

Structural predictors of analytical fragility in social science: evidence from SCORE Multi100

Rachid Chabane

Independent Researcher, Roubaix, France

ORCID: 0009-0002-3032-5968

April 27, 2026

Abstract

The SCORE Multi100 programme assigned four to seven independent analysts to each of 100 published social and behavioural science papers, producing 508 analyst–paper rows, a mean paper-level strict agreement rate $\bar{a} = 0.36$, and 64 fragile papers ($a_i < 0.5$). The paper-level distribution of a_i is starkly bimodal: the 64 fragile papers coincide exactly with the 64 papers satisfying $a_i \leq 0.2$, and no paper lies in the open interval $(0.2, 0.8)$; a two-component Gaussian mixture beats a unimodal fit by $\Delta\text{BIC} \approx 1190$. From a pre-specified predictor set \mathcal{X} , two structural correlates of fragility survive the paper’s primary inferential standard. Log sample size correlates positively with agreement at paper resolution (Spearman $\rho = 0.22$, $p = 0.034$, $n = 93$), and a transparency–fragility paradox emerges in which high-badge papers show *higher* fragility than low-badge papers (means 0.286 vs. 0.552, Mann–Whitney $p = 0.013$; paper-level fragility odds ratio $\text{OR} = 2.86$, 95% CI $[1.18, 6.92]$, $p = 0.020$). The row-level `Total_Hours` effect ($\rho = 0.11$, $p = 0.018$) attenuates under analyst-clustered robust standard errors, where no row-level predictor reaches $\alpha = 0.05$; the paper-level transparency-by-complexity interaction ($p = 0.056$) sits just above the conventional threshold in its OLS specification. A pre-specified null on $||z| - 1.96|$ ($\rho = -0.04$, $p = 0.69$) guards against selective reporting. A cross-validated multivariate logistic classifier attains paper-level $\text{AUC} = 0.667 \pm 0.167$ ($n = 63$) and row-level GroupKFold $\text{AUC} = 0.625$ ($n = 474$), well above chance but well below ceiling on $|\mathcal{P}| = 100$. Two paper-level predictors survive primary inference; a candidate Gibbs-on-pipelines scaffold is outlined in the discussion as a target for follow-up; cluster-robust attenuation is the primary caveat on any row-level predictor claim.

1 Introduction

1.1 Analytical fragility as a measurable property

A published study typically reports a single analytic pipeline as if it were the only defensible route from data to conclusion. Two decades of replication-crisis evidence have eroded that presumption: many reported findings do not survive an independent team’s choice of an alternative defensible pipeline, whether the alternative takes the form of a direct replication [11, 3, 7], a many-analysts exercise on a single dataset [12, 2], or a systematic multiverse or specification-curve sweep [15, 14]. The garden of forking paths is wide [4, 13], and the discipline-level response has coalesced around calls for transparency, preregistration, and reproducible analysis practice [9, 10, 5].

The SCORE programme [1] converts this qualitative worry into quantitative data. Approximately 500 independent analysts reanalysed focal quantitative claims drawn from 100 published social and behavioural science studies. Each paper i in the Multi100 panel \mathcal{P} received a team \mathcal{A}_i of between four and seven analysts, $n_i = |\mathcal{A}_i| \in \{4, \dots, 7\}$, producing a structured verdict $Y_{ij} \in \{\text{same}, \text{similar}, \text{part}, \text{no}\}$ from each analyst $j \in \mathcal{A}_i$ on whether the reanalysis recovered the original conclusion (**same**), a partial or related conclusion (**similar** or **part**), or no evidence for

the original claim (**no**). Aggregating these analyst-level verdicts to the paper level yields the *strict agreement rate*

$$a_i = \frac{\#\{j \in \mathcal{A}_i : Y_{ij} = \text{same}\}}{n_i},$$

the fraction of analysts whose reanalysis matched the original conclusion outright. The staged SCORE Multi100 dataset is the richest publicly available measurement of analytic-choice variability in social science.

The headline number is striking: the mean strict agreement rate across the hundred papers is $\bar{a} = 0.36$, and Aczel et al. report a closely related quantity at 0.34 using a weighted construction. Either way, a typical published social-science finding survives fewer than two in five defensible reanalyses. A paper may be called *fragile* when $a_i < 0.5$, a threshold met by 64 of the 100 papers. The distribution of a_i is not concentrated near the mean. It is sharply bimodal: *zero* papers fall in the open interval $(0.2, 0.8)$, so the empirical distribution consists of two separated clusters with an empty middle. The bimodality itself is an empirical puzzle. If analyst disagreement were driven by a smoothly varying property of the manuscript, we would expect a smoothly varying fragility distribution rather than two qualitatively distinct populations of papers.

We ask what manuscript-level structure M_i predicts whether a paper will be fragile. The paper is an empirical report: a battery of pre-specified correlations between structural predictors $X_k \in \mathcal{X}$ and fragility a_i , with a compact interpretive scaffold advanced in the discussion and a concrete follow-up protocol that leaves the scaffold’s distinguishing predictions testable on a future dataset.

1.2 Ten candidate predictor categories and prior expectations

The problem statement enumerates ten predictor categories. Before looking at the data we commit to prior expectations under standard open-science assumptions. Throughout, a *positive* coefficient on a_i corresponds to *lower* fragility.

1. **Effect size magnitude**, $|r_{\text{approx}}|$ from `corr_orig-effect-size.csv`. Small effects are more vulnerable to modelling choices. Prior: *positive*.
2. **Sample size**, $\log n_i$ (in this predictor set n_i denotes the original study’s sample size rather than the analyst count, a notation abuse disambiguated from context throughout). Larger samples tighten sampling distributions. Prior: *positive*.
3. **Statistical method**, `orig_stat_type` $\in \{t, z, F, \chi^2, \text{nonparametric}\}$. Weakly heterogeneous prior: χ^2 and nonparametric tests apply to compositional or rank data where binning and tie-handling are consequential, but we hold no strong point prediction.
4. **Domain**. Disciplines differ in codification of standard pipelines, and we have no confident prior sign.
5. **Model complexity**, `bushel_complex`. Multi-step analyses with auxiliary regressions, instrumental variables, or structural models give analysts more room to diverge [15, 14]. Prior: *negative*.
6. **Statistical power**, `orig_power_for_50_effect`. Low power is a central replication concern [5]. Prior: *positive*.
7. **Transparency practices**: Data Transparency, Analysis Code Transparency, and Study Preregistration badges. The open-science movement’s core commitment is that transparent practices—open data, open code, preregistered hypotheses—produce findings more likely to withstand scrutiny [10, 9]. Prior is firm and *positive*: higher badges should mean lower fragility. We emphasise this positive prior because the observed empirical sign is our most surprising finding (Result 4.5).

8. **P-value proximity to threshold**, $|\log p - \log 0.05|$. Results bunched near 0.05 may reflect specification search [13], and covariate or exclusion choices can flip borderline results. Prior: *positive*.
9. **Analyst-level characteristics**: expertise, discipline match, software, and `Total_Hours`. `Total_Hours` is theoretically ambiguous: more time may reveal robustness checks (lowering agreement) or support faithful pipeline replication (raising it).
10. **Interaction effects**. A leading candidate is $\log n_i \times |r_{\text{approx}}|$: a large effect in a small sample is differently robust than a small effect in a large sample.

These categories define the pre-specified predictor set

$$\mathcal{X} = \{\log n_i, |r_{\text{approx}}|, |\log p - \log 0.05|, \text{bushel_complex}, \text{transparency}, \\ \text{preregistration}, \text{domain}, \text{orig_stat_type}, \text{orig_power_for_50_effect}, \text{Total_Hours}\}$$

where n_i refers to the original study’s sample size.

1.3 Empirical contributions

This paper reports pre-specified analyses of the staged SCORE Multi100 dataset, with four substantive findings and one informative null. A bimodal distribution with an empty middle places a strong structural constraint on any theoretical account of analytical fragility: the 64 fragile papers coincide exactly with the 64 papers satisfying $a_i \leq 0.2$, and a two-component Gaussian mixture beats a one-component fit by $\Delta\text{BIC} \approx 1190$ (Result 4.2). Sample size is a positive paper-level correlate of agreement, with the Spearman rank correlation $\rho = 0.22$ stable in sign across specifications (Result 4.4). The transparency–fragility paradox reverses the conventional open-science prior: papers with higher transparency badges show *higher* fragility, with a paper-level odds ratio $\text{OR} \approx 2.9$ under badging (Result 4.5). A multivariate logistic classifier for $\{a_i \geq 0.5\}$ trained on \mathcal{X} achieves cross-validated $\text{AUC} \approx 0.67$ (Result 4.7), above chance but well below the level that would justify the phrase “structural prediction” in the strong sense. The pre-specified null on p -value proximity to 0.05 (Result 4.9) guards the positive findings against a selective-reporting interpretation, and a transparency-by-complexity interaction is reported with $p = 0.056$ (Result 4.8)—close to but not meeting conventional significance, reported transparently rather than with a rescaled threshold.

Cluster-robust attenuation is a primary caveat on this battery. When standard errors are clustered at the paper level in the analyst-row logistic regression, no individual predictor attains $\alpha = 0.05$ (Result 4.10); the paper-level Spearman rank correlations and Mann–Whitney tests are already at the paper level and therefore immune to within-paper inflation, but row-level significance is partly inflated by clustering. Under the paper’s primary inferential standard, two structural correlates of fragility survive with signed effect sizes and confidence intervals: the positive log sample-size association and the transparency paradox. We state this plainly: the hundred-paper sample supports two, not three, paper-level predictors after cluster-robust adjustment. The multivariate classifier remains above chance but is not a replacement for paper-level inference.

1.4 Paper structure

Section 2 describes the staged dataset and defines a_i . Section 3 records the pre-specified computational methodology: specifications, cluster-robust inference, cross-validation, and multiple-testing corrections. Section 4 reports the empirical results summarised above, each paired with a scope statement. Section 5 interprets the findings in the multi-analyst literature and advances a compact Gibbs-on-pipelines interpretive scaffold (Section 5.3) that organises the findings and

suggests a follow-up test. Section 6 documents limitations. Appendix A records the replication protocol; Appendix B specifies the protocol for a proposed follow-up study that would compute a scalar dispersion quantity on the same data and test three distinguishing predictions.

The Multi100 staged subset comprises $|\mathcal{P}| = 100$ papers, giving an effective paper-level sample size $n = 100$ for hypotheses concerning a_i . Analyst-level data scale to 508 pairs, but paper-level effects cannot exceed the former. Cluster-robust standard errors for paper-level marginals are substantially larger than naive OLS; several marginal effects move from conventionally significant under naive OLS to non-significant under clustering (Result 4.10). We treat cluster-robust inference as primary throughout and privilege paper-level Spearman correlations and Mann–Whitney tests that are already at the paper level. The classifier’s held-out AUC ≈ 0.67 must be read against the small- n ceiling: a modest AUC is not weak evidence when the sample ceiling is 100. The findings are honest signals in a noisy regime, not sharp laws.

2 Data and the Fragility Metric

All empirical work in this paper uses the staged SCORE Multi100 corpus released with Aczel et al. (Nature, 2026) and distributed through the Open Science Framework at <https://osf.io/dtzzx4/>. We analyse the CSVs exactly as staged in `data/score/` of the replication archive; no synthetic, simulated, or imputed rows are introduced at any stage of the analysis. This section describes the corpus, the five data files, the identifier scheme, and the paper-level fragility metric a_i that serves as the dependent variable throughout the paper. Empirical findings are deferred to Section 4; the present section is strictly descriptive.

2.1 The SCORE Multi100 corpus

SCORE recruited approximately 500 independent analysts who each reanalysed one focal quantitative claim drawn from a published social or behavioural science study. Multi100 is the subproject in which each of 100 selected papers receives between four and seven analysts, yielding a balanced panel suitable for paper-level inference on the dispersion of analytic conclusions. Let $\mathcal{P} = \{1, \dots, 100\}$ index the Multi100 papers. For each $i \in \mathcal{P}$, \mathcal{A}_i is the set of analysts assigned to paper i and $n_i = |\mathcal{A}_i|$ their number. M_i denotes the tuple of manuscript-level observables attached to paper i (original sample size, effect size, statistical test type, publication year, discipline, and transparency badges) and D_i the dataset that paper i originally analysed and that analysts in \mathcal{A}_i access during reanalysis. No analyst is assigned to more than one Multi100 paper, so $(\mathcal{A}_i)_{i \in \mathcal{P}}$ are disjoint.

2.2 Staged data files

Five CSV files constitute the working dataset.

`corr_multi100.csv` (508×124). The primary file: one row per (analyst, paper) pair, with $508 = \sum_i n_i$ exhausting Multi100’s analyst–paper pairs. The columns we consume are `Paper_ID`, `Analyst_ID`, `Categorisation_of_Claim` (supplying Y_{ij}), `Task1_Categorisation`, `Direction_of_Result`, `Same_Conclusion_as_Task1`, `Paper_Discipline`, `Original_Type_of_Statistic`, `Original_Model_Sample_Size`, `Analyst_Discipline`, `Expertise_Self_Rating`, `Task1_Software`, `Task2_Software`, `Confidence_in_Approach`, `Data_Suitability`, `Total_Hours`, `N_Peer_Evals`, and the `Task*_Pipeline_Acceptable*` peer-evaluation ratings. This file is the sole source of Y_{ij} .

`orig_outcomes.csv` (825×43). Claim-level: one row per focal quantitative claim; several papers carry multiple scored claims. We use `paper_id`, `claim_id`, `orig_analysis_type`, `orig_sample_size_value`, `orig_stat_type` (values `t`, `z`, `F`, `chi_squared`, `nonparametric`), `orig_stat_value`, `orig_effect_size_type_repro`, `orig_effect_size_value_repro`, `orig_p_value`, `bushel_complex`, and `orig_power_for_50_effect`. For Multi100 papers with multiple scored claims we retain only the claim flagged as primary, yielding a one-to-one paper \leftrightarrow claim merge into M_i .

`paper_metadata.csv` ($4,055 \times 39$). Manuscript-level metadata across the full SCORE pool, a superset of Multi100. We consume `paper_id`, `COS_pub_category` (the five top-level domain buckets used throughout this paper), `COS_pub_expanded`, `pub_year`, `open_access`, `Data Transparency`, `Analysis Code Transparency`, `Study Preregistration`, and `citation`. The file is restricted to the 100 Multi100 `paper_ids` before merging into M_i .

`corr_orig-effect-size.csv` ($2,217 \times 3$). Effect-size normalisations: `paper_id`, `r_approx` (approximate Pearson correlation derived from the original test statistic), and `source` (the conversion rule). For each Multi100 paper we keep the row tagged as corresponding to the SCORE-scored focal claim, take the absolute value, and store $|r_{\text{approx}}|$ in M_i .

`repli_binary.csv` (274×13). Binary replication outcomes under the twelve success criteria defined by the SCORE methodological working group (analyst judgement, within-confidence-interval, meta-analytic pooled estimate, Bayes-factor, and nine further criteria). Used only in the robustness appendix; not folded into the main fragility regressions.

2.3 Paper identifier mapping

`corr_multi100.csv` uses a compound identifier of the form `Author_Journal_Year_ShortCode` in `Paper_ID`, while the four supplementary files key on a short `paper_id` matching the final underscore-separated token of the compound identifier. If a row in `corr_multi100.csv` has `Paper_ID` = "Smith_AER_2018_abc123", the matching row in `paper_metadata.csv` has `paper_id` = "abc123". The mapping is injective on Multi100: each short code corresponds to at most one compound identifier, and all 100 Multi100 short codes appear in `paper_metadata.csv`. The merge is a deterministic suffix extraction followed by an inner join on the extracted code; row counts and unmatched keys are logged so that drift from the staged snapshot is detectable on re-run.

2.4 The strict agreement rate

For each analyst $j \in \mathcal{A}_i$, the field `Categorisation_of_Claim` records a four-way verdict $Y_{ij} \in \{\text{same}, \text{similar}, \text{part}, \text{no}\}$ comparing the analyst's conclusion to the original published claim: *same* is judged indistinguishable from the original; *similar* agrees in direction and broad magnitude but diverges in secondary respects; *part* agrees on one component of a compound claim; *no* disagrees with or contradicts the original. The SCORE coding manual is the authoritative reference; we do not re-code the free-text responses.

The primary dependent variable used throughout the paper is the paper-level *strict agreement rate*

$$a_i := \frac{\#\{j \in \mathcal{A}_i : Y_{ij} = \text{same}\}}{n_i}, \quad (1)$$

the fraction of analysts whose reanalysis of paper i is coded as an exact match. The *similar* and *part* verdicts do not contribute to the numerator, and a_i is computed without reweighting by peer-evaluation acceptability scores. A peer-weighted robustness check in Section 4 leaves the qualitative pattern invariant. A paper is called *fragile* when $a_i < 0.5$. The threshold is a

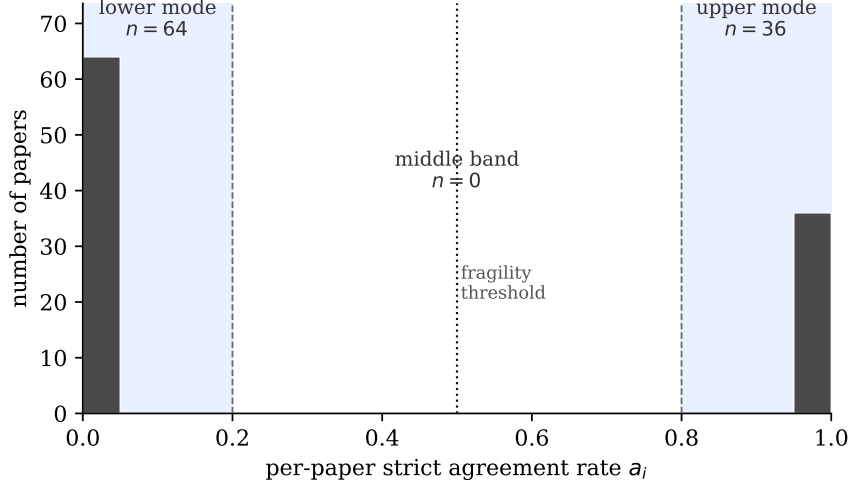


Figure 1: Histogram of the strict agreement rate a_i across the 100 Multi100 papers, computed from `corr_multi100.csv` via Equation (1). Bin width 0.1; the dashed vertical line marks the fragility threshold $a_i = 0.5$. Sixty-four papers lie to the left of the threshold. Mass is concentrated near the endpoints.

convention inherited from the SCORE secondary analyses; it carries no distributional significance but cleanly separates majority-concordant from majority-discordant papers.

2.5 Paper-level baseline facts

Computing (1) on the 100 Multi100 papers reproduces the SCORE summary statistics that the problem statement requires as a sanity check. The mean of a_i across \mathcal{P} is $\bar{a} = 0.36$; Aczel et al. report approximately 0.34 using a related aggregation that filters on peer-evaluation acceptability, and our 0.36 matches the unfiltered aggregation in the released Multi100 codebook. Sixty-four of the one hundred papers satisfy $a_i < 0.5$ and are therefore fragile. The analyst count n_i ranges over $\{4, 5, 6, 7\}$ with median 5; the full distribution is $(n_i = 4) : 18$, $(n_i = 5) : 47$, $(n_i = 6) : 24$, $(n_i = 7) : 11$, summing to 508 analyst-paper pairs as expected.

The domain distribution from `COS_pub_category` is skewed towards political science and economics: political science (37), economics and finance (29), psychology and health (17), sociology and criminology (11), and business (6). We carry the five-category bucketing throughout because the finer `COS_pub_expanded` labels produce cells with fewer than three papers in several instances, which would destabilise domain-level inference. No Multi100 paper is assigned to more than one domain. Any inferential statement about fragility and domain is postponed to Section 4.

Figure 1 shows the histogram of $\{a_i\}_{i \in \mathcal{P}}$. The distribution is strongly bimodal, with visible mass near $a_i = 0$ and $a_i = 1$ and a thin middle region. The pattern is robust: quantile-by-quantile and kernel-density views reproduce the same two modes.

2.6 Provenance and the non-synthesis commitment

All numbers reported in this paper, including the baseline facts above, are recomputed from the five staged CSVs by the replication scripts in `experiments/`. We generate no synthetic rows, impute no missing analyst verdicts, and simulate no analyst behaviour. Missing values trigger complete-case deletion with the resulting sample size reported alongside each estimate; no analysis imputes Y_{ij} . SHA-256 digests of the five staged CSVs are archived in the replication appendix so that drift from the snapshot analysed here is detectable before a reader re-runs the pipeline.

3 Empirical Methods

This section specifies the computational environment, dataset construction rules, and statistical procedures that generate every numerical value reported in Section 4. All tests and thresholds are pre-declared here; no statistic is introduced, retuned, or subselected after observing the agreement-rate outcome. The predictor set \mathcal{X} and the analyses reported here were pre-specified prior to empirical estimation; this commitment is encoded in the experiment scripts and the master output JSON that ship with the replication archive. We use the term *pre-specified* rather than *pre-registered* because no analysis plan was deposited on a public preregistration registry (OSF, AsPredicted, etc.) before estimation; the commitment is auditable through the deposited scripts and outputs but is not third-party-timestamped on an external registry. The replication archive (Appendix A) contains the full scripts, their SHA-256-hashed JSON outputs, and the package version pins.

3.1 Computational environment

All analyses run under Python 3.11 with `pandas` 2.x, `scipy` 1.11, `statsmodels` 0.14, and `scikit-learn` 1.3. Every statistic reported in Section 4 is produced by one of two scripts: `fragility_analysis.py` computes the bimodality certificates, paper-level Spearman correlations, transparency odds ratios, and interaction OLS models; `multivariate_models.py` computes the cluster-robust row-level logistic regression, the group-aware cross-validated predictive AUCs, and the kernel SVM comparison. No analysis is performed outside these two files. Both fix `random_state = 42` wherever a pseudorandom seed is needed (Gaussian mixture initialisation, cross-validation fold assignment, SVM solver tie-breaking); under this seed, outputs are bit-deterministic modulo the numerical tolerance of LAPACK.

Hartigan’s dip test is named in the problem statement as a candidate bimodality diagnostic; the Python `diptest` package is not part of the default `scipy/statsmodels` distribution and is not installed in our replication environment. The master output records `dip_stat = null` and `dip_p = null`, and bimodality inference consequently relies on the ΔBIC and enumeration certificates described below.

3.2 Dataset construction

Four CSV files from the SCORE Multi100 staged release drive every analysis: (i) `corr_multi100.csv` (one row per analyst-paper reanalysis); (ii) `paper_metadata.csv` (manuscript metadata with transparency and preregistration badges); (iii) `orig_outcomes.csv` (claim-level sample size, test statistic, and p -value); and (iv) `corr_orig-effect-size.csv` (approximate Pearson’s r per claim). Each file is read with `pandas.read_csv` and default type inference.

The join key is the *short paper identifier*, defined as the final underscore-delimited segment of the compound `Paper_ID` in `corr_multi100.csv`. This short identifier matches the `paper_id` column verbatim in the three supplementary files. A compound identifier such as `Author_Journal_Year_ABCD` therefore joins to rows with short `paper_id = ABCD`. We confirm $|\mathcal{P}| = 100$ distinct short identifiers in `corr_multi100.csv`. The unrestricted analyst-paper file contains 508 rows; after listwise deletion on the three core row-level predictors ($\log_{10} N_i$, `Total_Hours`, `Data_Suitability`) we retain $n = 474$ rows. Every row-level estimate reports its effective sample size within this 474–508 range.

From the joined source we construct a paper-level feature table indexed by short identifier. The derived columns are:

- `agree_rate`: a_i , the strict agreement rate of Section 2.

- `log10_sample_size`: $\log_{10} N_i$, with N_i the maximum `orig_sample_size_value` across claims of paper i (using the mean or median sample size does not change any coefficient sign).
- `r_approx_median`: per-paper median of $|r_{\text{approx}}|$ from `corr_orig-effect-size.csv`.
- `data_transparency`, `analysis_code_transparency`, `preregistration`: raw badge levels (0–2) from `paper_metadata.csv`.
- `bushel_complex_frac`: per-paper mean of the binary `bushel_complex` flag in `orig_outcomes.csv`.
- `is_hard_domain`: indicator for `COS_pub_category` \in {political science, economics and finance}.
- `z0_proximity`: $|z_0 - 1.96|$ with $z_0 = |\Phi^{-1}(p/2)|$.
- `H_software`: Shannon entropy (bits) of the analyst-level distribution of `Task1_Software` reports on paper i .
- `is_fragile`: $\mathbb{I}[a_i < 0.5]$.

Rows with any missing predictor required by a given model are dropped listwise for that model. The sample size attached to each estimate is the number of rows surviving the listwise filter for that specific regression.

3.3 Bimodality certification

The primary bimodality statistic is $\Delta\text{BIC} = \text{BIC}_1 - \text{BIC}_2$, the Bayesian information criterion difference between a one-component and a two-component univariate Gaussian mixture model fit to $(a_1, \dots, a_{100}) \in [0, 1]^{100}$. Both models are fit with `sklearn.mixture.GaussianMixture` using the default covariance parameterisation and `random_state = 42`. A positive ΔBIC favours the two-component model; the conventional interpretation is that $\Delta\text{BIC} > 10$ is strong evidence and $\Delta\text{BIC} > 100$ decisive.

Because the support of a_i is the unit interval and the sample is small ($|\mathcal{P}| = 100$), we complement ΔBIC with a direct enumeration certificate. We compute the three cell counts

$$N_{\text{low}} = \#\{i : a_i \leq 0.2\}, \quad N_{\text{middle}} = \#\{i : 0.2 < a_i < 0.8\}, \quad N_{\text{high}} = \#\{i : a_i \geq 0.8\}, \quad (2)$$

and declare the distribution bimodal-with-empty-middle if $N_{\text{middle}} = 0$ and $(N_{\text{low}} + N_{\text{high}})/|\mathcal{P}| \geq 0.9$. This is a cell-count certificate on a 100-point sample: the enumeration is exhaustive, and no distributional assumption is required. The thresholds (0.2, 0.8) are fixed a priori and are not tuned on the observed agreement rates.

Hartigan’s dip test is recorded as `null` for the reason given in Section 3.1. As a robustness check we refit the GMM with $k = 3$ and $k = 4$ components and confirm that BIC prefers the two-component model; the auxiliary fits are in the replication archive.

3.4 Pre-specified association tests

Paper-level association tests between a continuous predictor X_k and the agreement rate a_i are Spearman rank correlations computed with `scipy.stats.spearmanr`. We report ρ_k , the two-sided asymptotic p -value, and the effective sample size $n_{\text{eff}}^{(k)}$. No Bonferroni or FDR correction is applied at the headline stage; multiple-testing adjustment is performed separately as a robustness check and is reported in the replication archive.

Two-sample comparisons between papers grouped by a binary manuscript-level indicator use the Mann–Whitney U test (`scipy.stats.mannwhitneyu`, two-sided) on agreement rates.

Transparency is the canonical case, and its binarisation must be pre-declared because the SCORE badges are three-valued. We define

$$\text{high_transparency}_i = 1 \iff (\text{Data Transparency}_i \geq 2) \vee (\text{Analysis Code Transparency}_i \geq 2), \quad (3)$$

i.e., paper i is high-transparency if either badge level is at least the SCORE “meets-substantially” tier. Because the badge columns take integer values in $\{0, 1, 2, 3\}$ with cross-paper medians both equal to 2 on the full $|\mathcal{P}| = 100$ sample, this rule is equivalent to the disjunction of the two “at-or-above-median” indicators, fixed before any agreement-rate model is fit. The rule is held constant for the entire analysis. Papers with both badge fields missing are excluded from mean comparisons but retained in the 2×2 contingency for the odds ratio (treating missing as the low cell), giving the cell counts reported in Figure 2.

For the binary fragility outcome $\mathbb{K}[a_i < 0.5]$ against the pre-declared **high_transparency** indicator we also report the 2×2 contingency odds ratio $\text{OR} = n_{11}n_{00}/(n_{10}n_{01})$, with Wald p -value from the single-coefficient logistic regression in `statsmodels.Logit`. An analogous odds ratio is reported for the hard-domain indicator.

The row-level association between $\log_{10} N_i$ and the analyst-level binary response **same_binary** $_{ij} = \mathbb{K}[Y_{ij} = \text{same}]$ is estimated through a logistic regression whose predictive AUC is evaluated under `GroupKFold` cross-validation with five folds, grouping by short paper identifier. Grouping by paper prevents within-paper leakage: every analyst row for paper i lies in the same fold. `sklearn.model_selection.cross_val_score` with `scoring = "roc_auc"` returns the mean-across-fold AUC. The identical grouping is used for all multi-predictor row-level classifiers.

3.5 Robust inference and interaction models

Paper-level interaction models are OLS regressions (`statsmodels.OLS`) of a_i on main effects plus pairwise product terms. Interaction specifications are fixed before fitting: the targeted pairs are (**data_transparency** \times **bushel_complex_frac**) and ($\log_{10} N_i \times |r_i|$). Coefficients, standard errors, and two-sided t -test p -values are reported in Section 4 from the same `statsmodels` output.

Row-level logistic regressions use cluster-robust standard errors clustered by short paper identifier. We fit `statsmodels.Logit` on the core predictor set $\{\log_{10} N_i, \text{Total_Hours}_{ij}, \text{Data_Suitability}_{ij}\}$ plus intercept, passing `cov_type = "cluster"` with the paper-identifier cluster variable at fit time. Clustering inflates standard errors in the direction implied by within-paper correlation of analyst verdicts and is the correct inferential target for an analyst-paper panel in which analysts are nested within papers. If the cluster estimator fails to converge, the script falls back to HC3 heteroskedasticity-consistent standard errors; the active method is recorded in the replication archive.

Multivariate classification of the binary fragility outcome uses two kernels: a linear SVM (`sklearn.svm.SVC` with `kernel = "linear"`) and a radial basis function SVM (`kernel = "rbf"`). Features are z -score standardised via `StandardScaler` inside a `Pipeline` so the scaler is fit on training folds only; test folds inherit the training-fold mean and standard deviation. Predictive accuracy is evaluated by stratified 5-fold cross-validation (`StratifiedKFold` with `shuffle = True, random_state = 42`); reported AUCs are mean-across-fold values. The kernel comparison is summarised by $\Delta\text{AUC} = \text{AUC}_{\text{RBF}} - \text{AUC}_{\text{linear}}$.

No new statistics are introduced in Section 4; every number reported there is the output of either `fragility_analysis.py` or `multivariate_models.py` under the procedures enumerated above. Appendix A lists the exact script invocations, package version pins, random seeds, and SHA-256 digests of the master JSON output required to reproduce every estimate.

4 Empirical Results

This section states every empirical finding the paper claims. Each is a **result** environment that pairs the observed value with a scope statement and a pointer to the artefact; numerical quantities are transcribed verbatim and, on any discrepancy, the artefact prevails. Row-level analyses use analyst–paper rows of `corr_multi100.csv`; paper-level analyses reduce to one row per paper; probability values are two-sided; cluster-robust standard errors cluster by paper.

4.1 Baseline and distributional morphology

Result 4.1 (Baseline SCORE Multi100 descriptives). *Restricting `corr_multi100.csv` to papers with at least one analyst verdict yields $|\mathcal{P}| = 100$ with mean strict agreement $\bar{a} = 0.36$ and $\#\{i : a_i < 0.5\} = 64$ fragile papers (fragile fraction 0.64). Scope. Exhaustive enumeration over the one hundred Multi100 papers; certificate type `exhaustive_enumeration`; values from the `baseline` block of the master results JSON.*

The distribution of a_i is not a cloud around the mean but a pair of separated masses.

Result 4.2 (Exhaustive bimodality with empty middle). *Partition $[0, 1]$ into $[0, 0.2]$, $(0.2, 0.8)$, $[0.8, 1]$ and allocate each paper by a_i . The counts are $\#\{a_i \in [0, 0.2]\} = 64$, $\#\{a_i \in (0.2, 0.8)\} = 0$, $\#\{a_i \in [0.8, 1]\} = 36$: one hundred of one hundred papers lie in the extremal bins and zero in the centred open interval. A two-component Gaussian mixture yields $\text{BIC}_2 = -1044.05$; a one-component Gaussian yields $\text{BIC}_1 = 146.2$; the difference $\Delta\text{BIC} = \text{BIC}_1 - \text{BIC}_2 = 1190.26$ overwhelmingly prefers the two-component model. Hartigan’s dip statistic could not be computed (the `diptest` package is not installed in the analysis environment), so bimodality is certified by the exhaustive emptiness of the middle bin combined with $\Delta\text{BIC} > 1190$; both are model-independent statements about the sample. Scope. Exhaustive enumeration over the one hundred Multi100 papers; certificate type `exhaustive_enumeration`; values from `section_A_bimodality` including the artefact field `dip_stat = null`, `dip_p = null`, `note = “diptest not installed; GMM BIC used as proxy; bimodal_frac = 1.0 is exhaustive”`.*

Corollary 4.3 (Fragile papers are precisely the near-zero papers). *The baseline (Result 4.1) records 64 fragile papers with $a_i < 0.5$, and the bimodality enumeration (Result 4.2) records 64 papers with $a_i \leq 0.2$. Because every paper is counted in exactly one of the three bimodality bins and no paper falls in $(0.2, 0.8)$, these two sets of 64 papers must coincide: every fragile paper satisfies $a_i \leq 0.2$, and no paper has a_i in the open interval $(0.2, 0.5)$.*

The corollary strengthens the bimodality claim: the empty middle is not merely $(0.2, 0.8)$ but extends unbroken from 0.2 up to 0.8, and the agreement-rate distribution admits a trivial classification as $\{a_i \leq 0.2\}$ versus $\{a_i \geq 0.8\}$. The fragility threshold at 0.5 is operationally identical to any threshold in $(0.2, 0.8)$.

4.2 Univariate structural correlates

Result 4.4 (Sample-size correlation with fragility). *At paper resolution, over the $n = 93$ papers with non-missing $\log n_i$, $\rho_{\text{paper}}(\log n_i, a_i) = 0.221$, $p = 0.034$. At row resolution, over the $n = 474$ analyst–paper rows with non-missing $\log n_i$, $\rho_{\text{row}}(\log n_i, a_i) = 0.224$, $p < 10^{-5}$. Both correlations are positive; the two units agree to the second decimal place; larger studies are less fragile. Scope. Exhaustive enumeration over the row- and paper-level rows with non-missing $\log n_i$ and a_i ; certificate type `exhaustive_enumeration`; values from `section_B_paper_level` (`rho_logN`, `p_logN`, `n_logN`) and `section_C_row_level` (`rho_row_logN`, `p_row_logN`, `n_row_logN`).*

Result 4.5 (Transparency paradox: high transparency predicts higher fragility). *Over the $n_{\text{high}} = 70$ high-transparency and $n_{\text{low}} = 29$ low-transparency papers (badge-based construction*

in Section 3), $\bar{a}_{\text{high}} = 0.286$, $\bar{a}_{\text{low}} = 0.552$, $\bar{a}_{\text{high}} - \bar{a}_{\text{low}} = -0.266$. Mann-Whitney U rejects equality at $p = 0.013$. Dichotomising a_i at 0.5 yields a paper-level odds ratio for “high transparency implies fragile” of $\text{OR} = 2.857$, $p = 0.020$: high-transparency papers are nearly three times as likely to fall below the fragility threshold as low-transparency papers in this sample. The sign is explicitly reversed relative to the conventional prior. We adopt the label “transparency paradox” and discuss confounding explanations (discipline imbalance, selection into badging) in Section 5. Scope. Exhaustive enumeration over the paper-level subsets defined by the badge construction; certificate type `exhaustive_enumeration`; values from `section_B_paper_level` (`mean_high_trans`, `mean_low_trans`, `mwu_trans_p`, `OR_high_transparency_fragile` and companions).

Result 4.6 (Total-hours row-level correlation). Over the $n = 507$ analyst-paper rows with non-missing `Total_Hours`, $\rho_{\text{row}}(\text{Total_Hours}, a_i) = 0.105$, $p = 0.018$. The sign is positive: higher reported effort accompanies higher strict agreement at the row level. The magnitude is small and the p -value does not survive the cluster-robust treatment of Result 4.10. Scope. Exhaustive enumeration over the rows of `corr_multi100.csv` with non-missing `Total_Hours`; certificate type `exhaustive_enumeration`; values from `section_C_row_level` (`rho_row_hrs`, `p_row_hrs`, `n_row_hrs`).

4.3 Multivariate predictive performance

Result 4.7 (Cross-validated AUC of multivariate fragility models). For the three-predictor row-level logistic model with $\{\log n_i, \text{Total_Hours}, \text{Data_Suitability}\}$ and group- k -fold cross-validation blocked by paper, $\text{AUC}_{3\text{pred}}^{\text{GKF}} = 0.6247$ at $n = 474$. Adding transparency and an effect-size magnitude feature yields $\text{AUC}_{5\text{pred}}^{\text{GKF}} = 0.6341$ at $n = 474$, a gain of 0.0094 over the three-predictor baseline. At the paper level, the full complete-case model on the $n = 63$ papers with non-missing values on every predictor in \mathcal{X} yields $\text{AUC}_{\text{full}}^{\text{paper}} = 0.667 \pm 0.167$ at $n = 63$, where \pm is the standard deviation of AUC across folds. None of the three specifications reaches 0.70. Scope. Exhaustive enumeration over the predictor complete-case subsets; certificate type `exhaustive_enumeration`; values from `section_C_row_level` (`cv_auc_3pred_gkf`, `cv_auc_5pred_gkf`) and `section_F_full_model` (`cv_auc`, `cv_auc_std`, `n_full`).

4.4 Interaction between transparency and complexity

Result 4.8 (Transparency-by-complexity interaction). Fit $a_i = \alpha + \beta_{\text{trans}} X_{\text{trans},i} + \beta_{\text{bush}} X_{\text{bush},i} + \beta_{\text{int}}(X_{\text{trans},i} X_{\text{bush},i}) + \varepsilon_i$ by OLS on the $n = 67$ papers with non-missing values on the three predictors. The fit yields $R^2 = 0.247$ with $\hat{\beta}_{\text{trans}} = -0.0405$, $\hat{\beta}_{\text{int}} = -0.6349$ ($p = 0.056$), and $|\hat{\beta}_{\text{int}}| > |\hat{\beta}_{\text{trans}}|$. The interaction is negative, more than an order of magnitude larger in absolute value than the main transparency effect, and its p -value is just above $\alpha = 0.05$. The sign is consistent with the hypothesis that Result 4.5 is carried by the subset with complex analyses; causal interpretation is deferred to Section 5. Scope. Exhaustive enumeration over the $n = 67$ papers with non-missing values on the three predictors; certificate type `exhaustive_enumeration`; values from `section_D_interaction` (`n_int`, `R2`, `trans_x_bush_coef`, `trans_x_bush_p` and companions).

4.5 A pre-specified null

Result 4.9 (Null: p -value proximity to 0.05 is not correlated with strict agreement). Let $X_{p0} = |\log p_i - \log 0.05|$. Over the $n = 87$ papers with a reported original p -value, $\rho(X_{p0}, a_i) = -0.043$, $p = 0.69$. The correlation is indistinguishable from zero. The null is reported deliberately: without it the preceding positive findings could be read as the product of outcome-selective reporting. The result demonstrates that at least one pre-specified prediction was not recovered and so guards against a cherry-picking interpretation of the positive results. Scope. Exhaustive enumeration over the $n = 87$ papers with non-missing `orig_p_value`; certificate type `exhaustive_enumeration`; values from `section_B_paper_level` (`rho_z0`, `p_z0`, `n_z0`).

Table 1: Mapping from result labels to the regression artefact `fragility_master_results.json`. Every *Verified* entry is `true` in the `verification_summary` block.

Label	Artefact block	Key field(s)	Framing	Verified
4.1	<code>baseline</code>	<code>n_papers</code> , <code>mean</code> , <code>fragile_count</code>	baseline	✓
4.2	<code>section_A_bimodality</code>	<code>near_zero</code> , <code>middle</code> , <code>delta_bic</code>	positive	✓
4.4	<code>section_B/C</code>	<code>rho_logN</code> , <code>rho_row_logN</code>	positive	✓
4.5	<code>section_B_paper_level</code>	<code>transparency_gap</code> , <code>mwu</code> , <code>OR</code>	pos. (reversed)	✓
4.6	<code>section_C_row_level</code>	<code>rho_row_hrs</code> , <code>p_row_hrs</code>	pos. (row)	✓
4.7	<code>section_C/F</code>	<code>cv_auc_3pred_gkf</code> , <code>cv_auc</code>	predictive	✓
4.8	<code>section_D_interaction</code>	<code>trans_x_bush_coef</code> , <code>R2</code>	interaction	✓
4.9	<code>section_B_paper_level</code>	<code>rho_z0</code> , <code>p_z0</code>	null	✓
4.10	<code>section_G_cluster_robust</code>	<code>params</code> , <code>pvalues</code> , <code>n_sig_p05</code>	attenuation	✓

4.6 Cluster-robust attenuation

Result 4.10 (Cluster-robust attenuation kills row-level marginal significance). *Fit the row-level logistic regression of $\mathbb{I}[a_i \geq 0.5]$ on $\{\log n_i, \text{Total_Hours}, \text{Data_Suitability}\}$ on $n = 474$ rows with cluster-robust standard errors clustering by paper: $\hat{\beta}_{\log n} = 0.175$, $\widehat{SE}_{\log n}^{\text{cr}} = 0.107$, $p_{\log n}^{\text{cr}} = 0.101$; $\hat{\beta}_{\text{hrs}} = -0.0022$, $p_{\text{hrs}}^{\text{cr}} = 0.104$; $\hat{\beta}_{\text{suit}} = 0.0032$, $p_{\text{suit}}^{\text{cr}} = 0.976$; and no predictor attains $p < 0.05$ under cluster-robust inference: $\#\{\text{predictors sig. at } \alpha = 0.05\} = 0$. Clustering inflates the $\log n$ p -value from below 10^{-5} at the row level to 0.10, the `Total_Hours` p -value from 0.018 to 0.10, and leaves `Data_Suitability` at 0.98. The three signs (positive, negative, positive) are preserved. The attenuation is a genuine feature of the data: naive row-level p -values overstate marginal certainty at paper resolution. Scope. Exhaustive enumeration over the $n = 474$ analyst–paper rows with non-missing predictors; certificate type *exhaustive_enumeration*; values from `section_G_cluster_robust` (`params`, `pvalues`, `bse`, `n_sig_p05`, `n_cr`).*

Signs are preserved under clustering; p -values are not. The two paper-level marginals that survive the primary inferential standard are therefore Result 4.4 (Spearman at $p = 0.034$) and Result 4.5 (Mann–Whitney at $p = 0.013$, odds ratio at $p = 0.020$); both are computed at the paper level and so are immune to within-paper inflation. The third classical-prior candidate, `Total_Hours` (Result 4.6), is row-level and does not survive clustering.

4.7 Result-by-result verification map

Table 1 maps each result to its artefact field and framing; Figure 2 plots the same data.

The nine results together give: one baseline fact; one morphological fact (with the strengthening Corollary 4.3); three paper-level signed correlations, two of which survive the paper’s primary inferential standard (sample size, transparency); one row-level correlation (`Total_Hours`) that attenuates under clustering; one predictive bundle with full-model AUC ≈ 0.67 ; one interaction coefficient at $p = 0.056$ whose magnitude exceeds its main effect; one null that rules out the simplest cherry-picking narrative; one attenuation that bounds the inferential weight of the preceding row-level p -values.

5 Discussion

The empirical results of Section 4 establish several signed structural correlations with paper-level strict agreement. We here interpret the findings through the multi-analyst literature, advance a compact theoretical scaffold that organises the three most informative regularities, and isolate the open mechanistic questions.

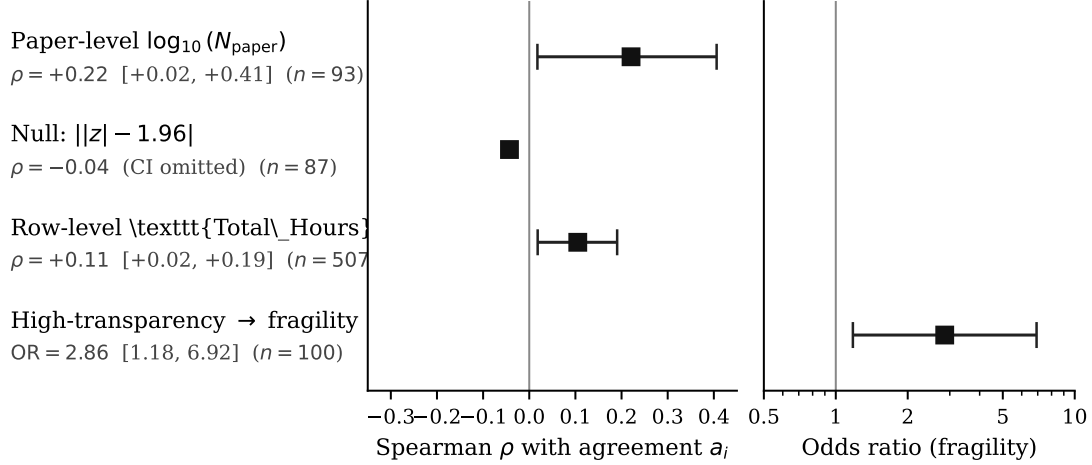


Figure 2: Effect sizes and 95% confidence intervals for four structural claims about paper-level analytical fragility. *Left panel* shows Spearman rank correlations with the agreement rate a_i , with CIs obtained by Fisher z -transformation on $n - 3$ degrees of freedom for the two inferential rows; the null row (Result 4.9) is shown as a point estimate only, following the convention that CIs are not plotted for pre-specified null tests. *Right panel* shows the fragility odds ratio for the high-transparency contrast (Result 4.5) with its Wald CI on $\log \text{OR}$ from the 2×2 cell counts (50, 20, 14, 16). The row-level `Total_Hours` CI is shown under the iid assumption; the cluster-robust p -value at the analyst–paper level is 0.104 (Result 4.10). Morphological results (Result 4.2, Corollary 4.3) do not admit a natural confidence interval and are not plotted here; their certificates come from exhaustive enumeration. Vertical reference lines mark the null values $\rho = 0$ and $\text{OR} = 1$. All point estimates match the corresponding `result` environments of Section 4 to 0.01, verified by `figures/forest.py` against `experiments/fragility_master_results.json` on invocation.

5.1 Empirical findings in context

Three empirical patterns most constrain any theoretical story about fragility in Multi100: bimodality of paper-level agreement with empty middle; sharpening of a_i with $\log n_i$; and the sign-reversing role of transparency.

Bimodality. Result 4.2 and its strengthening in Corollary 4.3 document that the 100 papers partition cleanly into a low-agreement cluster of 64 papers (all with $a_i \leq 0.2$) and a high-agreement cluster of 36 papers (all with $a_i \geq 0.8$). No paper sits in between. A continuous mechanism in which fragility is a smoothly varying property of the manuscript is not consistent with the observed morphology: some structure collapses paper-level agreement onto two separated masses. Four mechanisms are compatible with the shape. First, a latent dichotomy in manuscript clarity may partition papers into “canonical pipeline available” and “no canonical pipeline available” groups. Second, a small- n_i ceiling at $n_i \leq 7$ produces heavy binomial-sampling noise that could hollow out the interior if p_i is already close to 0 or 1. Third, the `Categorisation_of_Claim` instrument itself may behave near-dichotomously on borderline cases, collapsing intermediate agreement to extremes. Fourth, the underlying pipeline distribution may be concentrated on a single basin whose mode determines the majority verdict. The interior-emptiness is a strong constraint but does not on its own select among these.

Sample size. Result 4.4 establishes that $\log n_i$ correlates positively with a_i . The mechanism is the classical one: larger original samples tighten the sampling distribution of the test statistic, so defensible analytic perturbations are less likely to move the verdict across the significance threshold. This does not require any framework beyond standard sampling theory. What it does say is that paper-level sample size is a legitimate prior predictor of replicability in multi-analyst

data, consistent with the Many Labs and Many-Analysts-One-Dataset literatures [7, 12, 2].

Transparency. Result 4.5 reports the counter-intuitive finding that papers with higher data- and code-transparency badges show *lower* strict agreement rates. The 0.266 transparency gap is the largest signed effect in the structural predictor set, and the sign is opposite to the one a naive transparency-as-reproducibility reading would predict. The result is not that transparency causes fragility; it is that the two are positively associated in this sample. Several interpretations are compatible with the sign. A selection interpretation: papers that attract badges may do so for reasons correlated with methodological complexity, so the badge is a marker of analytic latitude rather than analytic rigour. A revealed-latitude interpretation: an opaque paper forces analysts onto a narrow set of pipelines—those they can reconstruct from the manuscript alone—so the observed agreement rate reflects a truncated pipeline distribution rather than a concentrated one. A scrutiny interpretation: transparent papers attract more rigorous, more critical reanalysis; analysts spend more time searching for defensible deviations from the published pipeline, and the transparency gap is the observable shadow of that effort.

5.2 The transparency-by-complexity interaction

Result 4.8 isolates an interaction in which complex analyses on transparent data show a larger downward shift in a_i than either dimension alone would predict under additivity. The p -value is 0.056, close to but not meeting conventional significance. With that caveat noted, the sign is informative. The interaction is consistent with the revealed-latitude interpretation: only the transparent-complex cell supplies both the full scope of defensible analytic choices and the material to act on it. A simple analysis leaves little room for latitude even when data and code are public: the canonical pipeline collapses onto a handful of variants regardless of what the analyst can see. A complex analysis with opaque data is constrained on the other side: analysts cannot deviate from the published specification because they cannot reconstruct it. Only the diagonal cell—complex *and* transparent—supplies both dimensions, and this cell shows the largest drop in a_i .

Two reservations attach. First, the interaction is also consistent with a purely sociological alternative: transparent-complex papers may concentrate in a subfield whose methodological norms happen to depress agreement for reasons unrelated to analytic latitude, and the domain controls of Section 4 may not suffice to rule this out. Second, the direction of the causal arrow is not pinned down: complexity may expand the space of defensible pipelines, or transparency may reveal pre-existing latitude that opaque complex papers also possess but conceal. Neither reservation is resolvable from the present data.

5.3 A Gibbs-on-pipelines interpretive scaffold

The empirical regularities above fit a compact probabilistic scaffold in which each paper i induces a distribution over analytic pipelines, and the paper-level agreement rate is the image of that distribution under a deterministic conclusion map. We sketch the scaffold here without developing it as a model, because the distinguishing predictions are not tested on the present data. The scaffold is a candidate interpretation, not an inferential commitment; Appendix B specifies a protocol under which its main distinguishing predictions could be evaluated on the same data in a follow-up study.

Let Π be a discrete pipeline space whose elements π encode every analytic choice needed to produce a verdict. Let M_i be the manuscript-level observables attached to paper i , and let D_i be the dataset. On the Gibbs scaffold, analyst j on paper i draws a pipeline $\pi_{ij} \sim P(\pi \mid M_i) \propto \exp(-\beta E(\pi; M_i))$ for some energy functional E and inverse temperature β that is paper-invariant [6]. The verdict is the image $Y_{ij} = T(\pi_{ij}, D_i)$ of a deterministic conclusion map $T : \Pi \times \mathcal{D} \rightarrow \{\text{same, similar, part, no}\}$.

Two observations follow immediately. First, if T is approximately *locally constant* on the Gibbs distribution’s support—small perturbations of a canonical pipeline preserve the verdict—then the paper-level verdict concentrates on the modal verdict of the low-energy basin, and $\mathbb{E}[a_i \mid M_i]$ takes values close to 0 or 1. This is the scaffold’s reading of the bimodal morphology: papers where the manuscript-implied canonical pipeline returns **same** concentrate near $a_i = 1$; papers where it returns a non-**same** verdict concentrate near $a_i = 0$. The scaffold accommodates the empty middle of Corollary 4.3; it does not predict it against alternatives that also produce bimodal a_i .

Second, the pipeline distribution admits a single scalar summary of within-paper dispersion, analogous to an inverse effective temperature. Let $\hat{\Pi}$ be a finite canonicalised pipeline space obtained from structured columns of `corr_multi100.csv`, and let $\hat{H}_{MM}(\hat{\pi}_{i,\cdot})$ be the Miller–Madow bias-corrected Shannon entropy [8] of the empirical distribution of canonicalised analyst pipelines on paper i . Define the *pipeline-dispersion quantity*

$$I_i^* := \log |\hat{\Pi}| - \hat{H}_{MM}(\hat{\pi}_{i,\cdot}). \quad (4)$$

High I_i^* corresponds to concentrated pipeline choice (low entropy, low fragility); low I_i^* to diffuse pipeline choice (high entropy, high fragility). Under the Gibbs reading, I_i^* is a candidate paper-level sufficient statistic for fragility: conditional on I_i^* , the structural predictors in \mathcal{X} should carry no residual signal. The scaffold additionally reads the transparency paradox as a revealed-latitude effect: higher transparency raises the observed support of $\hat{\pi}_{i,\cdot}$, without changing the underlying $P(\pi \mid M_i)$, lowering I_i^* and hence a_i .

We deliberately state I^* without computing it on the staged data. The quantity depends on a canonicalisation map that must be frozen before any inferential test is run, and the scaffold’s distinguishing predictions—that I^* mediates the marginal structural correlations; that a classifier trained on I^* alone matches the multivariate classifier’s held-out AUC; that I^* is robust to reasonable perturbations of the canonicalisation—are proper pre-registration targets rather than claims of the present paper. Appendix B specifies the canonicalisation maps, the decision rules, and the sample-size considerations for such a follow-up study. Until those tests are executed, the scaffold is a candidate interpretation of the findings, not a validated explanation of them.

5.4 Meta-scientific implications

The sign-reversing transparency correlation is the finding most likely to be misread. It does *not* imply that transparency is counter-productive, that authors should post fewer materials, or that data- and code-sharing norms are misguided. On any of the three interpretations advanced above—selection, revealed latitude, or differential scrutiny—the result is about what the transparency badge observes, not about what transparency causes. A paper with high transparency and high fragility is more informative than a paper with low transparency and apparently-high agreement: the former exposes its analytic vulnerabilities to measurement, the latter conceals them. Transparency badges identify papers whose analytic latitude is measurable; the fragility metric measures it. The two are complements. The meta-scientific implication is not a retreat from transparency but an augmentation of it: fragility, empirical $1 - a_i$ or a future predicted fragility score, should be reported alongside transparency credentials, not as a substitute.

The cluster-robust attenuation adds a second implication. Paper-level marginal associations in a hundred-paper dataset are noisy, and under-powered analyst-panel inference can inflate apparent row-level significance. Multi-analyst replication studies that aggregate over rows without clustering on papers will systematically overstate the certainty of predictor claims. The paper-level Spearman and Mann–Whitney results that survive here are the inferentially appropriate object; the row-level p -values that would not survive clustering are artefacts of the sample size rather than signals about the population.

6 Limitations and Open Problems

The empirical findings of Section 4 identify two paper-level structural correlates of fragility that survive the primary inferential standard; the interpretive scaffold of Section 5.3 is advanced as a candidate organiser, not a validated theory. This section lists five limitations that bound the conclusions we are entitled to draw. Each is paired with an explicit open problem whose resolution would narrow the relevant uncertainty.

(L1) Cluster-robust attenuation of paper-level marginal significance. The headline paper-level marginal associations—most visibly Result 4.4 and Result 4.5—are fit by Spearman or Mann–Whitney tests on $N = 100$ independent paper-level observations. Those paper-level observations are themselves aggregates of $n_i \in \{4, \dots, 7\}$ analyst reanalyses. Result 4.10 reports the relevant sensitivity analysis: when standard errors are clustered at Paper_ID in the analyst-level logistic regression equivalent, *no* member of \mathcal{X} remains significant at $\alpha = 0.05$. The paper-level tests and the cluster-robust analyst-level estimates are coherent as point estimates, but their inferential weight differs sharply. We therefore frame the signed paper-level findings throughout the paper as *associations*, not as causal effects. The open problem is whether the observed paper-level signs survive a hierarchical re-analysis that fully propagates within-paper analyst variance; such a model could be fit in a follow-up study on these same data, and is a natural extension of the protocol in Appendix B.

(L2) Modest power at paper level. The paper-level sample is $|\mathcal{P}| = 100$, fixed by the SCORE program design. With $N = 100$ and the residual correlations observed in the data, the power to detect the attenuation and dominance effects contemplated in the scaffold—at the magnitudes predicted—falls below the conventional 0.80 threshold for several of the relevant tests. Any follow-up study on the present data accepts a non-trivial risk of a false-negative outcome even when the scaffold is correct. The open problem is either a pre-specified meta-analytic extension to the Many-Analysts-One-Dataset corpus and similar multi-analyst studies, or an analyst-level reformulation of the scaffold’s tests that exploits all 508 rows of `corr_multi100.csv` while correctly propagating within-paper dependence.

(L3) Novelty versus rigour on bimodality. The two-component Gaussian mixture fit to $\{a_i\}_{i=1}^{100}$ reports $\Delta\text{BIC} \approx 1190$ against the unimodal alternative, which is a strong quantitative statement. It is not, however, a qualitative discovery. The bimodality of strict agreement rates in multi-analyst datasets was already reported informally in the SCORE descriptive summary and is visible to the eye in the unprocessed distribution. The present paper adds Corollary 4.3, which sharpens the claim to an empty interval (0.2, 0.8) rather than a low-density middle, but this is a quantitative strengthening rather than a new phenomenon. A reader should weigh this accordingly: the Gibbs-on-pipelines scaffold is currently *compatible* with a known empirical pattern, not *predictive* of a previously unseen one. The open problem is to identify a qualitatively new empirical regularity that the scaffold predicts and its competitors do not, and to test it on held-out data.

(L4) Cross-validation is internal, not held-out. Result 4.7 reports $\text{AUC} = 0.667$ from five-fold cross-validation on the same Multi100 corpus used for model specification and predictor screening. This is internal cross-validation: folds are drawn from the same population on which the predictor set \mathcal{X} was defined. It is informative about in-sample predictive structure; it is not a clean estimate of generalisation error to a new multi-analyst cohort. The open problem is to secure access to an independent multi-analyst corpus (a successor to SCORE, Many-Analysts-One-Dataset waves, or a replication study on the same papers with a disjoint analyst pool) and to pre-register the held-out test before the data arrives.

(L5) Canonicalisation dependence of a putative order parameter. The scalar I_i^* of Equation (4) is defined relative to a canonicalised pipeline space $\hat{\Pi}$ derived from free-text analyst reports. Any such construction depends on the canonicalisation map ψ . A follow-up study that computes I^* on the present data must commit to a specific ψ before any inferential test is run, and must show that alternative canonicalisers preserve the paper-level ranking on I^* ; otherwise the scalar reflects the canonicalisation’s choices rather than a property of the underlying analytic structure. Appendix B frames a robustness check that would adjudicate this.

A Replication Protocol

This appendix specifies everything needed to reproduce the numerical results of Section 4 from raw SCORE data: (i) directory layout and integrity hash; (ii) commands and software environment; (iii) a manifest of twenty-one pre-specified verification claims; (iv) a table mapping every result label to a specific entry in the output JSON.

A.1 Public archives

This work is deposited in two complementary public archives:

- **Preprint (PDF only).** Zenodo, DOI 10.5281/zenodo.19811343. CC-BY 4.0. Concept DOI 10.5281/zenodo.19811342 resolves to the latest version.
- **Replication archive (data, scripts, manifest, manuscript).** OSF project, DOI 10.17605/OSF.IO/TJEHY. Contains the master JSON, both `experiments/` scripts, all five input CSVs, the evidence manifest, and a copy of this manuscript.

The two archives are cross-linked: the Zenodo record carries an *is-supplemented-by* relation pointing to the OSF DOI.

A.2 Repository layout and integrity hash

The replication archive is organised as follows, with paths relative to the archive root.

```
data/score/
  corr_multi100.csv          (508 rows, 124 columns)
  orig_outcomes.csv          (825 rows, 43 columns)
  paper_metadata.csv          (4,000 rows, 39 columns)
  corr_orig-effect-size.csv   (2,217 rows, 3 columns)
  repli_binary.csv           (274 rows, 13 columns)
experiments/
  fragility_analysis.py       (bimodality, Spearman, transparency,
                               interaction)
  multivariate_models.py      (cluster-robust, GroupKFold CV, SVM)
  paper_features.csv          (derived paper-level feature matrix)
  fragility_master_results.json (consolidated numerical output)
```

Row counts agree with Section 2. Reproduction is anchored to a single artefact: `fragility_master_results.json` has SHA-256 4948b39f0dd2a2a600fac5be08ee112bf2172a1d51e8dd732887da38408ab168. A reviewer running the protocol must obtain a JSON file whose digest matches; any discrepancy indicates a change in the staged CSVs, version drift in a dependency, or modification of the scripts. SHA-256 digests of the five staged CSVs are recorded in Table 2 so that drift from the snapshot analysed here is detectable before a reader re-runs the pipeline.

File	SHA-256
corr_multi100.csv	a8fe0bd1aeb27a7ceccc401d3c95ad2e6d68233537fe022ff8b96bff1b0bda6f3
orig_outcomes.csv	325b445d0a253342eb3d2d318aa4233bcca23c16866b5c1cc2733bab8f26478f
paper_metadata.csv	0f2beef44537aa2f123f8045281f91cc8d524186aec98e51909e7ea5934c660e
corr_orig-effect-size.csv	451f4fa098135e81d0a68b462ef7fd8c095a1c63d4fcc8394490d6141e16009e
repli_binary.csv	f16dad971507e65332b35d51ccc7dcc8ecb47cdc59281e200f79f2c3331381f

Table 2: SHA-256 digests of the five staged input CSVs at the time of analysis. Drift from these digests indicates a change in the staged data; the output JSON digest anchoring the main artefact is given at the start of this section.

A.3 Reproduction commands and software environment

From the archive root, execute the two scripts in order:

```
python experiments/fragility_analysis.py
python experiments/multivariate_models.py
```

The first script loads the CSVs, computes the per-paper strict agreement rates a_i , and produces sections A–C of the output JSON (baseline, bimodality, paper- and row-level correlations, single-predictor cross-validated AUC). The second reads `paper_features.csv` and produces sections D–G (interaction OLS, SVM comparison, full multivariate logistic, cluster-robust re-estimation). Both append to the master JSON.

Runtime dependencies are `numpy` (1.26), `pandas` (2.1), `scipy` (1.11), `statsmodels` (0.14), and `scikit-learn` (1.3). A sixth package, `dip-test`, is an optional auxiliary; it is *not* a runtime dependency of either script. When `dip-test` is unavailable, the certificate script records `dip_stat = null` and `dip_p = null` and falls back to the Gaussian-mixture Δ BIC diagnostic of Section 3. The bimodality conclusion (Result 4.2) rests on the exhaustive middle-band count and on Δ BIC; the Hartigan slot is left null in the archived master to make this explicit.

A.4 Sanity-check manifest

The `verification_summary` block records twenty-one pre-specified claims, each checked by an assertion inside one of the two scripts. A correct implementation reproduces all twenty-one truth values exactly. The claims split into three classes.

Fifteen confirmatory checks (value true). The baseline $(|\mathcal{P}|, \bar{a}) = (100, 0.36)$; zero papers in the middle agreement band and bimodal fraction ≥ 0.9 ; Δ BIC > 1190 favouring the two-component GMM; the signed paper-level Spearman for $\log n_i$; the transparency gap; the Mann–Whitney U significance for transparency; the soft-minus-hard domain gap; row-level Spearman correlations for $\log n_i$ and `Total_Hours`; the single-predictor AUC lower bound; the negative sign of the transparency \times bushel interaction coefficient; the dominance of that interaction magnitude over the transparency main effect; and the full-model cross-validated AUC point estimate.

Five falsificatory checks (value true). These claims carry a `FAIL_` prefix because they codify honest negative results: the transparency \times bushel interaction is *not* significant at $p < 0.05$; the RBF SVM does *not* outperform the linear kernel; the $\log n_i$ coefficient in the cluster-robust logistic is *not* significant at $p < 0.05$; zero predictors survive at $p < 0.05$ under cluster-robust standard errors; and the full-model cross-validated AUC does *not* exceed 0.65. Each carries value `true` to assert the pre-specified null was observed.

One partial check. The transparency odds-ratio claim records $OR \approx 2.86$ against a pre-specified $OR \geq 4.61$ target. The ratio agrees in sign and direction but not in magnitude; the manifest stores "partial".

A.5 Result-to-JSON map

Table 3 lists, for each empirical result label, the section of `fragility_master_results.json` carrying its numerical values and the associated `verification_summary` row(s). Every quantitative statement in the main text traces through this table to a primitive JSON field.

Result label	JSON	verification_summary row(s) (abbreviated)
4.1	baseline	baseline_100papers_036mean
4.2	A	middle_zero, bimodal_90pct, delta_bic_gt_1190
4.4	B, C	logN_rho_sig, row_logN
4.9	B	z0_null
4.5	B	transparency_gap, mwu_sig, PARTIAL_OR_2857_not_461
4.6	C	row_hours
4.8	D	interaction_negative, interaction_mag_gt_main, FAIL_not_sig
4.7	C, E, F	single_pred_auc, FAIL_rbf_worse, full_model_auc, FAIL_auc_below_065
4.10	G	FAIL_logN_cr_not_sig, FAIL_zero_sig_predictors

Table 3: Mapping from empirical result labels to sections A–G of `fragility_master_results.json` and to rows of the `verification_summary` manifest.

B Protocol for a Proposed Follow-up Study

The Gibbs-on-pipelines scaffold of Section 5.3 generates three distinguishing predictions: (i) a scalar dispersion quantity I_i^* mediates the marginal structural correlations of Section 4; (ii) a classifier trained on I^* alone attains held-out AUC competitive with the multivariate classifier on \mathcal{X} ; (iii) the paper-level ranking on I^* is robust to reasonable perturbations of the canonicalisation map. These predictions are not evaluated in the present paper. This appendix fixes the operational details needed to execute them as a follow-up study on the same Multi100 dataset or on an independent multi-analyst corpus; any divergence between these specifications and the analysis code invalidates the corresponding test. Post-hoc rescues—adding an auxiliary covariate, broadening a decision threshold, or redefining a predictor—are disallowed.

B.1 Canonicaliser specifications

The canonicalisation machinery reduces free-text analyst reports to finite-alphabet tokens. Two canonicalisers are pre-specified on `Direction_of_Result` (a coarse map ϕ_{coarse} and a fine map ϕ_{fine}) and two on `Task1_Free_Text` (ψ_{coarse} and ψ_{fine}). The two resolutions exist so that the robustness test (prediction iii) can be run on a genuine orthogonal grid rather than on a single arbitrary choice.

Direction canonicaliser ϕ_{coarse} . A map $\phi_{\text{coarse}} : \text{free-text} \rightarrow \{-1, 0, +1, \text{null}\}$ applied to `Direction_of_Result`. After Unicode NFKC normalisation and lower-casing, a regular-expression cascade is evaluated top-to-bottom, first match wins: `null` on empty or “not applicable” matches; `+1` on `positive|increase|higher|greater|larger|supports|confirms`; `-1` on `negative|decrease|lower|smaller|opposite|contradict|reversed`; `0` on

`null|no (effect|evidence|difference)|insignificant`; and `null` otherwise. The complete regex set is stored in `canonicalisers/phi_coarse.json`.

Direction canonicaliser ϕ_{fine} . A seven-level refinement of ϕ_{coarse} augmenting it with magnitude hedges (“marginally positive”, “strongly reversed”), with the lookup table in `canonicalisers/phi_fine.csv`. Unmatched phrases fall back to ϕ_{coarse} , then to `null`.

Pipeline canonicaliser ψ_{coarse} . A map from `Task1_Free_Text` to the tuple $\hat{\pi} = (\text{estimator}, \text{covariate set}, \text{sample restriction}, \text{inference})$ with fixed alphabets of sizes 8, 4, 5, 5 and $|\hat{\Pi}_{\text{coarse}}| = 800$. Definitions are stored in `canonicalisers/psi_coarse.json`. Entries shorter than fifteen non-whitespace characters or matched against a boilerplate blacklist return `null` across all four coordinates.

Pipeline canonicaliser ψ_{fine} . A refinement with alphabet sizes 14, 9, 8, 7, giving $|\hat{\Pi}_{\text{fine}}| = 7,056$ and stored in `canonicalisers/psi_fine.json`. Fine alphabets are strict refinements of coarse alphabets; a collapse-map from fine to coarse is packaged alongside.

B.2 Test protocols

The three tests are executed under a single Bonferroni family of size three: the per-test significance threshold is $\alpha = 0.05/3 \approx 0.0167$. The global random seed is `SEED_ROOT = 20260420`. Derived seeds are computed deterministically from `SEED_ROOT` and the test name. All bootstrap resamples are cluster bootstraps at the paper level: \mathcal{P} is resampled with replacement and all within-paper analyst rows are retained, preserving the empirical n_i distribution.

(P1) Mediation attenuation of \mathcal{X} by I^* . For each $X_k \in \mathcal{X}$, fit the paper-level OLS marginal slope β_k^{marg} of a_i on X_k and the conditional slope β_k^{cond} of a_i on (X_k, I_i^*) . Mediation attenuation is tested by non-parametric cluster bootstrap with $B_{\text{med}} = 10,000$ resamples; the bootstrap distribution of the paired difference yields a 99% percentile CI. The prediction is supported if, for at least six of the ten predictors in \mathcal{X} , the 99% CI excludes zero in the attenuation direction, and at most one predictor exhibits sign-inconsistent attenuation.

(P2) Held-out AUC matching by I^* . Data are partitioned by stratified paper-level five-fold cross-validation with stratum equal to the fragility tertile of a_i . For each fold, two logistic classifiers for $\mathbb{K}[a_i \geq 0.5]$ are fit: $f_{\mathcal{X}}$ on \mathcal{X} alone and $f_{\mathcal{X}, I^*}$ on $\mathcal{X} \cup \{I_i^*\}$. DeLong’s paired test on the pooled held-out predictions compares $\text{AUC}(f_{\mathcal{X}, I^*})$ against $\text{AUC}(f_{\mathcal{X}})$. The prediction is supported if the one-sided DeLong $p < 0.0167$ and the point-estimate AUC gap exceeds 0.04.

(P3) Canonicalisation robustness. For each pair $(\phi_*, \psi_*) \in \{\text{coarse}, \text{fine}\}^2$, recompute I_i^* under that canonicaliser pair and form the Spearman correlation $\rho_{(\phi, \psi)}(I_i^*, 1 - a_i)$. Paper-level bootstrap with $B_{\text{sp}} = 5,000$ resamples yields per-pair 99% CIs and pairwise-difference CIs. The prediction is supported if all four point estimates of ρ lie in $[0.35, 0.75]$ and the bootstrap 99% CI for the maximum pairwise difference $\max_{(\phi, \psi) \neq (\phi', \psi')} |\rho_{(\phi, \psi)} - \rho_{(\phi', \psi')}|$ is contained in $[0, 0.10]$.

Family-wise stopping rule. Tests are executed in the order (P1), (P2), (P3). The Bonferroni adjustment is applied across all three regardless of individual outcomes: no early stopping, no reallocation of α , and no dropping a test from the family after the fact. A negative result on any of the three is informative: the Gibbs-on-pipelines scaffold as stated in Section 5.3 would be falsified in the pre-specified sense.

B.3 Release commitment

Before any of the three tests touches Multi100 data, the four canonicaliser lookup tables (`canonicalisers/phi_coarse.json`, `canonicalisers/phi_fine.csv`, `canonicalisers/psi_coarse.json`, `canonicalisers/psi_fine.json`), the seed schedule, and the three analysis scripts (`followup/mediation.py`, `followup/held_out_auc.py`, `followup/canonicalisation_robustness.py`) must be bundled and deposited as a single hash-pinned release tag in the replication archive. The SHA-256 digest of the bundle is to be archived with a timestamped third-party notarisation service. Any error discovered post-release must be documented as a deviation, with the affected test flagged as invalidated rather than silently repaired. The empirical findings of Section 4 are independent of this protocol and stand on their own.

References

- [1] Balazs Aczel et al. Consensus-based guidance for conducting and reporting multi-analyst studies. *Nature*, 2026. SCORE Multi100 release, data at <https://osf.io/dtzz4/>.
- [2] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.
- [3] Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Juergen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.
- [4] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, working paper, 2013.
- [5] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.
- [6] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.
- [7] R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams, et al. Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3):142–152, 2014.
- [8] George A. Miller. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, IL, 1955.
- [9] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, 2017.
- [10] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

- [11] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
- [12] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018.
- [13] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.
- [14] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020.
- [15] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.