

Diagnosing Causal Structure in Meta-Analysis Using Artifact Meta-Regression

Jacob L Orquin

Aarhus University & Reykjavik University

Author Note

Correspondence concerning this article should be addressed to Jacob L. Orquin, Department of Management, Aarhus University, Fuglesangs alle 4, 8210 Aarhus V - Denmark. E-mail: jalo@mgmt.au.dk. This research was supported by the Independent Research Fund Denmark, grant number: 8046-00014A. The author declares to have no conflicts of interest. Code and data are available on

https://osf.io/5m2uv/?view_only=2d24cffb200b4ce6b9cc01b663ace8ac.

Abstract

Meta-analyses are widely used to estimate average effects, yet their credibility is often undermined by publication bias and selective reporting. Existing methods focus on correcting these biases, but pay less attention to whether variation across studies is consistent with an underlying causal process. This paper introduces a diagnostic framework that uses dose–response patterns in meta-analysis to assess causal structure. The key idea is that if a real effect exists, studies that measure it with less attenuation should report larger effect sizes. I operationalize this using artifact meta-regression, which relates observed effect sizes to study-level indicators of attenuation, such as measurement validity or reliability. Simulations show that the resulting diagnostic is conservative, exhibiting low false positive rates even under substantial publication bias. Evidence from replication–meta-analysis pairs suggests that dose–response patterns are more common in literatures that replicate successfully. An application to nudging interventions illustrates how the absence of a clear gradient can reflect limited variation rather than the absence of an effect. By treating heterogeneity as potentially informative rather than purely noise, the framework complements existing approaches and provides a new lens for evaluating the credibility of meta-analytic findings.

Keywords: meta-analysis, meta-regression, publication bias, causality, nudging

Diagnosing Causal Structure in Meta-Analysis Using Artifact Meta-Regression

Meta-analysis is widely regarded as a cornerstone of cumulative science, providing a quantitative synthesis of empirical findings across studies. Its appeal lies in increased statistical power, the ability to summarize heterogeneous literatures, and its central position in evidence hierarchies. However, the inferential status of meta-analytic estimates remains contested. A growing body of evidence suggests that meta-analyses may systematically overestimate effect sizes, particularly in domains characterized by publication bias and selective reporting (Carter et al., 2019). Kvarven et al. (2020) compared 15 meta-analytic estimates with corresponding estimates obtained from large-scale preregistered replication studies and found that meta-analytic effect sizes were almost three times larger than replication estimates, with bias corrections doing little to close this gap. Moreover, in several cases, meta-analyses reported significant effects where replications failed to reject the null hypothesis, raising concerns about false positives.

This concern became salient in a recent debate on behavioral interventions, where a meta-analysis reported a robust average effect of nudging (Mertens et al., 2022), yet a reanalysis that explicitly modeled publication bias suggested that the effect was not different from zero (Maier et al., 2022). These discrepancies highlight a fundamental limitation: meta-analysis, as typically practiced, focuses on estimating average effect sizes, yet the central challenge in many domains is not aggregation but assessing whether observed patterns are consistent with an underlying causal process. In the absence of tools to evaluate this, meta-analytic estimates risk losing their interpretability.

In this paper, I propose a framework that treats meta-analysis as diagnostic evidence consistent with a causal model based on the recovery of *dose-response relationships*. Building on the epidemiological tradition associated with Hill’s criteria (Hill, 1965) and their formalization in modern causal inference (Hirano & Imbens, 2004), I argue that systematic variation in effect sizes across theoretically meaningful dimensions can serve as a diagnostic for causal interpretation. Rather than viewing

heterogeneity as a nuisance, the framework evaluates whether variation across studies follows patterns consistent with attenuation of a common underlying effect.

To operationalize this idea, I introduce *artifact meta-regression*. In the psychometric meta-analysis tradition (Hunter & Schmidt, 2004), study artifacts such as measurement error, imperfect construct validity, range restriction, and dichotomization are treated as commensurable sources of attenuation that combine multiplicatively. I build on this insight but reinterpret their role: rather than representing differences in treatment intensity, artifact multipliers index the degree to which an underlying effect is observable in the data. Under the assumption that artifacts attenuate effect sizes toward zero, variation in artifact multipliers induces systematic variation in how strongly a common effect is expressed. This provides the basis for evaluating whether observed effect sizes follow a theoretically predicted gradient: If a real effect exists, studies that measure it with less attenuation should show larger effects. If no effect exists, improving measurement should not systematically increase effect sizes. Artifact meta-regression tests for this pattern. The contribution is not the regression model itself, but the reinterpretation of heterogeneity as a testable implication of a causal attenuation model. The diagnostic value of this approach depends critically on the source of variation in artifact multipliers. Some forms of variation, such as differences in measurement reliability or scale length driven by practical constraints, are more plausibly orthogonal to underlying effect heterogeneity and sources of bias. Other forms, such as differences in study design, populations, or research contexts, may reflect broader heterogeneity that is not solely attributable to attenuation. Accordingly, the proposed framework is most informative when variation in artifact multipliers primarily captures differences in the observability of a common underlying effect, rather than substantive differences in the effect itself.

The paper adopts a triangulation approach, combining controlled simulation, comparative evidence from existing literatures, and an applied reanalysis. Study 1 uses simulation to isolate the core mechanism underlying the proposed diagnostic, examining how dose-response patterns emerge when variation in artifact multipliers reflects

attenuation under well-defined conditions. Study 2 extends the analysis to real meta-analytic contexts by examining replication–meta-analysis pairs, assessing whether dose–response patterns are more likely to appear in literatures that successfully replicate. Study 3 applies the framework to the literature on nudging interventions, illustrating how the presence or absence of a dose–response pattern can inform the interpretation of contested empirical findings. Taken together, these studies provide complementary evidence on whether variation in effect sizes is consistent with attenuation of a common underlying effect, thereby evaluating the usefulness of dose–response patterns as a diagnostic for causal structure in meta-analysis.

Hill Criteria and the Central Role of Dose–Response

Hill’s criteria for causal inference have long provided a heuristic framework for evaluating whether observed associations reflect underlying causal relationships. Among these criteria, the biological gradient, or dose–response relationship, is distinctive in imposing a structural constraint on the data: outcomes must vary systematically with exposure. For example, if smoking causes cancer, then cancer risk should increase with smoking intensity. In contrast to criteria such as strength or consistency, which are vulnerable to selective reporting and heterogeneity, dose–response relationships require a coherent pattern of variation across levels of an explanatory variable. This makes them particularly informative for distinguishing causal effects from spurious associations.

Despite its central role in causal reasoning, dose–response has been underutilized in meta-analysis. Conventional approaches focus on estimating average effects, treating heterogeneity primarily as a nuisance. However, heterogeneity often reflects systematic variation in study characteristics—variation that can, in principle, be leveraged to assess whether the data are consistent with an underlying causal process. Reframing meta-analysis around dose–response relationships therefore shifts the inferential target from estimating average effects to testing diagnostic evidence consistent with a causal model. The relevant question becomes whether effect sizes vary predictably with theoretically meaningful dimensions.

From Hill to Formal Causal Inference: Hirano and Imbens

The idea of a dose–response relationship was formalized within the framework of modern causal inference by Hirano and Imbens, 2004. Although this framework is more commonly used in statistics and econometrics, its core intuition aligns closely with Hill’s original reasoning: causal claims are strengthened when changes in an explanatory variable are systematically associated with changes in outcomes. Let $Y(t)$ denote the outcome that would be observed if the treatment were set to level t , and define the dose–response function as

$$\mu(t) = E[Y(t)].$$

Interpreting such relationships causally requires assumptions about how treatment levels are assigned. In particular, variation in treatment intensity must not be systematically confounded with other factors that influence the outcome. The generalized propensity score (GPS) provides a framework for addressing this problem by modeling the probability of treatment levels given observed characteristics. When treatment intensity is independent of potential outcomes conditional on these characteristics, the causal dose–response function can be recovered.

This framework clarifies an important point: observing a gradient is not sufficient for causal interpretation; one must also consider whether the gradient reflects a causal mechanism or underlying differences between units. In the context of meta-analysis, the relevant question is whether variation across studies—such as differences in measurement attenuation—induces systematic variation in how strongly an underlying effect is expressed, and whether this variation can be used to diagnose causal structure under plausible assumptions.

Artifact Meta-Regression as Dose–Response Diagnostics

The usefulness of dose–response relationships in meta-analysis is constrained by the nature of treatment exposure and the availability of suitable data. I therefore propose to use study artifacts—such as imperfect construct validity, measurement error, range restriction, and dichotomization—to construct dose–response diagnostics. Study

artifacts attenuate effect sizes and, unless accounted for, introduce bias in meta-analytic estimates. Psychometric meta-analysis provides a framework for modeling these artifacts and recovering unbiased estimates under appropriate assumptions (Hunter & Schmidt, 2004).

Following Hunter and Schmidt, 2004, let $A_i > 0$ denote an artifact multiplier capturing the extent to which measurement and design attenuate the true effect ρ . The observed effect size in study i is given by

$$r_i = A_i\rho + \eta_i + \varepsilon_i, \quad (1)$$

where ε_i represents sampling error and η_i captures systematic deviations due to bias or unmodeled heterogeneity. Sampling error is assumed to be mean-zero conditional on A_i , so that it affects precision but not bias.

Within this framework, variation in A_i does not correspond to changes in the treatment itself, but to systematic differences in the extent to which the underlying effect is observed. When artifacts attenuate effects toward zero, higher values of A_i imply that a larger share of the underlying effect is revealed in the data. If a nonzero effect exists, this implies a systematic relationship between A_i and observed effect sizes; if no effect exists, no such relationship should emerge.

This interpretation allows artifact meta-regression to be used as a diagnostic for causal structure: it evaluates whether observed variation in effect sizes follows a pattern consistent with attenuation of a common underlying effect. Under stronger assumptions—most notably that variation in A_i is independent of systematic bias—the relationship can also be interpreted causally. However, the primary role of the approach is diagnostic, assessing whether the structure of the data is consistent with a causal model under plausible conditions.

Study 1. Simulation Analysis

The simulation study examines how artifact meta-regression behaves as a diagnostic for causal structure under controlled conditions. Rather than evaluating

performance across all possible data-generating processes, the simulation is designed to isolate the core mechanism underlying the proposed framework: attenuation of a common underlying effect. Specifically, it assesses whether systematic variation in artifact multipliers induces the expected dose–response pattern when an effect is present, and whether such patterns are absent when no effect exists. For comparison, we also evaluate the performance of the PET-PEESE publication bias correction method.

We manipulate three factors: the presence or absence of an effect, the number of studies included in the meta-regression ($k = 10, k = 30, k = 60, k = 100$), and the degree of publication bias (none, moderate, or severe). In the effect condition, the true average effect size is set to $r = .16$, while in the null condition the true effect is zero. Publication bias is introduced by selectively retaining studies based on statistical significance and effect direction. In the no-bias condition, all studies are retained. In the biased conditions, studies with $p < .025$ and $r > 0$ are always retained, studies with $.025 < p < .05$ and $r > 0$ are retained with an exponentially decreasing probability, and studies with $p > .05$ are retained with a uniform probability. The parameters are calibrated such that the moderate bias condition yields approximately 20% non-significant results, and the severe bias condition approximately 5%, consistent with empirical benchmarks (Carter et al., 2019; Hagger et al., 2010; Hartgerink et al., 2017). Figure 1 illustrates the relationship between the artifact multiplier and observed effect sizes across conditions. Appendix Table 3 reports descriptive statistics, including average effect sizes $\bar{r}$, heterogeneity $\bar{\tau}$, and the proportion of non-significant studies. The data-generating process simulates individual studies in which outcomes are generated as $y = \beta z + \epsilon$, where $\epsilon \sim \text{Norm}(0, \sigma_1)$ and z is a noisy measure of a latent variable x , given by $z = x + e$ with $e \sim \text{Norm}(0, \sigma_2)$. This setup induces classical measurement attenuation. Sample sizes N are drawn uniformly between 25 and 350. For each study, the effect size is computed as the correlation r between y and z , and the reliability coefficient is given by $r_{xx} = 1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$. The artifact multiplier is computed as $a = \sqrt{r_{xx}}$, capturing the extent to which measurement attenuates the underlying effect.

Artifact meta-regressions are estimated using the `metafor` package (Viechtbauer,

2010), weighting effect sizes by sample size and including the artifact multiplier as a continuous moderator. The diagnostic is coded as indicating an effect if the slope on the artifact multiplier is positive and statistically significant, and as indicating no effect otherwise.

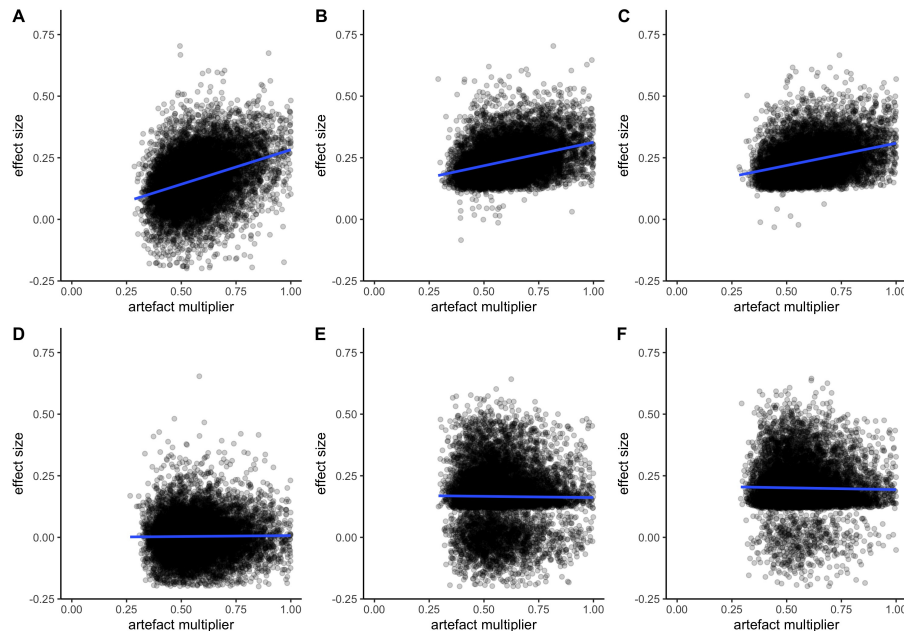


Figure 1. Effect of artifact multiplier on effect sizes under different conditions: A) $r_{xy} > 0$, no publication bias, B) $r_{xy} > 0$, moderate publication bias, C) $r_{xy} > 0$, severe publication bias, D) $r_{xy} = 0$, no publication bias, E) $r_{xy} = 0$, moderate publication bias, F) $r_{xy} = 0$, severe publication bias

For comparison, PET-PEESE analyses are conducted under the same conditions. The PET stage regresses effect sizes on their standard errors, weighted by inverse variance. If the intercept is not significant ($p > .05$), the result is coded as no effect. Otherwise, the PEESE stage regresses effect sizes on squared standard errors, and the intercept determines the final classification¹.

Table 1 reports the probability that artifact meta-regression indicates an effect across conditions. When a true effect is present, the probability of detecting a positive dose-response pattern increases with the number of studies, exceeding .81 at $k = 60$ and .96 at $k = 100$. When no effect is present, the method rarely produces false positives, with rates between .01 and .03 across all conditions. These results reflect the

¹ Simulation code is available at https://osf.io/5m2uv/?view_only=2d24cffb200b4ce6b9cc01b663ace8ac.

conservative nature of the diagnostic: it requires sufficient variation and sample size to detect a systematic gradient, but is unlikely to indicate one spuriously. Numbers in parentheses show corresponding results for PET-PEESE. While PET-PEESE performs similarly in detecting true effects, it exhibits elevated false positive rates under severe publication bias, reaching .14 at $k = 100$.

The simulation provides a baseline demonstration of how the proposed diagnostic behaves under well-defined conditions. At the same time, it does not capture all possible dependencies between artifacts and bias that may arise in practice. Rather than attempting to exhaustively model such dependencies in simulation, the following study complements this analysis by examining whether similar dose-response patterns emerge in real meta-analytic contexts where these assumptions are less likely to hold exactly.

Table 1

Meta-regression results with PET-PEESE results in parentheses.

condition	$k = 10$	$k = 30$	$k = 60$	$k = 100$
$r_{xy} > 0$, no publication bias	0.28 (0.33)	0.60 (0.83)	0.85 (0.99)	0.96 (1.00)
$r_{xy} > 0$, moderate publication bias	0.22 (0.17)	0.55 (0.57)	0.85 (0.88)	0.98 (0.98)
$r_{xy} > 0$, severe publication bias	0.20 (0.16)	0.52 (0.54)	0.81 (0.86)	0.96 (0.97)
$r_{xy} = 0$, no publication bias	0.02 (0.04)	0.02 (0.04)	0.03 (0.04)	0.02 (0.04)
$r_{xy} = 0$, moderate publication bias	0.03 (0.04)	0.01 (0.04)	0.02 (0.05)	0.01 (0.06)
$r_{xy} = 0$, severe publication bias	0.01 (0.07)	0.01 (0.10)	0.01 (0.11)	0.01 (0.14)

Study 2. Dose-Response Patterns in Replication-Meta-Analysis Pairs

Study 2 examines whether the dose-response diagnostic yields informative patterns in real meta-analytic contexts where the assumptions underlying the simulation are unlikely to hold exactly. Specifically, I analyze the replication-meta-analysis pairs compiled by Kvarven et al., 2020, which provide a unique opportunity to compare meta-analytic findings with corresponding large-scale preregistered replications. The goal is not to validate the diagnostic in a strict sense, but to assess whether dose-response patterns tend to emerge in literatures that successfully replicate, and are less prevalent in those that do not. Study 2 is best interpreted as exploratory rather than confirmatory.

Across the 15 meta-analyses examined by Kvarven et al., 2020, all reported significant positive average effects, whereas only seven of the corresponding replication

Table 2

Overview of dose-response variables identified in the meta-analyses included in Kvarven et al. (2019).

Original study	Replication	Meta-analysis	Evidence of dose-response relationships
Oppenheimer et al. (2009)	0.27*	0.50*	The sunk-cost effect decreases with longer delays between initial and subsequent investment, with medium and long delays producing smaller effects than short or no delay ($p's < .05$).
Tversky & Kahneman (1981)	0.60*	0.57*	The magnitude of the framing effect increases linearly with similarity to the original paradigm by Tversky & Kahneman (1981) ($p = .001$).
Husnu & Crisp (2010)	0.13*	0.46*	No evidence that longer imagined contact ($p = .172$) or greater detail about the interaction partner ($p = .395$) affects intergroup bias. However, greater contextual detail significantly moderates the effect ($p = .005$).
Schwarz et al. (1991)	-0.07	0.29*	No evidence that excluding less valid studies affects item-order effects in judgment.
Hauser et al. (2007)	0.79*	0.88*	No evidence that experimental design (within- vs. between-subjects) affects moral judgments ($p's > .05$). However, effects are moderated by personal contact ($p's < .01$).
Critcher & Gilovich (2008)	0.04	0.31*	No evidence that delay between exposure to the anchor and subsequent judgment affects the anchoring bias ($p > .05$).
Graham et al. (2009)	0.29*	0.68*	No evidence that attention checks or more reliable scales influence the relationship between morality and political orientation.
Jostmann et al. (2009)	0.03	0.57*	No evidence of differences between lab and field studies in the effect of weight on judgments of importance ($p = .868$).
Monin & Miller (2001)	0.15*	0.31*	No evidence of differences between actual and hypothetical choices in moral licensing ($p = .791$).
Schooler & Engstler-Schooler (1990)	0.17*	0.24*	Longer exposure to faces reduces the own-race bias ($p < .001$), whereas longer delays between exposure and recall increase it ($p < .001$).
Sripada et al. (2014)	0.04	0.62*	No evidence that depletion task duration affects the ego depletion effect ($p = .07$).
Rand et al. (2012)	-0.02	0.12*	No evidence that including versus excluding noncompliant participants affects the relationship between intuition and cooperation.
Strack et al. (1988)	0.02	0.20*	No evidence that delay between facial expression manipulation and emotional response affects the facial feedback effect ($p = .65$).
Srull & Wyer (1979)	0.06	0.35*	No evidence that delay between trait priming and evaluation affects assimilation ($p = .08$), anchoring contrast ($p = .65$), or correction contrast effects ($p = .54$).
Mazar et al. (2008)	-0.04	0.43*	No evidence of differences between actual and hypothetical behavior in the effect of moral reminders ($p = .444$).

Note. Asterisks indicate statistically significant effect sizes.

studies found effects that were not statistically different from zero. If replication success is taken as a proxy for the presence of an underlying effect, then the proposed framework implies that dose–response patterns should be more likely to appear in those meta-analyses that replicate successfully. To evaluate this, I conducted an exploratory and deliberately inclusive coding of dose–response information in the original meta-analyses. Twelve of the original meta-analyses contained sufficient information to identify such variables. In one case, relevant data was available online for reanalysis (Rabelo et al., 2015). In the remaining two cases, relevant information was not available in the original reports or online. For these, I relied on more recent meta-analyses conducted either by the original authors (Kivikangas et al., 2021) or by other research teams (Schley & Weingarten, 2025), which included the necessary information to assess potential dose–response relationships. The coding was performed by the author and was not blinded to the study hypotheses. To minimize the risk of selective reporting, the coding strategy was deliberately broad: all variables that could plausibly reflect systematic variation in attenuation or effect expression were included. These variables encompassed both continuous and categorical indicators related to validity and reliability, such as strength of experimental manipulations (e.g., duration in minutes), delays between manipulation and measurement, fidelity to original study designs, study context (lab vs. field), experimental design (within- vs. between-subjects), inclusion or exclusion of noncompliant participants, and scale length (short vs. long versions). This inclusive approach reduces the risk of cherry-picking favorable moderators, but it also implies that some coded variables may be weak or noisy proxies for attenuation. As a result, the analysis should be interpreted as a conservative test of whether informative dose–response patterns are present in the literature.

An overview of the findings is presented in Table 2. Among the seven meta-analyses with successful replications, three exhibit clear evidence of dose–response patterns and two show mixed evidence. In contrast, among the ten meta-analyses without replication support, dose–response patterns are largely absent. These results are consistent with the interpretation of dose–response relationships as a conservative

diagnostic. The presence of a systematic gradient appears to be associated with literatures that replicate successfully, suggesting that such patterns are informative about the existence of an underlying effect. At the same time, the absence of a dose–response pattern does not provide definitive evidence against a causal effect, particularly given the inclusive coding strategy and the potential noisiness of the available moderators.

Taken together with the simulation results, this study provides complementary evidence for the proposed framework. While Study 1 demonstrates how dose–response patterns emerge under controlled conditions where attenuation is the primary source of variation, Study 2 shows that similar patterns can be detected in real meta-analytic contexts, albeit imperfectly and conservatively. This supports the view that examining systematic variation in effect sizes can provide an additional lens on the credibility of meta-analytic findings. Because coding was not blinded, results should be interpreted as suggestive rather than confirmatory.

Study 3. Artifact Meta-Regression of Nudging Interventions

Study 3 provides an applied illustration of the proposed framework in a setting where the existence of an underlying effect is contested. Rather than attempting to resolve this debate directly, the analysis examines whether the structure of the data is consistent with attenuation of a common underlying effect. In the triangulation logic of the paper, this study complements the simulation (Study 1), which isolates the mechanism under controlled conditions, and the replication analysis (Study 2), which examines real-world variation across literatures, by applying the diagnostic to a high-profile empirical case.

We reanalyze the data from Mertens et al. (2022), who report a medium average effect size of choice architecture interventions ($d = .43$). A subsequent reanalysis by Maier et al. (2022), using more conservative publication bias adjustments, finds no evidence for an overall effect. This discrepancy raises the question of whether the observed effects reflect an underlying causal process or are primarily driven by bias. The dataset contains two variables that can be used to construct an artifact multiplier:

study type (conventional laboratory vs. field studies) and outcome type (continuous vs. binary). Following Mitchell, 2012, field studies can be interpreted as noisier measurements of the same underlying constructs, implying attenuation relative to laboratory studies. Specifically, laboratory studies are treated as having no attenuation ($A = 1$), while field studies are assigned a reliability of $r_{xx} = .71$, corresponding to an artifact multiplier of $A = \sqrt{.71}$. For outcome type, binary outcomes are treated as dichotomized versions of underlying continuous variables. Under standard psychometric assumptions (Hunter & Schmidt, 2004), dichotomization attenuates effect sizes. Continuous outcomes are therefore assigned $A = 1$, while binary outcomes are assigned an attenuation factor based on standard corrections for dichotomization.² The overall artifact multiplier for each effect size is computed multiplicatively as

$$A_i = \sqrt{r_{xx}} \times A_{binary},$$

reflecting the combined attenuation from study type and outcome type. This construction follows the logic of psychometric meta-analysis, where multiple sources of attenuation are treated as commensurable and combine multiplicatively. In this framework, variation in A_i captures systematic differences in how strongly an underlying effect is expressed in observed data, rather than differences in the treatment itself. Effect sizes for which reliability information could not be identified were excluded from the analysis.

To examine whether the data exhibit patterns consistent with attenuation, Figure 2 presents unweighted mean effect sizes across study types and outcome measures. As expected, effect sizes are smaller in field studies than in laboratory studies, and smaller for binary outcomes than for continuous outcomes. These patterns are consistent with the interpretation that such study characteristics attenuate the observable expression of an underlying effect. At the same time, these differences are accompanied by substantial variability, particularly for field studies, suggesting that the signal is

² The exact correction depends on assumptions about the underlying distribution and cut-point; here we apply a standard approximation consistent with psychometric meta-analysis.

relatively weak relative to noise.

We estimate a multilevel random-effects meta-regression model (REML; $k = 210$), with random intercepts for publication and effect size within publication, following Mertens et al. (2022). The overall intercept is significantly different from zero ($b = 0.582$, $SE = 0.224$, $z = 2.59$, $p = .010$, 95% CI [0.143, 1.022]), indicating a positive average effect. However, there is substantial residual heterogeneity ($Q_E(208) = 21185.22$, $p < .001$), with large variance at the publication level ($\sigma^2 = 0.198$), suggesting considerable unexplained variability across studies. Crucially, the artifact multiplier is not a statistically significant predictor of effect size ($Q_M(1) = 0.38$, $b = -0.165$, $SE = 0.268$, $z = -0.62$, $p = .538$, 95% CI [-0.691, 0.360]). In other words, the meta-regression does not detect a reliable dose–response relationship.

This result should be interpreted in light of the descriptive patterns and the properties of the diagnostic. The unweighted data exhibit variation consistent with attenuation, but this variation is noisy and unevenly distributed. Because artifact meta-regression weights observations by precision, the effective signal available to estimate the dose–response relationship is limited. As a result, the absence of a statistically significant slope does not imply the absence of an underlying effect, but rather indicates that the available evidence is insufficient to support a strong diagnostic of causal structure. Artifact multipliers are negatively associated with sample size ($b = -2.35 \times 10^{-7}$, $SE = 5.57 \times 10^{-8}$, $t = -4.21$, $p < .001$), indicating that studies with greater attenuation tend to have larger samples. Because the meta-regression weights by sampling variance, this relationship is partially accounted for in estimation, but it further reduces the effective variation available to identify systematic patterns.

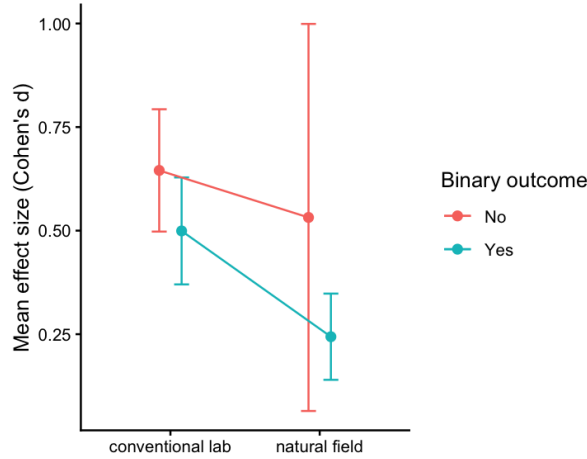


Figure 2. Unweighted mean effect sizes (Cohen’s d) by study type and outcome type. Error bars represent confidence intervals. Effect sizes are smaller in field studies than in laboratory studies and smaller for binary outcomes than for continuous outcomes, consistent with attenuation of the observable effect. Substantial variability, particularly for field studies, indicates that these patterns are relatively noisy.

Taken together, the findings illustrate the conservative nature of the proposed diagnostic. While the data contain patterns consistent with attenuation, these patterns are not sufficiently strong or precise to yield a statistically reliable dose–response relationship. In the context of the broader triangulation, this result complements the simulation evidence (which shows that detection requires sufficient variation and sample size) and the replication analysis (which suggests that clear dose–response patterns are relatively rare). Rather than providing definitive evidence for or against a causal effect, the analysis highlights the limits of what can be inferred from the available data when variation in attenuation is weak relative to noise.

Discussion

Meta-analysis is generally used as a tool for estimating average effects, yet the central challenge in many domains is not aggregation but assessing whether observed patterns are consistent with an underlying causal process. This paper reframes meta-analysis accordingly. Rather than focusing exclusively on correcting publication bias, we propose that systematic variation in effect sizes—specifically, dose–response relationships—can serve as a diagnostic for causal structure. Building on Hill’s biological gradient criterion (Hill, 1965) and its formalization in modern causal

inference (Hirano & Imbens, 2004), we show how artifact meta-regression operationalizes this idea by leveraging variation in measurement attenuation. In this framework, artifacts do not represent differences in treatment intensity, but rather determine how strongly an underlying effect is expressed in observed data.

Across three complementary approaches, the findings support this perspective while also clarifying its limits. First, the simulation study isolates the core mechanism: when variation in artifact multipliers reflects attenuation of a common underlying effect, systematic dose-response patterns emerge, and their detection becomes more likely as the number of studies increases. Importantly, the diagnostic is conservative: it exhibits low false positive rates even under substantial publication bias, but requires sufficient variation and precision to detect a gradient. Second, the analysis of replication-meta-analysis pairs shows that dose-response patterns are more likely to appear in literatures that successfully replicate. Given the inclusive coding strategy, which incorporates a broad set of potentially relevant moderators, this evidence should be interpreted conservatively. The presence of a dose-response pattern appears to be informative about the existence of an underlying effect, whereas its absence is less conclusive. This asymmetry reinforces the interpretation of dose-response relationships as a conservative diagnostic rather than a definitive test. Third, the application to nudging interventions illustrates how the framework operates in practice. The data exhibit patterns consistent with attenuation—effect sizes are smaller in field studies and for binary outcomes—but these patterns are noisy and unevenly distributed. As a result, artifact meta-regression does not detect a statistically reliable dose-response relationship. This finding should not be interpreted as evidence against a causal effect, but rather as indicating that the available variation in attenuation is too weak, relative to noise, to support a strong diagnostic. In this sense, the absence of a significant gradient reflects the limits of the data rather than a definitive conclusion about the underlying process.

Taken together, these results highlight a shift in emphasis for meta-analytic practice. Rather than treating heterogeneity solely as a nuisance to be averaged over or

explained away, the approach developed here treats it as potentially informative. The central question becomes whether variation in effect sizes follows patterns consistent with attenuation of a common underlying effect. A key implication is that the presence of a systematic dose–response relationship strengthens the case for a causal interpretation, whereas its absence—particularly in settings with limited variation or precision—should be interpreted cautiously.

At the same time, the approach relies on several important conditions. Most critically, artifact multipliers must primarily capture attenuation rather than qualitative differences in constructs, and variation in artifacts should not be strongly correlated with sources of bias. Violations of these conditions can weaken or obscure the expected dose–response pattern. In addition, the method requires sufficient variation in artifact multipliers and a reasonably large number of studies to achieve adequate power, making it less informative in small or homogeneous meta-analyses. Measurement also poses practical challenges: coding artifacts such as reliability, validity, or design features requires detailed reporting and may introduce subjectivity, particularly when information is incomplete or must be inferred.

When and How to Apply Artifact Meta-Regression

The considerations above imply that artifact meta-regression should not be applied mechanically to any available source of variation. Its usefulness depends on both the context in which it is applied and the way artifact multipliers are constructed and interpreted.

When to apply the method. Artifact meta-regression is most informative in settings where there is (i) substantial variation in artifact multipliers across studies and (ii) a plausible basis for interpreting this variation as reflecting attenuation of a common underlying effect. This condition is most likely to hold when artifact multipliers are derived from measurement-related features such as reliability, scale length, range restriction, or dichotomization, which primarily affect how strongly an effect is expressed in observed data. The method is less informative in settings where variation in study characteristics is limited, or where available moderators primarily

capture differences in design, populations, or contexts rather than attenuation. In such cases, the absence of a detectable dose–response relationship is difficult to interpret.

How to construct artifact multipliers. The construction of A_i should be guided by substantive considerations about the sources of attenuation in the literature under study. In line with psychometric meta-analysis (Hunter & Schmidt, 2004), multiple sources of attenuation can be treated as commensurable and combined multiplicatively when they plausibly affect the same underlying construct. Priority should be given to artifacts that index measurement quality or information loss, such as reliability coefficients or known effects of dichotomization. Greater caution is warranted when incorporating design-based features, such as study setting or experimental design, which may reflect both attenuation and substantive heterogeneity. In such cases, researchers should justify why the included feature primarily affects observability rather than the underlying effect, and assess robustness to alternative specifications of A_i .

How to interpret results. The interpretation of artifact meta-regression results is inherently asymmetric. A statistically significant positive relationship between A_i and observed effect sizes is consistent with attenuation of a common underlying effect and therefore strengthens the case for a causal interpretation. In contrast, the absence of a statistically significant relationship is ambiguous: it may reflect the absence of an underlying effect, but it may also arise from insufficient variation in A_i , limited statistical power, or violations of the underlying assumptions. Negative relationships between A_i and effect sizes may indicate more severe violations, such as systematic correlations between artifact multipliers and sources of bias or heterogeneity. Accordingly, the method is best understood as providing diagnostic evidence that complements, rather than replaces, existing approaches to bias correction and causal inference.

These considerations highlight that the credibility of artifact meta-regression depends less on the statistical model itself than on the plausibility of the assumptions linking artifact variation to attenuation. In this respect, the framework is analogous to other approaches that rely on quasi-exogenous variation: its inferential value is strongest

when the source of variation can be plausibly interpreted as affecting the observability of an effect without systematically altering the effect itself or introducing bias.

Despite these limitations, the framework offers a useful complement to existing approaches. Methods such as PET-PEESE focus on correcting bias in average effect estimates, whereas the approach proposed here evaluates whether the structure of the data is consistent with a causal model. These perspectives address different questions and are therefore best viewed as complementary. In contexts such as nudging, where conclusions are sensitive to modeling assumptions, examining dose-response patterns provides an additional lens through which to assess the credibility of meta-analytic findings.

More broadly, the results suggest that causal claims in meta-analysis should not rely solely on bias correction, but also on whether effect sizes exhibit theoretically predicted patterns of variation. This perspective does not eliminate uncertainty, but it provides a structured way to assess it by linking heterogeneity to underlying mechanisms. Future work could extend this approach by developing more precise measures of attenuation, exploring robustness to different artifact specifications, and integrating dose-response diagnostics with existing causal inference methods.

In conclusion, artifact meta-regression provides a principled way to leverage systematic variation across studies to diagnose causal structure in meta-analysis. While it does not replace existing methods for estimating average effects, it complements them by addressing a different question: whether the observed heterogeneity is consistent with an underlying causal process. Integrating this perspective into meta-analytic practice may help shift the field from a sole focus on bias correction toward a broader assessment of causal credibility.

References

- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144.
- Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological bulletin*, 136(4), 495.
- Hartgerink, C. H., Wicherts, J. M., & Van Assen, M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology*, 3(1), 9.
- Hill, A. B. (1965). The environment and disease: Association or causation?
- Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In *Applied bayesian modeling and causal inference from incomplete-data perspectives*. John Wiley & Sons, Ltd.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.
- Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J.-E. (2021). Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological bulletin*, 147(1), 55.
- Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4), 423–434.
- Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31), e2200300119.
- Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1), e2107346118.
- Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, 7(2), 109–117.

- Rabelo, A. L., Keller, V. N., Pilati, R., & Wicherts, J. M. (2015). No effect of weight on judgments of importance in the moral domain and evidence of publication bias from a meta-analysis. *PloS one*, *10*(8), e0134808.
- Schley, D., & Weingarten, E. (2025). 50 years of anchoring: A meta-analysis and meta-study of anchoring effects.
- Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, *36*, 1–48.

Appendix

Formal Properties of Artifact Meta-Regression

This appendix provides a formal characterization of artifact meta-regression and clarifies the conditions under which the estimated relationship between artifact multipliers and observed effect sizes can be interpreted as reflecting an underlying causal effect.

Data-Generating Process

Following the main text, let the observed effect size in study i be given by

$$r_i = A_i\rho + \eta_i + \varepsilon_i, \quad (2)$$

where $A_i > 0$ is an artifact multiplier capturing attenuation, ρ is the underlying effect, η_i represents systematic deviations due to bias or unmodeled heterogeneity, and ε_i is sampling error. We assume

$$E[\varepsilon_i \mid A_i] = 0,$$

so that sampling error affects precision but not bias.

The key feature of this formulation is that variation in A_i determines how strongly the underlying effect ρ is expressed in observed effect sizes. Under the assumption that artifacts attenuate effects toward zero, larger values of A_i correspond to less attenuation and thus a stronger observable signal.

Population Relationship Between r_i and A_i

Consider the conditional expectation of observed effect sizes:

$$E[r_i \mid A_i] = A_i\rho + E[\eta_i \mid A_i].$$

If the systematic component is independent of the artifact multiplier, that is,

$$E[\eta_i \mid A_i] = 0,$$

then

$$E[r_i \mid A_i] = A_i \rho,$$

which implies a linear relationship between A_i and r_i . In this case, variation in artifact multipliers induces systematic variation in observed effect sizes proportional to the underlying effect.

If, in addition, $\rho = 0$, then

$$E[r_i \mid A_i] = 0,$$

so that no systematic relationship between A_i and r_i should be observed.

Regression Interpretation

Let β_A denote the population slope from a linear regression of r_i on A_i :

$$\beta_A = \frac{\text{Cov}(r_i, A_i)}{\text{Var}(A_i)}.$$

Substituting the data-generating process yields

$$\text{Cov}(r_i, A_i) = \rho \text{Var}(A_i) + \text{Cov}(\eta_i, A_i) + \text{Cov}(\varepsilon_i, A_i).$$

Under the assumption that sampling error is independent of A_i ,

$$\text{Cov}(\varepsilon_i, A_i) = 0.$$

The slope can therefore be written as

$$\beta_A = \rho + \frac{\text{Cov}(A_i, \eta_i)}{\text{Var}(A_i)}. \quad (3)$$

This expression highlights that the estimated slope reflects a combination of the underlying effect and any systematic dependence between artifact multipliers and bias.

Diagnostic Interpretation

The framework developed in the main text does not rely on the strict recovery of ρ . Instead, it uses the relationship between A_i and r_i as a diagnostic for whether the data are consistent with an underlying causal effect.

The key implication of the model is:

- If $\rho \neq 0$ and $\text{Cov}(A_i, \eta_i)$ is small, then r_i should increase systematically with A_i .
- If $\rho = 0$ and $\text{Cov}(A_i, \eta_i) = 0$, then no systematic relationship should be observed.

Thus, the presence of a monotonic relationship between A_i and r_i is consistent with attenuation of a common underlying effect, whereas its absence—particularly in large samples—weakens the case for such an effect.

Conditions for Causal Interpretation

Stronger conclusions can be drawn under additional assumptions. In particular, if

$$\text{Cov}(A_i, \eta_i) = 0,$$

then

$$\beta_A = \rho,$$

and the regression slope recovers the underlying effect.

This condition requires that variation in artifact multipliers is independent of systematic sources of bias. Violations may arise if, for example, lower-quality studies both attenuate measurement and inflate effect sizes through selective reporting, or if artifact multipliers are correlated with omitted moderators.

In such cases, the slope reflects both causal signal and bias, and its interpretation must be correspondingly cautious.

Boundary Conditions

The diagnostic value of artifact meta-regression depends on several conditions:

- **Attenuation:** Artifacts must primarily reduce the magnitude of observed effects rather than induce qualitative changes in constructs.

- **Variation:** There must be sufficient variation in A_i across studies to detect systematic patterns.
- **Independence:** Variation in A_i should not be strongly correlated with sources of bias or omitted heterogeneity.
- **Monotonicity:** The relationship between attenuation and observed effects should be approximately monotonic.

These conditions define a continuum rather than a binary threshold. Violations weaken the interpretation of observed dose–response patterns but do not necessarily invalidate them entirely.

Relation to Continuous Treatment Frameworks

The interpretation of A_i as indexing the strength with which an underlying effect is expressed is analogous to continuous treatment frameworks in causal inference. However, unlike standard settings where treatment intensity reflects variation in exposure, A_i captures variation in attenuation of a common effect.

Accordingly, the role of artifact meta-regression differs from that of conventional dose–response estimation. Its primary function is to evaluate whether observed heterogeneity is consistent with a model in which a common underlying effect is differentially attenuated across studies, rather than to identify causal effects under minimal assumptions.

Table 3

Descriptive results under different publication conditions.

condition	$\bar{r}$	$\bar{\tau}$	% n.s.
$r_{xy} > 0$, no publication bias	0.16	0.06	46.52
$r_{xy} > 0$, moderate publication bias	0.23	0.01	0.47
$r_{xy} > 0$, severe publication bias	0.22	0.01	0.16
$r_{xy} = 0$, no publication bias	-0.00	0.01	95.51
$r_{xy} = 0$, moderate publication bias	0.14	0.06	21.96
$r_{xy} = 0$, severe publication bias	0.17	0.01	6.03