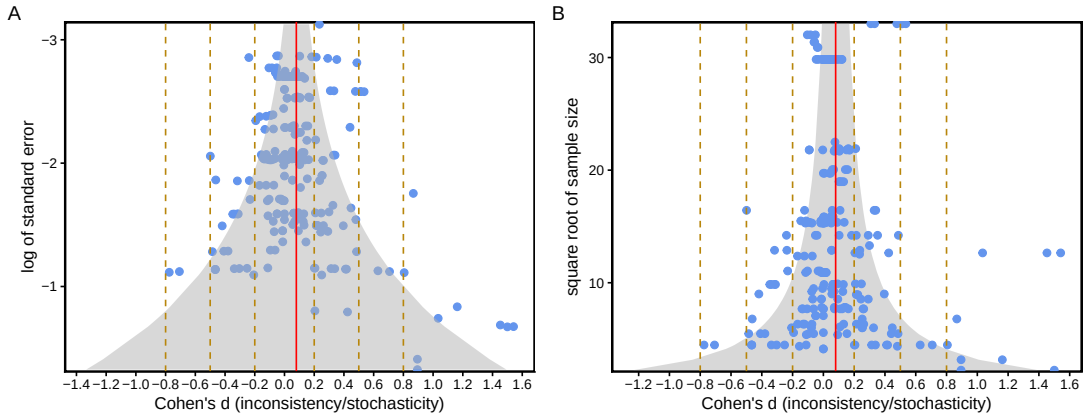

Figure 11: Funnel plot of Cohen's d against its log-standard error

The figure plots the raw effect sizes d_i against $-ln(se_i)$ (panel A) and against $\sqrt{N}$ (panel B). The gray area in panel A indicates a zone containing non-significant results. The gray area in panel B provides a similar measure, given by $\frac{1.5}{\sqrt{N}}$. The scaling factor of 1.5 is used because it approximates the average *standard deviation* in the sample.

PET–PEESE:

The slope is significantly positive only under the random-effects specification, whereas it becomes statistically indistinguishable from zero under the fixed-effects specification.

$\lambda_{\text{PET,FE}} = -0.16$ ($p = 0.134$); $\lambda_{\text{PEESE,FE}} = 0.45$ ($p = 0.130$).

$\lambda_{\text{PET,RE}} = 0.75$ ($p < 0.01$); $\lambda_{\text{PEESE,RE}} = 2.15$ ($p < 0.001$).

Vevea-Hedges and Andrews-Kasy models:

The selection mechanisms displayed in Figure 12 are similar to those for the preference outcomes.

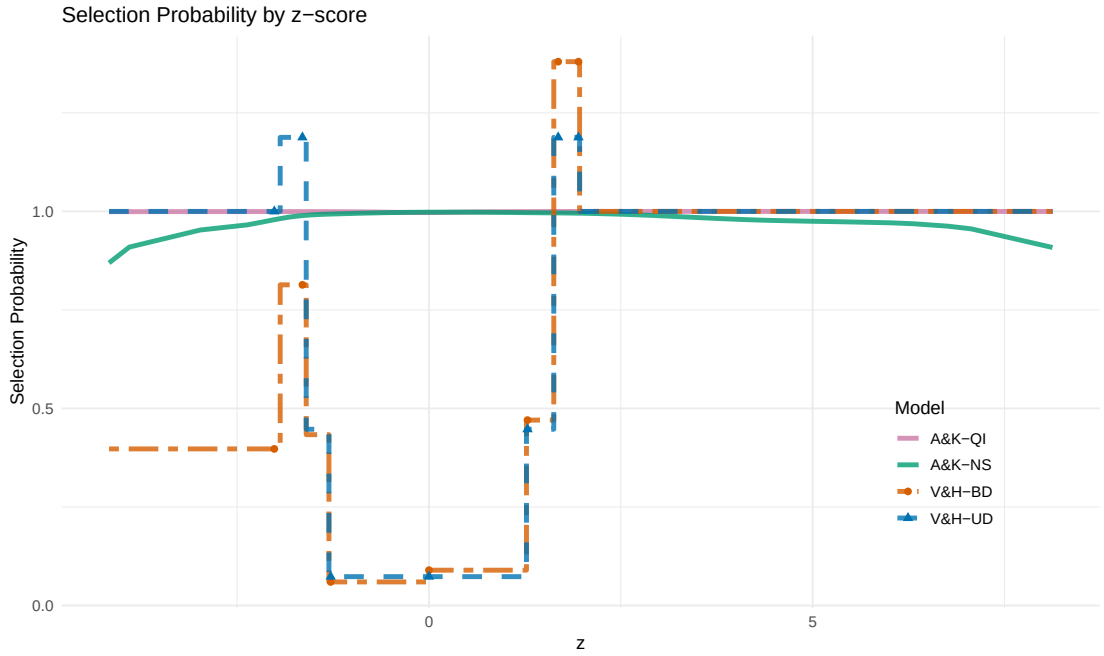

Figure 12: Posterior mean selection probabilities for the Vevea & Hedges and Andrews & Kasy models under random-effects specifications.

PoBMA

The LOO differences, stacking weights, and posterior means of μ for all included models under PoBMA are reported in Table 7. Across all random-effects models, the bias-adjusted population mean remains small. Posterior means range from approximately -0.029 (PET) to 0.072 (A&K-NS), and all associated 95% credible intervals lie entirely within the "negligible" range of $[-0.2, 0.2]$.

The resulting model-averaged posterior estimate of the underlying effect size under PoBMA is

$$\mu_{\text{PoBMA}} = 0.034 \quad (95\%\ \text{CrI}\ [0.009,\ 0.059]),$$

which again falls squarely within the negligible region.

Model	Δelpd	weight	μ
Random PEESE	0.0 (0.0)	0.349	0.018 (-0.040, 0.077)
Random PET	-1.2 (1.6)	0.048	-0.029 (-0.125, 0.065)
Random A&K-NS	-5.4 (3.6)	0.193	0.072 (0.019, 0.132)
Random MEM	-5.4 (3.6)	0	0.070 (0.017, 0.129)
Random A&K-QI	-7.1 (3.6)	0	0.071 (0.019, 0.128)
Random V&H-UD	-13.5 (6.0)	0.210	0.052 (0.010, 0.098)
Random V&H-BD	-15.8 (4.9)	0	0.055 (-0.010, 0.123)
Fixed A&K-NS	-169.8 (43.1)	0	0.067 (0.054, 0.081)
Fixed MEM	-169.9 (43.0)	0	0.068 (0.054, 0.081)
Fixed A&K-QI	-170.0 (43.1)	0	0.067 (0.054, 0.081)
Fixed PEESE	-171.3 (43.7)	0.064	0.062 (0.046, 0.078)
Fixed PET	-171.5 (42.7)	0	0.082 (0.052, 0.110)
Fixed V&H-BD	-172.9 (41.7)	0	0.077 (0.059, 0.094)
Fixed V&H-UD	-212.0 (58.3)	0.137	0.002 (-0.001, 0.007)

Table 7: Leave-one-out cross-validation (LOO) results and stacking weights for the 14 model specifications included in the Prediction-optimized Model Averaging (PoBMA) framework. The 'Model' column lists each specification, indicating fixed- or random-effects assumptions and the type of publication bias correction applied: hierarchical Bayesin measurement error model (MEM), Precision-Effect Test (PET), Precision-Effect Estimate with Standard Error (PEESE), Vevea & Hedges selection models under unidirectional (V&H-UD) or bidirectional (V&H-BD) specifications, and Andrews & Kasy selection models using natural spline (A&K-NS) or quadratic interpolation (A&K-QI). Δelpd denotes the difference in expected log predictive density relative to the best-fitting model (standard error in parentheses), 'weight' indicates the model stacking weight, and μ gives the posterior mean of the bias-corrected effect size (with 95% credible intervals in parentheses).

Subgroup statistics summery

Table 8 presents the systematic subgroup analyses, disaggregating the raw effect sizes and PoBMA estimates across substantively distinct contexts. Across all categories, the PoBMA estimates remain small. Mean effect sizes range from 0.015 to 0.092, while median effect sizes range from 0.002 to 0.077, well within the conventional negligible-effect interval of $[-0.2, 0.2]$.

Subgroup characteristics	Raw Effects			PoBMA Estimates			Obs.
	Mean	Med.	SD	Mean	Med.	SD	*N*
Proxy Category							
Inconsistency	0.061	0.042	0.285	0.050	0.026	0.104	190
Stochasticity	0.166	0.045	0.390	0.081	0.031	0.135	40
Topic Domain							
Temporal discounting	0.160	0.067	0.356	0.088	0.069	0.136	41
Risk taking	0.062	0.033	0.294	0.049	0.025	0.103	189
Source of Data							
Lab experiment	0.090	0.063	0.351	0.059	0.039	0.114	165
Field experiment	0.091	0.001	0.221	0.080	0.002	0.156	24
Online experiment	0.032	0.033	0.065	0.027	0.026	0.037	41
Subject Pool							
University students/staff	0.060	0.055	0.342	0.046	0.025	0.105	125
Non-university participants	0.104	0.033	0.260	0.067	0.031	0.116	105
Location of Experiment							
North America	0.104	0.004	0.435	0.059	0.019	0.127	89
Europe	0.037	0.029	0.181	0.033	0.027	0.075	100
Africa	0.041	0.061	0.097	0.041	0.077	0.054	22
Treatment Design							
Within-subjects	0.146	0.071	0.348	0.080	0.043	0.142	52
Between-subjects	0.060	0.044	0.293	0.049	0.026	0.099	178
Elicitation Task							
Binary choice	0.046	0.011	0.286	0.040	0.023	0.092	172
Choice list	0.123	0.058	0.214	0.089	0.035	0.127	49
Bidding	0.107	-0.035	0.525	0.060	0.020	0.117	39
Trial Procedure							
Static procedure	0.094	0.056	0.350	0.063	0.026	0.125	168
Iterative adjustment procedure	0.040	0.017	0.135	0.035	0.031	0.047	62
Incentive Realization Scheme							
All subjects paid	0.099	0.063	0.340	0.067	0.042	0.122	178
Random subjects paid	0.013	0.006	0.130	0.015	0.020	0.031	52
All decisions realized	0.161	0.104	0.474	0.092	0.051	0.141	69
Random decisions realized	0.045	0.020	0.189	0.040	0.025	0.091	161
Payoff Domain							
Gain	0.077	0.046	0.243	0.056	0.031	0.110	183
Loss	0.012	0.053	0.209	0.017	0.018	0.062	20
Mixed	0.114	0.027	0.531	0.065	0.022	0.116	38
Estimation Method							
Parametric	0.049	0.024	0.193	0.036	0.026	0.097	31
Non-parametric	0.084	0.053	0.322	0.059	0.030	0.112	199
Publication Status							
Published in a peer-reviewed journal	0.077	0.019	0.318	0.053	0.025	0.114	214
Published in an econ Top5 journal	0.111	0.046	0.172	0.087	0.031	0.124	33
Published in an economics journal	0.042	0.046	0.272	0.044	0.025	0.101	137
Published in another field journal	0.135	0.018	0.348	0.072	0.044	0.122	93

Table 8: Summary of effect sizes by subgroup characteristics. Raw effects correspond to the original Cohen's d, while PoBMA estimates refer to the posterior estimates of true effect sizes. Subsets with fewer than 20 observations are excluded.

Attention: response and dwell time

An additional 27 effects, drawn from 4 out of the 73 papers (2 on temporal discounting and 2 on risk taking), capture attention-related measures, such as response time and dwell time. These measures were coded such that positive values indicate greater attentional effort under real incentives, as reflected in longer response or dwell times.

For the raw effects, 44.4% of the effect sizes are negligible, while the remaining 55.6% fall into Cohen's "small" category. For the signed effects, both the median (-0.119) and the mean (-0.143) are negative, suggesting that subjects even spend less time making decisions under real incentives than under hypothetical scenarios, although the aggregate effect is negligible.

The BHMEM estimate of the meta-analytic mean is $\mu = -0.011$, with a 95% credible interval (CrI) of $[-0.295,\ 0.269]$, indicating an overall null effect.

L List of Included Papers

Below, we include a list of papers that are currently included in the meta-analysis.

Abdellaoui, M., Baillon, A., Placido, L., & Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review, 101*(2), 695–723.

Alsharawy, A., Zhang, X., Ball, S. B., & Smith, A. (2021). Incentives Affect the Process of Risky Choice. *Available at SSRN 3943681*.

Baker, F., Johnson, M. W., & Bickel, W. K. (2003). Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. *Journal of Abnormal Psychology, 112*(3), 382.

Barreda-Tarrazona, I., Jaramillo-Gutiérrez, A., Navarro-Martínez, D., & Sabater-Grande, G. (2011). Risk attitude elicitation using a multi-lottery choice task: Real vs. hypothetical incentives. *Spanish Journal of Finance and Accounting/Revista Espanola De Financiación Y Contabilidad, 40*(152), 613–628.

Battalio, R. C., Kagel, J. H., & Jiranyakul, K. (1990). Testing between alternative models of choice under uncertainty: Some initial results. *Journal of Risk and Uncertainty, 3*(1), 25–50.

Beattie, J., & Loomes, G. (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty, 14*(2), 155–168.

Bickel, W. K., Pitcock, J. A., Yi, R., & Angtuaco, E. J. C. (2009). Congruence of BOLD response across intertemporal choice conditions: fictive and real money gains and losses. *Journal of Neuroscience, 29*(27), 8839–8846.

Bohm, P. (1994). Time preference and preference reversal among experienced subjects: The effects of real payments. *The Economic Journal, 104*(427), 1370–1378.

Brañas-Garza, P., Estepa-Mohedano, L., Jorrat, D., Orozco, V., & Rascón-Ramírez, E. (2021). To pay or not to pay: Measuring risk preferences in lab and field. *Judgment and Decision Making, 16*(5), 1290–1313.

Brañas-Garza, P., Jorrat, D., Espín, A. M., & Sánchez, A. (2023). Paid and hypothetical time preferences are the same: Lab, field and online evidence. *Experimental Economics, 26*(2), 412–434.

Butler, D. J., & Loomes, G. C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review, 97*(1), 277–297.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics, 2*(2), 107–127.

Cox, J. C., & Grether, D. M. (1996). The preference reversal phenomenon: Response mode, markets and incentives. *Economic Theory, 7*(3), 381–405.

Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics, 1*(2), 115–131.

Etchart-Vincent, N., & l'Haridon, O. (2011). Monetary incentives in the loss domain and behavior toward risk: An experimental comparison of three reward schemes including real losses. *Journal of Risk and Uncertainty, 42*(1), 61–83.

Fan, C.-P. (2002). Allais paradox in the small. *Journal of Economic Behavior & Organization, 49*(3), 411–421.

Ferrey, A. E., & Mishra, S. (2014). Compensation method affects risk-taking in the Balloon Analogue Risk Task. *Personality and Individual Differences, 64*, 111–114.

Fidanoski, F., Dixit, V., & Ortmann, A. (2025). Risky intertemporal choices have a common value function, but a separate choice function. *I4R Discussion Paper Series*.

Freeman, D. J., & Mayraz, G. (2019). Why choice lists increase risk taking. *Experimental Economics, 22*(1), 131–154.

Gneezy, U., Halevy, Y., Hall, B., Offerman, T., & van de Ven, J. (2024). How real is hypothetical? a high-stakes test of the allais paradox. *Technical report*.

Green, R. M., & Lawyer, S. R. (2014). Steeper delay and probability discounting of potentially real versus hypothetical cigarettes (but not money) among smokers. *Behavioural Processes, 108*, 50–56.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review, 69*(4), 623–638.

Hackethal, A., Kirchler, M., Laudenbach, C., Razen, M., & Weber, A. (2023). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty, 66*(2), 189–213.

Hinvest, N. S., & Anderson, I. M. (2010). The effects of real versus hypothetical reward on delay and probability discounting. *Quarterly Journal of Experimental Psychology, 63*(6), 1072–1084.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655.

Horn, S., & Freund, A. M. (2022). Adult age differences in monetary decisions with real and hypothetical reward. *Journal of Behavioral Decision Making, 35*(2), e2253.

Huck, S., & Müller, W. (2012). Allais for all: Revisiting the paradox in a large representative sample. *Journal of Risk and Uncertainty, 44*(3), 261–293.

Irwin, J. R., McClelland, G. H., & Schulze, W. D. (1992). Hypothetical and real consequences in experimental auctions for insurance against low-probability risks. *Journal of Behavioral Decision Making, 5*(2), 107–116.

Johnson, M. W., & Bickel, W. K. (2002). Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the Experimental Analysis of Behavior, 77*(2), 129–146.

Kachelmeier, S. J., & Shehata, M. (1992). Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *The American Economic Review*, 1120–1141.

Keren, G., & Gerritsen, L. E. M. (1999). On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica, 103*(1–2), 149–172.

Kirby, K. N., & Maraković, N. N. (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes, 64*(1), 22–30.

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Hypothetical and real. *Organizational Behavior and Human Decision Processes, 89*(2), 1162–1175.

Lagorio, C. H., & Madden, G. J. (2005). Delay discounting of real and hypothetical rewards III: Steady-state assessments, forced-choice trials, and all real rewards. *Behavioural Processes, 69*(2), 173–187.

Laury, S. K., & Holt, C. A. (2008). Further reflections on the reflection effect. *Research in Experimental Economics, 12*, 405–440.

Lawyer, S. R., Schoepflin, F., Green, R., & Jenks, C. (2011). Discounting of hypothetical and potentially real outcomes in nicotine-dependent and nondependent samples. *Experimental and Clinical Psychopharmacology, 19*(4), 263.

Lawyer, S. R., Prihodova, T., Prihodova, K., Rasmussen, E., Doubkova, N., & Preiss, M. (2022). Steeper delay discounting for potentially real versus hypothetical cigarettes (but not money) in Czech Republic smokers. *The Psychological Record, 72*(2), 167–175.

Löckenhoff, C. E., Rutt, J. L., Samanez-Larkin, G. R., O'Donoghue, T., Reyna, V. F., & Ganzel, B. (2016). Dread sensitivity in decisions about real and imagined electrical shocks does not vary by age. *Psychology and Aging, 31*(8), 890.

Loomes, G., & Taylor, C. (1992). Non-transitive preferences over gains and losses. *The Economic Journal, 102*(411), 357–365.

Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology, 11*(2), 139.

Madden, G. J., Raiff, B. R., Lagorio, C. H., Begotka, A. M., Mueller, A. M., Hehli, D. J., & Wegener, A. A. (2004). Delay discounting of potentially real and hypothetical rewards: II. Between- and within-subject comparisons. *Experimental and Clinical Psychopharmacology, 12*(4), 251.

Magen, E., Dweck, C. S., & Gross, J. J. (2008). The hidden zero effect: Representing a single choice as an extended sequence reduces impulsive choice. *Psychological Science, 19*(7), 648.

Mentzakis, E., & Sadeh, J. (2021). Experimental evidence on the effect of incentives and domain in risk aversion and discounting tasks. *Journal of Risk and Uncertainty, 62*(3), 203–224.

Miller, J. R. (2019). Comparing rapid assessments of delay discounting with real and hypothetical rewards in children. *Journal of the Experimental Analysis of Behavior, 111*(1), 48–58.

Morgenstern, R., Heldmann, M., & Vogt, B. (2014). Differences in cognitive control between real and hypothetical payoffs. *Theory and Decision, 77*(4), 557–582.

Noussair, C. N., Trautmann, S. T., Van de Kuilen, G., & Vellekoop, N. (2013). Risk aversion and religion. *Journal of Risk and Uncertainty, 47*(2), 165–183.

Noussair, C. N., Trautmann, S. T., & Van de Kuilen, G. (2014). Higher order risk attitudes, demographics, and financial decisions. *Review of Economic Studies, 81*(1), 325–355.

Okouchi, H. (2023). Real, potentially real, and hypothetical monetary rewards in probability discounting. *Journal of the Experimental Analysis of Behavior, 120*(3), 406–415.

Rabin, M., & Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review, 99*(4), 1508–1543.

Reyes-Huerta, H. E., Calvillo, A., Tovar, E., Balbontín, K., & Robles, E. (2025). A pilot study of delay discounting of real and hypothetical Lego bricks by children: Systematicity of data depends on task type. *The Psychological Record, 75*(1), 105–110.

Robertson, S. H., & Rasmussen, E. B. (2018). Comparison of potentially real versus hypothetical food outcomes in delay and probability discounting tasks. *Behavioural Processes, 149*, 8–15.

Robinson, P. J., & Botzen, W. J. W. (2019). Determinants of probability neglect and risk attitudes for disaster risk: An online experimental study of flood insurance demand among homeowners. *Risk Analysis, 39*(11), 2514–2527.

Robinson, P. J., & Botzen, W. J. W. (2020). Flood insurance demand and probability weighting: The influences of regret, worry, locus of control and the threshold of concern heuristic. *Water Resources and Economics, 30*, 100144.

Rommel, J., Hermann, D., Müller, M., & Mußhoff, O. (2019). Contextual framing and monetary incentives in field experiments on risk preferences: evidence from German farmers. *Journal of Agricultural Economics, 70*(2), 408–425.

Rotella, A., Fogg, C., Mishra, S., & Barclay, P. (2019). Measuring delay discounting in a crowdsourced sample: An exploratory study. *Scandinavian Journal of Psychology, 60*(6), 520–527.

Scheres, A., Sumiya, M., & Thoeny, A. L. (2010). Studying the relation between temporal reward discounting tasks used in populations with ADHD: a factor analysis. *International Journal of Methods in Psychiatric Research, 19*(3), 167–176.

Schoemaker, P. J. H. (1990). Are risk-attitudes related across domains and response modes? *Management Science, 36*(12), 1451–1463.

Schunk, D., & Betsch, C. (2006). Explaining heterogeneity in utility functions by individual differences in decision modes. *Journal of Economic Psychology, 27*(3), 386–401.

Slovic, P. (1969). Differential effects of real versus hypothetical payoffs on choices among gambles. *Journal of Experimental Psychology, 80*(3p1), 434.

Taylor, M. P. (2013). Bias and brains: Risk aversion and cognitive ability across real and hypothetical settings. *Journal of Risk and Uncertainty, 46*(3), 299–320.

Taylor, M. P. (2017). Information acquisition under risky conditions across real and hypothetical settings. *Economic Inquiry, 55*(1), 352–367.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Ubfal, D. (2016). How general are time preferences? Eliciting good-specific discount rates. *Journal of Development Economics, 118*, 150–170.

Veettil, P. C., Yashodha, Y., & Vecci, J. (2025). Hypothetical bias and cognitive ability: Farmers' preference for crop insurance products. *American Journal of Agricultural Economics, 107*(3), 888–924.

Vieider, F. M. (2011). Separating real incentives and accountability. *Experimental Economics, 14*(4), 507–518.

Vieider, F. M. (2018). Violence and risk preference: experimental evidence from Afghanistan: comment. *American Economic Review, 108*(8), 2366–2382.

Von Gaudecker, H.-M., Van Soest, A., & Wengström, E. (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review, 101*(2), 664–694.

Wiseman, D. B., & Levin, I. P. (1996). Comparing risky decision making under conditions of real and hypothetical consequences. *Organizational Behavior and Human Decision Processes, 66*(3), 241–250.

Xu, S., Fang, Z., & Rao, H. (2013). Real or hypothetical monetary rewards modulates risk taking behavior. *Acta Psychologica Sinica*.

Xu, S., Pan, Y., Wang, Y., Spaeth, A. M., Qu, Z., & Rao, H. (2016). Real and hypothetical monetary rewards modulate risk taking in the brain. *Scientific Reports, 6*(1), 29520.

Xu, S., Pan, Y., Qu, Z., Fang, Z., Yang, Z., Yang, F., Wang, F., & Rao, H. (2018). Differential effects of real versus hypothetical monetary reward magnitude on risk-taking behavior and brain activity. *Scientific Reports, 8*(1), 3712.

Xu, S., Xiao, Z., & Rao, H. (2019). Hypothetical versus real monetary reward decrease the behavioral and affective effects in the Balloon Analogue Risk Task. *Experimental Psychology*.

Yang, X.-L., Chen, S.-T., & Liu, H.-Z. (2022). The effect of incentives on intertemporal choice: Choice, confidence, and eye movements. *Frontiers in Psychology, 13*, 989511.