Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Why Is Reviewer 2 Always the Villain? Myth, Bias, or Statistical Reality?

Muh. Akbar Idris

muhakbaridris@apps.ipb.ac.id

IPB University

Research Article

Keywords: peer review, reviewer position, polarization, ordinal measurement, editorial decisions, cross-source triangulation

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11740871/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Why Is Reviewer 2 Always the Villain? Myth, Bias, or Statistical Reality?

Large-scale evidence on reviewer position, ordinal disagreement, polarization, and editorial associations

Muh. Akbar Idris[a,b,*]

[a]*Department of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, Indonesia*

[b]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat, Majene, Indonesia*

*Corresponding author. Email address: muhakbaridris@apps.ipb.ac.id (Muh. Akbar Idris)

Abstract

The Reviewer 2 stereotype conflates recommendation severity, polarization, disagreement, and editorial influence. We examined these dimensions using metadata-only records reconstructed from the Open Review-Based (ORB) archive, focusing on ordinal rating behavior rather than review-text tone or civility. After deduplication, the database contained 62,935 manuscripts and 157,070 review records. The primary explicit-label sample included 8,885 manuscripts and 29,166 reviews, concentrated predominantly in OpenReview machine-learning conferences (7,966 from ICLR 2018–2021). Manuscript-balanced ordinal contrasts, venue-specific mixed models, endpoint analyses, and independent editorial outcomes separated observable positions from decision mechanisms. Label two showed no clear excess harshness: -0.008 (95% CI -0.022 to 0.006). Its excess probability of either scale endpoint was 0.003 (95% CI -0.004 to 0.009); median-distance and broader-tail analyses likewise did not support general polarization. Public labels matched creation chronology on only 30.1% of eligible reviews. Among 5,964 decision manuscripts, greater second-reviewer severity was associated with rejection, but the severity-by-peer-disagreement interaction was inconclusive (P=0.637). In a retrospective audit of 1,180 held-out final archived snapshots, an equal-weight mean outperformed a separate-R2 model in Brier loss (0.0830 vs. 0.0853). A paired MIDL subset showed reduced disagreement after rebuttal across 85 manuscripts, without clear second-reviewer-specific change. The supplied small F1000Research case was treated solely as an unverified historical audit, not an independent replication. Within these observed open-conference settings, the evidence favors a measurement- and institution-dependent account over universal Reviewer 2 ordinal harshness or polarization; generalizability to traditional cross-disciplinary journals, linguistic hostility, and prospective editorial weighting remain unmeasured.

Keywords: peer review; reviewer position; polarization; ordinal measurement; editorial decisions; cross-source triangulation

Introduction

“Reviewer 2” names a recognizable experience: a manuscript receives one encouraging assessment and another that seems disproportionately severe, obstructive, or dismissive. The label is compelling because it gives an otherwise impersonal editorial process a recurring antagonist. Yet three propositions are easily conflated. Authors may experience an especially difficult reviewer; that reviewer may receive the number two in a particular workflow; and reviewers carrying that number may, on average, behave differently from their colleagues. Evidence for the first proposition does not establish the second or third. A statistical investigation must therefore begin by specifying which reviewer number is observable and which behavior is being measured.

The existing direct evidence does not justify assuming a universal numbered-reviewer effect. Worsham et al. (2022) examined initial reports on 794 published BMJ manuscripts and found no statistically significant differences in their selected linguistic outcomes between the second reviewer and other reviewers. Their numbering represented the order in which comments were returned, rather than invitation order. Peterson (2020) likewise addressed the stereotype through editorial records in political science. These studies make reviewer position an empirical question, but they do not establish that all journals, all numbering systems, or all manifestations of criticism are interchangeable. A numerical recommendation, a demanding methodological objection, and disrespectful language are distinct outcomes.

Recent qualitative work reinforces that distinction. Caner and Asma (2026) studied experiences reported by 93 academics and described the emotional and professional consequences associated with the Reviewer 2 phenomenon. Such evidence is relevant to how the stereotype is understood and reproduced, but it does not estimate the difference between publicly numbered reviewers. Conversely, a small average difference in recommendations would not invalidate an author's experience of an abusive report. Watling et al. (2021) discuss constructive reviewing in terms that concern how criticism is communicated. The present study concerns the distribution of recorded ratings and their relationship to editorial outcomes, without attempting to infer civility or the substantive correctness of a criticism from its score.

There is also a measurement problem upstream of statistical modelling. Public review databases often retain recommendations while discarding timestamps, anonymous reviewer labels, or the distinction between editorial decisions and recommendations. A database can be large enough to estimate very precise associations while remaining unable to identify the reviewer position named in the research question. Storage order is especially tempting because it produces an apparently complete variable at negligible computational cost. It is also scientifically inadequate unless the source establishes what that order means. Adding a sophisticated model after assigning unsupported positions would make the resulting inference more precise without making it more valid.

This investigation reconstructs an existing small-sample project around those identification requirements. The supplied project proposed a Reviewer Harshness Index, a standardized variant, empirical Bayes shrinkage, leave-one-position-out disagreement attribution, a Reviewer 2 Disproportionate Influence Index, and a Reviewer 2 Effect Decomposition Model. Their arithmetic and interpretation are audited separately. The reconstructed empirical study uses the Open Review-Based dataset, ORB, together with metadata recovered from its separately published raw OpenReview component. No manuscript PDFs, abstracts, report text, or reviewer identities are needed in the final analytical database. The primary position variable is an explicit anonymous reviewer number; creation chronology is a separate secondary definition.

The central objective is to distinguish five questions: whether the publicly numbered second reviewer is systematically less favorable, more likely to use end categories, or associated with greater disagreement than peers evaluating the same manuscript. A secondary objective concerns the conditional association between that reviewer's rating and an explicit editorial decision. These are exploratory research questions rather than preregistered directional hypotheses. Harshness, polarization, disagreement, editorial association, and stage-related change are evaluated as distinct dimensions. Revision and rebuttal questions require actual paired measurements and are screened for feasibility before analysis. The intended contribution is an auditable operationalization and coherent statistical workflow, not the invention of new names for established estimators.

Theoretical Background and Literature Review

Peer-review research provides several reasons to resist interpreting disagreement as evidence of an unusually problematic reviewer. Bornmann et al. (2010) synthesized inter-rater reliability studies using a multilevel approach, establishing reliability as a question that varies across settings rather than a property guaranteed by expert judgment. Tennant and Ross-Hellauer (2020) identify limits in the evidence available about peer review itself. Their broader implication for the present design is methodological: conclusions should be conditioned on the editorial systems, records, and populations actually observed. Publicly accessible reviews offer valuable evidence, but accessibility is a sampling mechanism, not a guarantee of representativeness.

The distinction between disagreement and bias is particularly important. Lee et al. (2013) organize concerns about bias in peer review across several mechanisms. Tomkins et al. (2017) examine reviewer bias under single- and double-blind review, while Squazzoni et al. (2021) investigate gender-related patterns across 145 journals. These studies address substantive attributes and institutional arrangements, whereas a reviewer-number analysis concerns a label generated by an editorial workflow. The existence of bias along one dimension cannot be used as evidence of bias attached to the number two. Equally, a null position contrast would leave other forms of unequal treatment unresolved.

Social and editorial context can also shape recommendations. Teplitskiy et al. (2018) investigate professional networks and peer judgment, and Siler et al. (2015) study scientific gatekeeping. Their questions concern relationships between assessments, selection, and scientific outcomes that extend beyond a reviewer's displayed number. Such contextual differences provide alternative explanations for apparent position effects. Editors might recruit an additional specialist for a difficult manuscript, different venues might number reviewers differently, and more demanding cases might accumulate more reviews. Within-manuscript comparisons remove differences in which manuscript was evaluated, but they do not eliminate differences in the expertise or role assigned to each reviewer.

The recent literature also cautions against treating peer review as a simple vote. De las Cuevas (2026) analyzes recommendation disagreement across review rounds and editorial outcomes in anonymized Healthcare records. Their study distinguishes published and rejected manuscripts and follows manuscript-level trajectories. Mutz et al. (2025) compare aspects of eLife's old and new publishing arrangements, including inter-rater agreement and reliability. Both studies motivate attention to institutional processes and the meaning of an outcome. Neither supplies evidence that the publicly numbered second reviewer causes rejection. Their contribution here is to clarify why agreement, recommendation severity, and editorial adjudication require separate estimands.

Reform proposals do not resolve these measurement distinctions automatically. Waltman et al. (2023) organize peer-review reform into four schools of thought, while Aczel et al. (2025) review ideas, interventions, and evidence on its present and future. Malički and Mehmani (2024) examine structured peer review across 23 journals, directly connecting the design of review questions to the way agreement can be investigated. A system can improve the usefulness or transparency of reports without increasing numerical consensus. Conversely, apparent consensus can reflect a narrow set of publicly observed manuscripts. For a position study, these considerations favor reporting source-specific eligibility and preserving the underlying evaluation scales.

Open peer review enables analysis while introducing its own definitional variability. Ross-Hellauer (2017) reviews the concept systematically, and Wolfram et al. (2020) examine transparency in open science. “Open” can refer to identities, reports, interactions, or other

elements of a process; it should not be interpreted as a uniform data-generating regime. ORB combines OpenReview, SciPost, and PeerJ, but a common export structure does not make their practices identical. The extraction therefore preserves source and venue indicators, original rating descriptions, and version labels. The analytical sample is an explicitly justified subset of that database, rather than the entire collection relabeled as a homogeneous peer-review experiment.

Computational resources further demonstrate the difference between possessing review text and possessing the metadata needed for a position analysis. Ghosal et al. (2022) provide a benchmark for computational analysis of peer reviews, and Severin et al. (2023) investigate report thoroughness and helpfulness in relation to journal impact factor. Their outcomes concern features that cannot be inferred reliably from a recommendation alone. The present work deliberately excludes supplementary language analysis: the supplied excerpts lack a retrieval trail sufficient to establish that they are complete, verbatim reports, and report text is unnecessary for the primary ordinal questions. This keeps the research aligned with the available evidence and limits the amount of content redistributed.

Reviewer assignment supplies another reason to avoid treating a label as a reviewer trait. Ragone et al. (2013) examine conference review effectiveness, scales, and rating behavior, while Shah (2022) surveys challenges, experiments, and computational approaches to peer review. A position pattern could arise because different expertise is selected for different roles, because reports arrive at different times, or because an additional reviewer is recruited for a difficult case. These mechanisms require assignment records to distinguish them. The current archive allows label and chronology comparisons but does not reveal the editor's reason for recruiting a particular reviewer.

Evidence about evaluating reviews also complicates an interpretation based entirely on recommendation severity. Pier et al. (2018) found low agreement when reviewers evaluated the same grant applications, in a setting distinct from conference acceptance. Goldberg et al. (2025) experimentally investigated review-length bias and also found observational author-outcome bias in evaluations of reviews. That work makes a cognitive or evaluative account of the stereotype plausible without proving that the number two causes it. A recommendation that favors the author's outcome can affect the perceived quality of a review; identifying the cultural mechanism would require linked perceptions, text, and assignment information that is absent here.

The statistical literature supplies the necessary tools without requiring a collection of purportedly new indices. McCullagh (1980) established regression models for ordinal responses, and Agresti (2010) provides the framework for interpreting ordered categories without assuming equal distances. Hedeker and Gibbons (2006) discuss models for clustered and longitudinal data; Hedeker et al. (2008) show why location and scale can be distinct modelling targets. These foundations motivate an ordinal location model and separate categorical end-point analyses. A Gaussian variance comparison on a handful of ordered categories would not, by itself, establish polarization or distinguish it from a shift in the distribution's location.

Agreement measures answer related but nonidentical questions. Cohen (1968) introduced weighted agreement, while Gwet (2008) addresses chance-corrected reliability under high agreement. Ordinal Krippendorff agreement uses the marginal distribution to determine category distances; quadratic weighting instead assigns distances from ordered category positions. Reporting both makes that dependence visible. Neither coefficient identifies who is right, and deleting a reviewer changes the marginal distribution as well as observed

disagreement. This is why leave-one-position-out results are calculated on a matched set of manuscripts and treated as descriptive perturbations rather than causal attribution.

Grant-review costs also concern institutions rather than numbered reviewer traits. Pollitt et al. (2026) analyze costs and benefits in funding peer review; that setting informs the distinction between review effort and decision utility, but its outcomes are not pooled with manuscript acceptance.

The remaining gap is therefore precise. Existing work supports studying reliability, reviewer experience, and editorial adjudication, but these strands do not identify a universal Reviewer 2 effect across incompatible numbering conventions. The present study connects a provenance audit, a documented public-label estimand, manuscript-balanced ordinal contrasts, source-specific mixed models, and matched disagreement diagnostics. A literature comparison matrix records how the selected studies relate to those choices and distinguishes direct position evidence from contextual and methodological sources. This comparison does not establish methodological priority; it defines the narrower claim that the reconstructed project can defend.

Conceptual Framework and Research Questions

Observable reviewer positions

Four position concepts must remain separate. Formal assignment order records which invitation or assignment came first. An explicit displayed number is a publicly visible label attached to a reviewer. Submission chronology orders reports by a recorded time, with its meaning dependent on the timestamp field. Unknown order is the correct classification when none of those concepts can be established. ORB's harmonized review object does not contain dates or reviewer-number fields, so an integer obtained by enumerating its review list belongs only to storage provenance. It is retained for traceability and never used as a reviewer-position measurement.

The raw OpenReview component preserves anonymous signatures and true creation timestamps. Numeric signatures permit the explicit-number estimand even when the order in the exported list differs. The signature is evidence of a public label, not proof that the corresponding reviewer was the second person invited or assigned. True creation time permits a separate chronology estimand. Dates with incomplete coverage or exact ties are not broken using row order or the review identifier. Anonymous labels are manuscript-local; they are not linked across papers to create reviewer histories, and personal names are not retained.

Locates these measurement in Fig. 1 choices before statistical estimation. Its separation of position definition, severity, end-category propensity, disagreement, and decision association is substantive: no one outcome is treated as a validated substitute for the others. The same diagram also indicates where the research design ceases to support an inference. Without repeated ratings, for example, a final snapshot cannot establish response to rebuttal. Without an independently recorded editorial decision, a recommendation-based rule cannot establish editorial weighting.

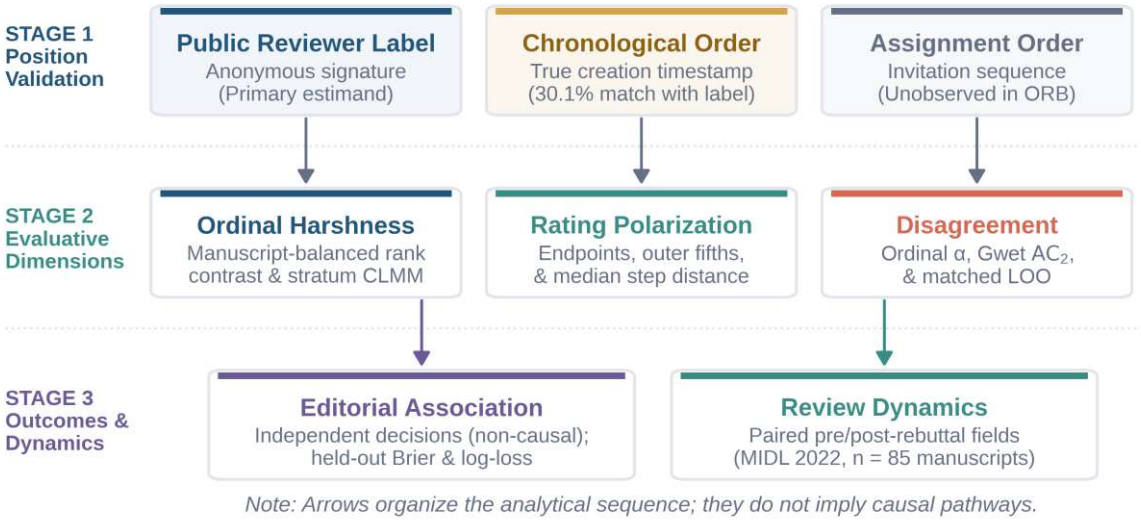


Fig 1. Conceptual framework separating measurement, ordinal location, polarization, disagreement, editorial association, and paired review dynamics. Arrows organize the analysis; they do not assert causal pathways.

Harshness as a manuscript-balanced ordinal contrast

Let Y_{mi} denote the recorded rating for reviewer i on manuscript m, with higher categories representing a more favorable evaluation in a validated venue-specific ordering. Let n_m denote the number of eligible reviews on that manuscript. A transparent descriptive contrast compares the second reviewer's score with the average score of peers. Its raw form is retained within a venue-year and written as Equation 1.

$$R_{m2} = Y_{m2} - \frac{1}{n_m - 1} \sum_{i \neq 2} Y_{mi}. \quad (1)$$

Negative values of R_{m2} indicate a less favorable score than peers. This contrast uses the numerical spacing of a particular scale. It is therefore unsuitable as the sole primary measure across different venues, and rescaling it by a pooled standard deviation does not restore information about ordinal distances. With only two reviewers, a peer standard deviation is undefined; with unanimous peers it can be zero. An arbitrary denominator floor would introduce another tuning decision. The reconstructed study reports raw contrasts only within their original scale and does not elevate the standardized variant into a new methodological contribution.

The primary comparison uses ordering alone. Equation 2 averages signed pairwise differences between the second reviewer and each peer, and then gives every manuscript equal weight. Here M is the number of eligible manuscripts and sgn returns minus one, zero, or one.

$$H_2 = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n_m - 1} \sum_{i \neq 2} \operatorname{sgn}(Y_{mi} - Y_{m2}). \quad (2)$$

Positive values of H_2 mean that Reviewer 2 is more often below peers than above them. Ties contribute zero. The statistic is bounded by minus one and one and is unchanged by a strictly increasing recoding within a rating scale. Equal manuscript weights prevent papers with many reviewers from dominating the target. Averaging pairwise signs is an established rank-based idea; its use here is an operational choice, not a claim of mathematical novelty. Confidence intervals are obtained from manuscript-level contrasts, with a manuscript bootstrap as a check on the large-sample approximation.

Conditional ordinal location modelling

The secondary model is a cumulative-link model with a manuscript random intercept. It is fitted separately within eligible venue-year strata, allowing each setting to retain its own categories and thresholds. Equation 3 expresses the logistic version; k indexes an ordered cutpoint, s denotes a stratum, and u_m captures manuscript-level heterogeneity.

$$\text{logit}\Pr(Y_{mi} \leq k \mid u_m, s) = \theta_{sk} - \beta_{s2} I(i = 2) - \beta_{s3} I(i \geq 3) - u_m. \quad (3)$$

The random intercept follows a mean-zero normal distribution with variance σ_s^2. The reference category is the explicitly labeled first reviewer. A positive β_{s2} corresponds to a more favorable distribution for Reviewer 2; its exponential is a conditional odds ratio for being above any cutpoint under the proportional-odds assumption. Thresholds supply the model intercept, so an additional free intercept is not included. Ordered threshold increments are parameterized exponentially. The likelihood integrates over the manuscript effect using Gauss-Hermite quadrature, and the implementation reports the numerical gradient and observed-information diagnostics.

This model does not identify stable individual reviewer severity because the retained identifiers do not link anonymous persons across manuscripts. It provides partial pooling through manuscript-level heterogeneity and estimates position contrasts across manuscripts in the same stratum. A position-average shrinkage formula based on only three aggregated means is not treated as a validated empirical Bayes reviewer model. Proper reviewer-specific shrinkage would require reliable repeated identities and a defensible hierarchical variance model, neither of which can be created by treating Reviewer 2 on different papers as the same person.

Proportional odds is evaluated rather than assumed to follow from model convergence. Brant (1990) motivates examining threshold-specific associations, but the present clustered marginal diagnostic is not mislabeled as a conditional Brant test for the random-intercept model. Threshold-specific binary regressions are jointly assessed with manuscript-clustered covariance. Sensitivity fits use more quadrature points and a probit link. When threshold patterns differ materially, the manuscript-balanced rank contrast remains the primary estimand and the single mixed-model coefficient is interpreted as a restrictive summary of a distributional relationship.

End-category tendencies and disagreement

Polarization is operationalized as a difference in the probability of the lowest and highest observed categories within each venue-year. The two tails are reported separately, along with their union. These are observed end categories, not an assertion that every nominal scale endpoint appears in the archive. For each tail, the second reviewer's indicator is compared with the mean indicator among peers on the same manuscript. This pairing controls manuscript composition without assuming equal category distances. Increased use of one tail alone should be described as a tail-specific difference; increased use of both tails would provide stronger evidence for a polarization interpretation.

Disagreement is summarized using ordinal Krippendorff alpha, quadratic-weighted Gwet AC2, and observed weighted agreement. Equation 4 gives the generic chance-correction form for alpha, where D_o is observed disagreement and D_e is expected disagreement under its marginal reference distribution.

$$\alpha = 1 - \frac{D_o}{D_e}. \quad (4)$$

The corresponding AC2 calculation uses its own chance-agreement construction and quadratic agreement weights. Its numerical value need not match alpha. Comparisons are made within a

stratum rather than across pooled category labels from incompatible scales. For position removal, the same complete manuscripts with exact labels one, two, and three contribute to every comparison. Exactly one rating is removed at a time. This avoids an apparent improvement that arises merely because removing Reviewer 2 eliminates all two-review manuscripts from the eligible set. Changes in either chance-corrected coefficient remain sensitive to the altered marginals.

Editorial association and the limits of decomposition

An explicit editorial acceptance or rejection is modelled once per manuscript. Equation 5 conditions on the second reviewer's score, the mean of the other scores, the number of reviews, and venue-year indicators. A_m is the acceptance indicator, z_{m2} and $\bar{z}_{m,-2}$ are scores scaled within a stratum's observed range, and γ_s represents stratum effects.

$$\text{logit}\Pr(A_m = 1) = \gamma_s + \delta_2 z_{m2} + \delta_o \bar{z}_{m,-2} + \delta_n n_m. \quad (5)$$

This is an associational model of archived snapshots. It does not establish that editors saw the archived rating before deciding, that they weighted reviewers according to the fitted coefficients, or that changing a reviewer's recommendation would change the outcome. A standardized probability contrast compares predictions at the first and third quartiles of the second-reviewer score while retaining the observed covariates. It is more interpretable than calling a ratio of selected conditional probabilities "disproportionate influence." A same-manuscript holdout compares predictive performance with a model using an equal-weight mean score, without presenting retrospective prediction as a prospective decision tool.

The original decomposition is not retained as an identified causal model. An outcome gap between manuscripts with favorable and unfavorable second-reviewer ratings can be decomposed algebraically by a linear probability regression, but the residual may be zero as an accounting consequence of fitted group means. That does not separate intrinsic harshness, manuscript quality, editor sensitivity, and unobserved confounding. Similarly, F1000Research indexing eligibility constructed from review outcomes is not an independent editorial response. Keeping these distinctions explicit is a substantive correction to the original project rather than the loss of a valid causal contribution.

Data and Methods

Source hierarchy and triangulation design

The fixed ORB package is the primary empirical source. Its current manuscript supplies the initial structure. The supplied F1000Research manuscript, older package, and integration memorandum supply hypotheses and an arithmetic-audit case. File hashes establish that the supposedly newer small manuscript is identical to the previously supplied version. The redesigned ORB Word document has the same paragraph text as the frozen ORB manuscript; the accompanying PDF is a rendering, not another dataset. Repeated descriptions of the same 21 articles and 71 reports are counted once.

The supplied small-study acquisition script constructs observations directly, and no retrieved report history establishes the authenticity of its recommendation records. Attempts to inspect representative F1000Research pages returned access errors. The case is therefore designated unverified supplied metadata throughout, rather than an independent empirical replication. Its numerical calculations can be reproduced, but its proposed explanations cannot be transferred as findings. The ten supplied ICLR 2026 papers are all accepted and cannot support an acceptance model. Neither small sample is pooled with ORB, and short report excerpts are excluded from linguistic inference.

Source inspection and minimal acquisition

The ORB record described by Szumega et al. (2024) is the fixed Zenodo release dated 30 August 2024, version 1.2.0, DOI 10.5281/zenodo.13628746, accessed on 14 November 2026. Its downloadable objects are a schema image, a harmonized JSON export, a transformed dataset pickle, and three source-specific extraction pickles. They are alternative and source-specific representations, not a single four-gigabyte compressed archive. The JSON export is 1,215,940,562 bytes. The raw OpenReview component is 968,765,991 bytes. The other large components are not needed to recover the fields used here and are not acquired.

The server supports byte ranges, verified through HTTP status 206 and the exact Content-Range header. Acquisition uses resumable ranges with retries and records source checksums and transfer provenance. Every byte of the two selected components must be transferred for this extraction route; projecting fields after parsing does not mean that only those fields were downloaded. The process temporarily holds these components in the workspace, validates their published MD5 checksums, and deletes their temporary chunks after successful metadata extraction. No linked manuscript PDF, supplementary attachment, figure, or source manuscript is downloaded.

Database construction and validation

Six compressed Parquet tables separate paper metadata, review records, evaluation fields, editorial decisions, reviewer-position evidence, and the integrated analytical dataset. A review key combines the paper, version, source review identifier, and source list index so that placeholder identifiers in one source cannot collide globally. Duplicate source keys are audited separately from the uniqueness of this constructed key. Paper identifiers are checked before a many-to-one merge, review joins are checked as one-to-one, and the pipeline rejects unsupported position assignments. Data dictionaries retain the meaning of derived fields, missingness, and the difference between source-level and analytical counts.

Editorial fields receive a specific provenance check. The published OpenReview converter can substitute a meta-review recommendation when an explicit decision is absent. The SciPost converter can derive its decision string from recommendations. Neither substituted field is accepted as an independent final editorial decision in the main analysis. For OpenReview, an explicit raw decision invitation is required; PeerJ editorial statuses are retained as source-specific metadata. Unknown status is not recoded as rejection. A source Boolean indicating acceptance is preserved for auditing but does not override a missing or ambiguous decision string.

The interpretation of review rounds also depends on source. ORB collapses OpenReview records into one final article version, even where the underlying system supports revisions. That version label cannot be interpreted as an initial round. PeerJ and SciPost may retain multiple article versions, but their available rating and position information differs from OpenReview. Presence of a rebuttal record is distinguishable from absence of a recorded rebuttal, not from proof that no rebuttal occurred. These constraints are recorded before defining the eligible samples and prevent the nominal size of the archive from being mistaken for the size of every analysis.

Eligibility and rating preservation

The primary numbered-reviewer sample includes OpenReview manuscripts with at least two official reviews, usable numerical overall ratings for all retained reviews, complete explicit numeric labels, and exactly one label two. Duplicate labels within a manuscript make that label-based assignment ineligible. The chronological sample independently requires complete,

unique true creation timestamps and usable ratings. It orders the reviews by those timestamps without borrowing the explicit labels. These samples target different observable positions and are never combined into one undifferentiated Reviewer 2 coefficient.

The harmonized overall field is named “final” in ORB. It is described here as a harmonized archived snapshot because the converter's selection among source evaluation keys does not guarantee a rating measured after rebuttal. Original short evaluation names are retained when recoverable from the raw component. Rating categories remain within their venue-year ordering for ordinal modelling, and numerical means are not compared across venues without explicit qualification. Confidence and auxiliary evaluation dimensions are retained where present; they are not imputed simply to enlarge a model's covariate list.

The primary rank contrast is invariant to monotone recoding, but eligibility can still depend on missingness. Complete-case restriction may exclude manuscripts with incomplete reports, hidden decisions, or nonstandard formats. Source-coverage and venue-year tables therefore report how the analytical population differs from the full archive. Sensitivity analyses include an exact first-versus-second comparison, reviewer-count strata, year-specific estimates, and the chronology-defined second report. These comparisons diagnose dependence on sample construction and estimand definition; they are not independent replications from randomly sampled editorial systems.

Estimation and diagnostics

The manuscript is the independent sampling unit for the primary analysis. Standard errors for the mean manuscript contrast use its empirical variability, and a 2,000-replication manuscript bootstrap provides percentile intervals, following the resampling principles discussed by Efron and Tibshirani (1994). This avoids treating reviews of the same manuscript as independent manuscripts. Repeated records from different sources are not pooled into the primary model. Anonymous reviewer labels do not support cross-manuscript reviewer clustering, and the resulting limitation is retained in the interpretation rather than resolved through a fabricated stable reviewer identifier.

Marginal threshold diagnostics and end-category analyses examine whether a single location shift misses relevant distributional patterns. Because these constitute a set of exploratory comparisons, isolated small probabilities are interpreted with their effect sizes and intervals. Benjamini and Hochberg (1995) provide the rationale for false-discovery-rate control when screening families of hypotheses. The core primary contrast is presented separately from those screens; no uncorrected collection of tail, venue, and year findings is promoted to a confirmatory discovery. All exact numerical probabilities are retained in machine-readable tables rather than rounded to zero in the manuscript.

For the secondary chronology-defined rebuttal analysis, original preliminary and explicitly post-rebuttal fields are paired by review identifier. Matching category descriptions are required before interpreting a change. Each paired review is coded as decreased, unchanged, or increased; its signed change is minus one, zero, or one. The chronological second review is compared with the mean signed change of paired peers within the same manuscript. Only manuscripts with both a paired second review and at least one paired peer contribute. Manuscript-level intervals respect this pairing. Raw category-score changes are a supplementary sensitivity analysis within the single matched scale. This design identifies stage-related changes in a selected subset, not a causal effect of rebuttal.

Enriched distributional and editorial analyses

Polarization extensions retain manuscript-balanced endpoint contrasts and add an outer-category indicator: the bottom or top fifth of ordered category ranks in each stratum. The fifth refers to the scale's ordered positions, not a sample quantile of manuscripts. Median distance uses the absolute difference between a review's category rank and the manuscript median rank, divided by the largest category rank. This is a transparent ordered-step distance, not a claim of equal substantive intervals. Numerical within-manuscript variance is reported only as a coding-dependent supplement. Five dispersion contrasts receive a common false-discovery-rate adjustment. Two thousand manuscript bootstrap draws and a manuscript-preserving label permutation check provide complementary uncertainty assessments; the permutation assumes label exchangeability under the null and does not establish random assignment.

An adjusted logistic endpoint model includes public-label groups, venue-year indicators, review count, and relative creation chronology; covariance is clustered by manuscript. Constant or very sparse endpoint strata are excluded from that particular model and documented separately. The largest ordinal stratum additionally includes relative chronology and centered review count in the cumulative-link random-intercept model. Relative chronology is zero for the first and one for the last report. Its coefficient concerns creation order conditional on the displayed labels. Four coefficients in this joint model form a separately declared exploratory multiplicity family.

The extended editorial outcome is rejection, so more severe ratings have the same positive orientation as adverse outcomes. Let s_{m2} be one minus the stratum-scaled favorable score, $\bar{s}_{m,-2}$ the corresponding mean peer severity, and $d_{m,-2}$ mean pairwise distance among peers' normalized ordinal ranks. With one peer, pairwise distance is set to zero and review count remains controlled. Equation 6 adds a reviewer-severity-by-disagreement interaction to the earlier associational specification; λ_s is a venue-year intercept.

$$\text{logit}\Pr(J_m = 1) = \lambda_s + \kappa_2 s_{m2} + \kappa_o \bar{s}_{m,-2} + \kappa_d d_{m,-2} + \kappa_{2d} s_{m2} d_{m,-2} + \kappa_n n_m. \quad (6)$$

Here J_m denotes recorded rejection. Standardization compares predictions at the observed first and third quartiles of second-reviewer severity while retaining other observed covariates and recomputing the interaction. Delta-method intervals use the robust covariance, following the general sandwich-variance principle of White (1980). Marginal contrasts within disagreement or minority-status subgroups describe selected populations and are not causal effects of moving a manuscript between those groups. A separate complete-profile model compares labels one, two, and three-plus on the same manuscripts. The third-plus predictor averages all such reviews; it is not equivalent to a single individual's score. Expanded editorial scope allows unlabeled peers when a scored label two is observable, with exclusions and sample counts reported.

Predictive evaluation fixes a test set by the SHA256 hash of manuscript identifier: the first eight hexadecimal digits modulo five equal zero. The split never uses scores or outcomes, is shared across three models, and allocates roughly one fifth of papers to testing. Models use an equal-weight score mean, separate R2 and peer scores, or separate scores plus disagreement and its interaction. Brier score, logarithmic loss, AUROC, calibration intercept and slope, and grouped calibration curves assess different aspects of performance (Van Calster et al. (2019)). Paired test-paper bootstraps provide intervals for differences in Brier loss. Snapshot timing can still leak information about the final process state; the evaluation is explicitly retrospective. A sensitivity analysis requires every retained rating to have been last modified before the recorded decision. Fixed cubic score terms check linear coding sensitivity without selecting knots from outcomes.

Review dynamics extend the paired MIDL analysis by holding the exact paired-reviewer set constant within each manuscript and comparing mean pairwise ordered-category distance before and after rebuttal. At least two paired reviews are required. Manuscripts with only one paired review cannot estimate disagreement and are excluded from that analysis. A fall in distance indicates convergence of the recorded assessments, but does not identify improvement in manuscript quality, reviewer learning, or a causal rebuttal effect.

Monte Carlo validation and research status

The simulation design follows the separation of aims, data-generating mechanisms, estimands, methods, and performance measures emphasized by Morris et al. (2019). Thresholds and manuscript heterogeneity are calibrated from the largest eligible empirical ordinal stratum. In the baseline mechanism a latent favorable evaluation is the sum of a manuscript effect, a reviewer-position shift, and logistic error, as in Equation 7. The second-reviewer harshness parameter Δ lowers the favorable latent evaluation when positive.

$$L_{mi} = u_m - \Delta I(i = 2) + \varepsilon_{mi}, u_m \sim N(0, \sigma^2), \varepsilon_{mi} \sim \text{Logistic}(0,1). \quad (7)$$

Ordered thresholds discretize the latent evaluation into the empirical categories. Principal rank-estimator scenarios vary manuscript count, a null or positive harshness shift, unequal reviewer counts, missing ratings, category imbalance, and venue-level location heterogeneity. Missing completely at random is separated from a mechanism in which unfavorable second-reviewer scores are more likely to be absent. The target rank contrast is calculated by numerical integration under the generating mechanism. Bias is therefore evaluated against the ordinal estimand, rather than incorrectly comparing a discretized score contrast with the latent shift parameter.

This reconstruction is exploratory. It was not preregistered before source inspection, and the original small project had already exposed the investigator to preliminary hypotheses and results. Simonsohn (2015) motivates attention to the sensitivity of replication evidence to precision, while Altman and Bland (1995) distinguish failure to reject from affirmative evidence of no effect. Lakens (2017) provides a framework for equivalence questions, but no retrospective equivalence margin is imposed here. The conclusions therefore depend on estimated magnitudes, uncertainty, cross-stratum consistency, and the validity of the position definition, rather than on a binary significant or nonsignificant label.

Four additional mechanism simulations use 1,000 replications each. A symmetric three-category latent model separates dispersion from location; increasing only label-two logistic-error scale generates polarization without a mean shift. Swapping its public label in half the manuscripts tests attenuation of the measured tail contrast. Selecting a low-rating tie breaker from otherwise exchangeable candidates demonstrates how recruitment can generate apparent label-three severity. A symmetric two-of-three endorsement rule constructs an adverse outcome without any special editor weight. These are explicitly hypothetical mechanisms. They test identification and estimator behavior, not whether editors actually follow those processes. A further 1,000 empirical subsamples of 21 ORB manuscripts illustrate precision at the small study's scale; resampling does not create another independent corpus.

Results

Data characteristics and observable positions

The source reconstruction reproduced 62,938 submission rows and 157,082 reviews. Removing three identical PeerJ submissions and their duplicate reviews left 62,935 manuscripts and 157,070 review records. The six core Zstandard tables occupy 22.84 MB. Table 1 separates

archive coverage from the analytical populations. The primary label sample comprises 8,885 manuscripts and 29,166 reviews; creation chronology supplies a broader sample of 25,960 manuscripts and 91,795 reviews. Most primary manuscripts are ICLR conference submissions, limiting generalization to other editorial systems. Missing titles in the broader archive do not affect title completeness in the primary sample.

Table 1. Source coverage and distinct analytical populations

Population	Count
Source submission rows	62,938
Deduplicated manuscripts	62,935
Source / deduplicated reviews	157,082 / 157,070
Primary explicit-label manuscripts / reviews	8,885 / 29,166
Chronology manuscripts / reviews	25,960 / 91,795
Complete-label independent decision manuscripts	5964
Expanded decision scope	6037
Paired disagreement manuscripts / reviews	85 / 235

Public labels and creation order coincide for only 30.1% of eligible reviews, and only 31.8% of explicitly numbered second reviewers are chronologically second. Fig. 2 shows the cross-classification and compares estimands. On the identical 8,885 manuscripts, the chronology contrast is 0.003 (95% CI -0.011 to 0.017), while the paired label-minus-chronology difference is -0.011 (95% CI -0.031 to 0.009). The broader chronology estimate is 0.002 (95% CI -0.006 to 0.010). These small average contrasts do not make the definitions interchangeable: they index different reviews even when their population means are similar. Assignment order and stable reviewer identity remain unobserved.

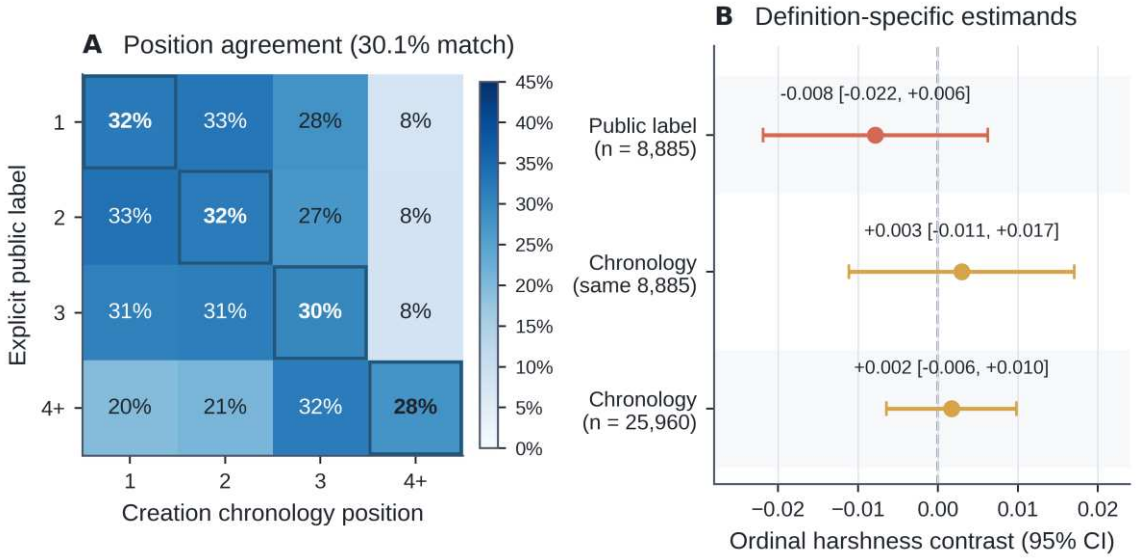


Fig. 2. Explicit public labels and creation chronology identify different reviews. A: row-normalized proportions in the primary sample, grouping positions four and above. B: manuscript-balanced contrasts with 95% intervals; the middle row uses the identical primary manuscripts.

Ordinal harshness and position heterogeneity

The manuscript-balanced explicit-label estimate was -0.008 (95% CI -0.022 to 0.006), P=0.275314. Its 2,000-resample bootstrap interval was -0.022 to 0.006. Positive values would indicate more lower-than-peer ratings; the estimate is small and slightly negative. Fig. 3 displays the largest venue-year strata, with their own uncertainty rather than one common category-scale coefficient. The exact R1-versus-R2 comparison and review-count restrictions also leave their intervals spanning zero. These are estimates of observed public-label patterns, not evidence that all difficult reviewer experiences are imaginary.

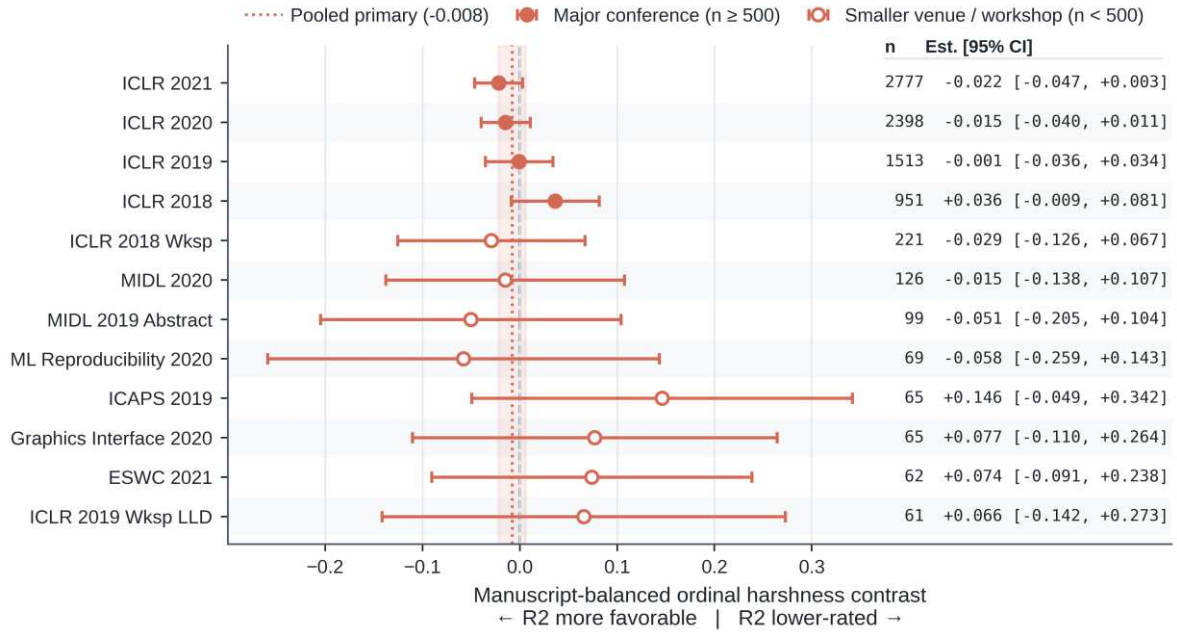


Fig. 3. Venue-year heterogeneity in the manuscript-balanced harshness contrast. The largest nine strata are shown. Positive values mean label two rates below peers more often than above. Intervals are manuscript-level 95% intervals.

The largest stratum, ICLR 2021, contains 2,777 manuscripts and 10,875 reviews. Its joint ordinal model includes labels, relative creation chronology, review count, and a manuscript random intercept. Label two has favorable log odds 0.130 (95% CI 0.032 to 0.228), and label three-plus has favorable log odds 0.096 (0.010 to 0.182). The latter direction differs from the supplied small study's adverse label-three association. Table 2 reports all four coefficients and their exploratory multiplicity adjustment. The joint model converged, its minimum information eigenvalue was 11.006, and the manuscript standard deviation was 1.440. Neither numbering nor timing establishes recruitment purpose. A selected tie-breaker explanation remains a hypothesis, not an observed cause.

Table 2. Joint public-label and chronology ordinal model, ICLR 2021

Term	Favorable log odds	95% interval	*P*	BH P
R2 vs R1	0.130	0.032 to 0.228	0.0093	0.0371
R3+ vs R1	0.096	0.010 to 0.182	0.0292	0.0536
Chronology first to last	0.096	0.004 to 0.187	0.0402	0.0536
Review count (centered)	-0.054	-0.193 to 0.085	0.4469	0.4469

The six original venue-specific fits were independently rerun. The largest-stratum threshold diagnostic gave $P = 0.486173$, after excluding sparse end cutpoints; its marginal nature prevents it from certifying conditional proportional odds. Increasing quadrature to 31 points and changing to a probit link preserved the favorable direction for label two. The rank contrast remains the less scale-dependent primary target when a common location shift is questionable.

Polarization beyond the mean

Fig. 4 shows the full ordinal distribution in ICLR 2021, with manuscript-balanced proportions and intervals. Fig. 5 places original-scale position means beside the broader primary-sample distributional contrasts. It deliberately avoids placing incompatible means from different platforms on a shared axis. Similar means could conceal heavier tails, but the executed contrasts do not show that pattern generally. Across the primary sample, the lowest endpoint contrast was 0.001 (95% CI -0.003 to 0.006), the highest endpoint contrast was 0.001 (95% CI -0.003 to 0.006), and either endpoint was 0.003 (95% CI -0.004 to 0.009). All are R2-minus-peer probabilities. Their small positive point estimates are compatible with modest differences in either direction.

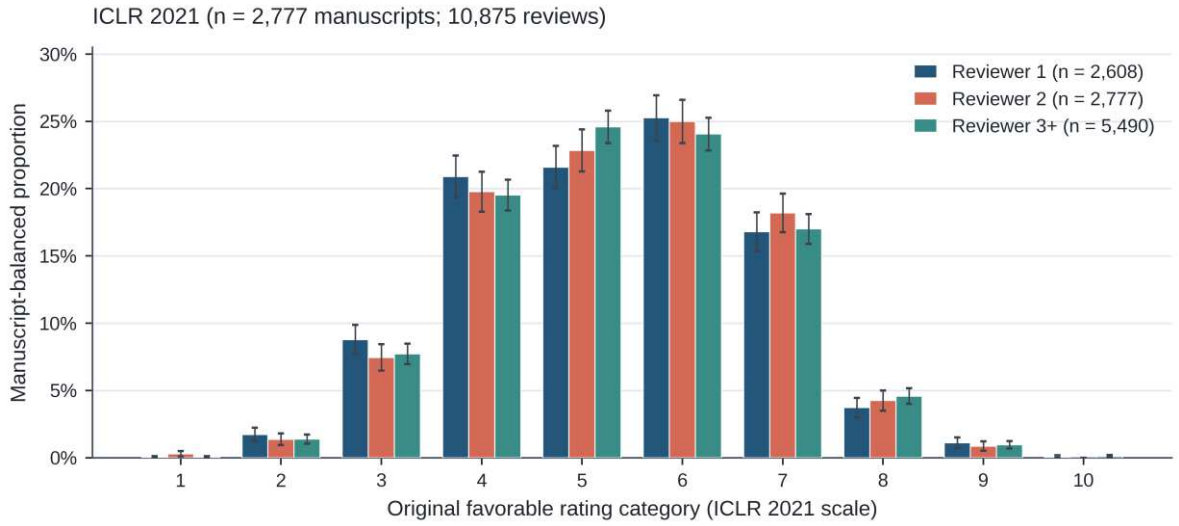


Fig. 4. Complete original-scale recommendation distributions in ICLR 2021. Proportions balance manuscripts within each position group; bars show 95% intervals truncated to the probability range. Label three-plus may contain several reviews per manuscript.

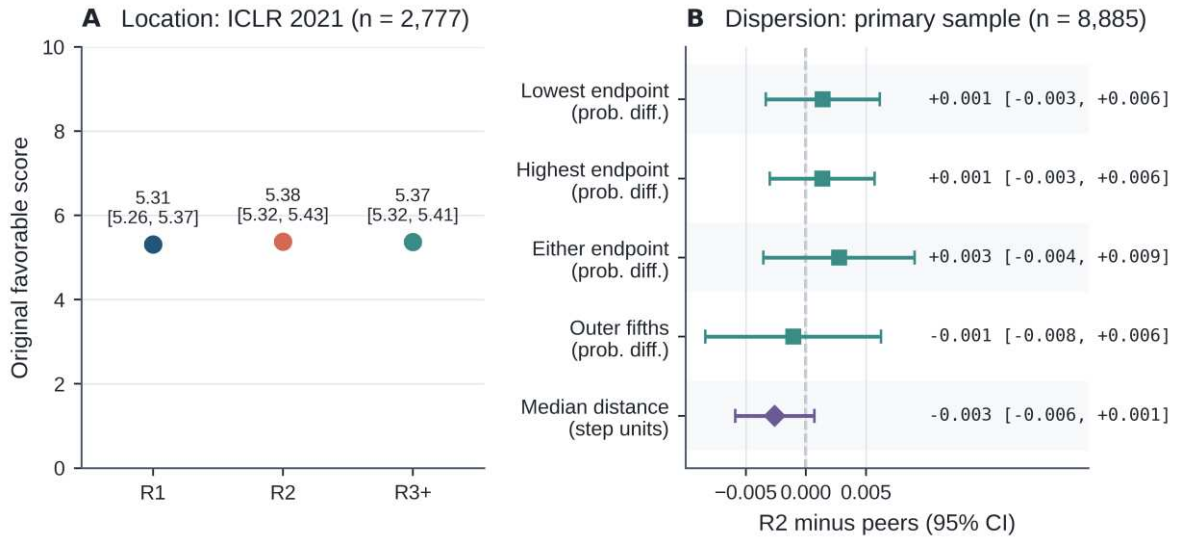


Fig. 5. Location and polarization are distinct targets, but neither shows a general adverse label-two pattern. A: original favorable scores in ICLR 2021 with manuscript-level 95% intervals. B: R2-minus-peer contrasts across all 8,885 primary manuscripts; endpoints and outer fifths are probability differences, while median distance uses normalized ordinal steps.

The broader outer-fifth contrast was -0.001 (95% CI -0.008 to 0.006); normalized median distance was -0.003 (95% CI -0.006 to 0.001). The five-test adjusted probabilities all exceeded 0.60. Table 3 reports the manuscript-bootstrap intervals and makes the distinct units explicit. The within-manuscript label permutation for the outer-fifth contrast gave $P = 0.7821$, conditional on its exchangeability assumption. In adjusted endpoint models, the R2-versus-R1 odds ratio was 0.998 (95% CI 0.894 to 1.113) for either observed endpoint and 0.933 (0.851 to 1.023) for the broader outer fifths. Venue-year, chronology, and review count were included. Thus the smaller study's polarization suggestion was tested rather than adopted, and was not reproduced as a general ORB pattern.

Table 3. Manuscript-balanced R2-minus-peer distributional contrasts

Metric	Estimate	Bootstrap 95% interval	BH P
Lowest endpoint	0.001	-0.003 to 0.006	0.7046
Highest endpoint	0.001	-0.003 to 0.006	0.7046
Either endpoint	0.003	-0.004 to 0.009	0.7046
Outer fifths	-0.001	-0.008 to 0.006	0.7755
Median distance	-0.003	-0.006 to 0.001	0.6099

Inter-reviewer disagreement

Disagreement exists across all positions. In ICLR 2021, ordinal alpha was 0.381800 and its independent library implementation agreed to numerical precision. Quadratic AC2 was substantially higher, reflecting its different weighting and chance reference. Removing label two from the same 2,417 exact-label triples changed alpha by approximately -0.007, with a manuscript-bootstrap interval from -0.032 to 0.016. This neither identifies a uniquely disruptive reviewer nor measures correctness. Changes to marginal frequencies remain part of the removal calculation.

Fig. 6 relates the distance between R2 and peers to disagreement among the peers themselves. Each manuscript contributes once. A reviewer can be isolated when peers agree, while a large peer spread makes several apparently extreme relative positions possible. Distances are normalized ordered-category steps rather than commensurate substantive score gaps across fields. The geometry is descriptive; its cells do not show how an editor responded. The next analysis tests that association explicitly.

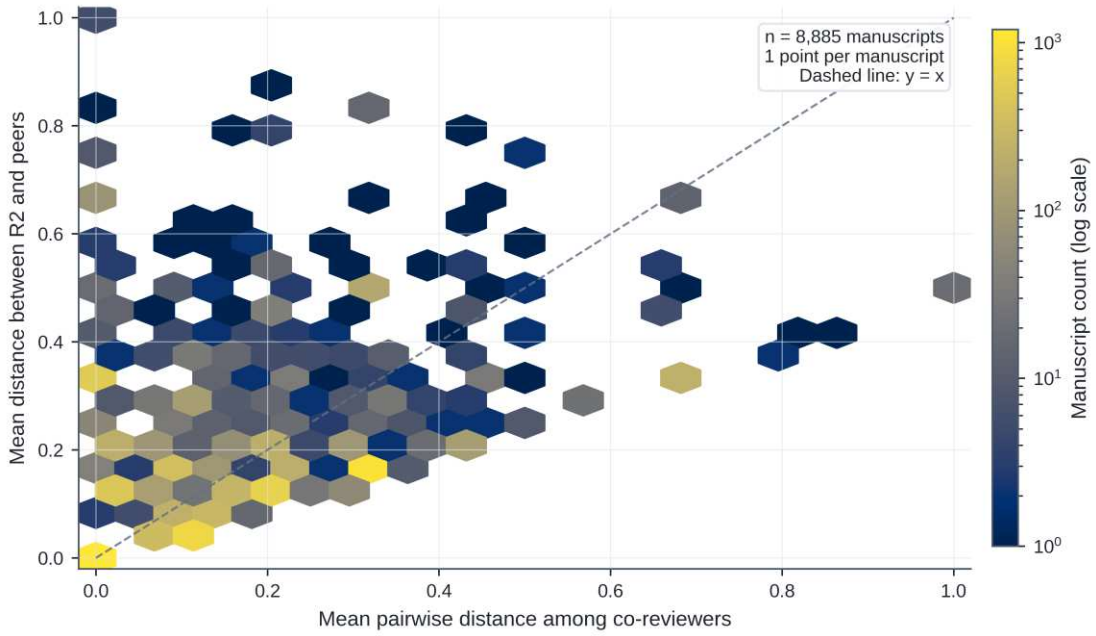


Figure 6. R2-versus-peer distance in relation to disagreement among peers. One point per primary manuscript is aggregated into hexagonal cells. Ordered-category ranks are normalized within venue-year. Manuscripts with one peer have zero peer-pair distance by definition; color indicates counts, not decision risk.

Editorial associations and predictive comparison

The complete-label editorial model includes 5,964 manuscripts: 2,188 accepted and 3,776 rejected. Outcomes come from explicit decision records. An interquartile increase in R2 severity is associated with a rejection probability difference of 0.247 (95% CI 0.227 to 0.267) after standardization over observed covariates. Fig. 7A plots adjusted probabilities at selected levels of peer disagreement. The interaction coefficient is 0.767, $P = 0.636540$; the evidence does not establish that the R2 association becomes stronger under disagreement. Subgroup contrasts are larger among manuscripts where R2 is uniquely lowest, but those groups have different covariate distributions and cannot be interpreted as causal moderation.

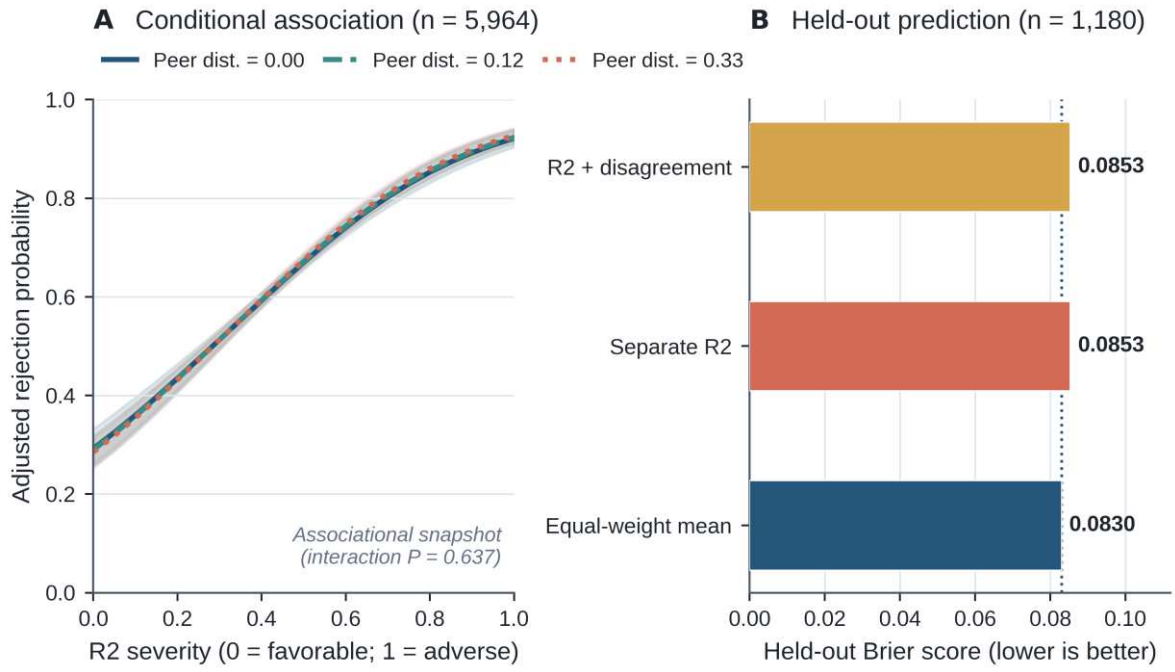


Fig. 7. Conditional editorial association and retrospective held-out prediction. A: adjusted rejection probabilities with pointwise 95% intervals at selected peer-distance quantiles, standardized over 5,964 observed profiles; the tested interaction is inconclusive. B: Brier loss on the same hash-selected test manuscripts. These are archived snapshots, not a prospective policy evaluation.

The same-sample position model has 5,492 complete profiles. Interquartile rejection contrasts are 0.288 for label one, 0.245 for label two, and 0.308 for the average label-three-plus score. These contrasts involve different score ranges and, for the last predictor, an average of several assessments. They are not editor-weight estimates. The expanded scope audit permits unlabeled co-reviewers and includes 6,037 decision manuscripts with complete numerical ratings, yielding 0.243 (95% CI 0.224 to 0.262). The stricter timing subset retains 5,489 papers and gives 0.242 (95% CI 0.221 to 0.263). Fixed cubic score terms give 0.264 (95% CI 0.242 to 0.286). Association is therefore not confined to the complete-label sample or one linear coding, but remains retrospective and observational. A joint venue-year slope test detects heterogeneity ($P = 1.13e - 23$), which rules out interpreting one pooled coefficient as an invariant editorial rule.

The hash split allocates 4,784 manuscripts to training and 1,180 to testing. Table 4 reports performance on exactly the same held-out papers. Equal-weight mean Brier loss was 0.0830, compared with 0.0853 for separate R2 and 0.0853 after adding disagreement. The separate-R2 minus equal-mean difference was 0.0023 (paired bootstrap 95% CI 0.0005 to 0.0042). Separate-R2 AUROC was 0.947; its calibration slope was 0.952. Calibration curves and intercepts are supplied in the supplement. A significant recommendation coefficient did not produce better held-out prediction. Because predictors are archived process snapshots, these scores do not validate prospective editorial decision support.

Table 4. Retrospective prediction on 1,180 held-out manuscripts

Model	Brier	Log loss	AUROC	Calibration slope
Equal-weight mean	0.0830	0.2781	0.949	0.949
Separate R2	0.0853	0.2819	0.947	0.952
R2 with disagreement	0.0853	0.2819	0.948	0.952

Review-round dynamics

MIDL 2022 supplies 247 matched preliminary and explicitly post-rebuttal review records on 97 manuscripts. There are no explicit numeric public labels in this subset. Its chronology-specific second-reviewer signed-change contrast, based on 71 manuscripts with a paired second reviewer and at least one paired peer, was -0.110 (95% CI -0.271 to 0.050), $P = 0.174611$. Thus a general change in assessments should not be described as a demonstrated Reviewer 2 response.

For disagreement, 85 manuscripts retain at least two paired reviewers, totaling 235 paired review records. Fig. 8 holds that reviewer set constant across stages. Normalized pairwise disagreement decreased by -0.121 (95% CI -0.156 to -0.085). Among all 247 paired reports, 52.6% increased their favorable rating, 46.6% were unchanged, and 0.8% decreased. These are separate denominators: the disagreement analysis excludes single-paired-review papers. The archive supports convergence of recorded ratings at this stage, not a causal account of revision, persuasion, selective response, or changes in review quality.

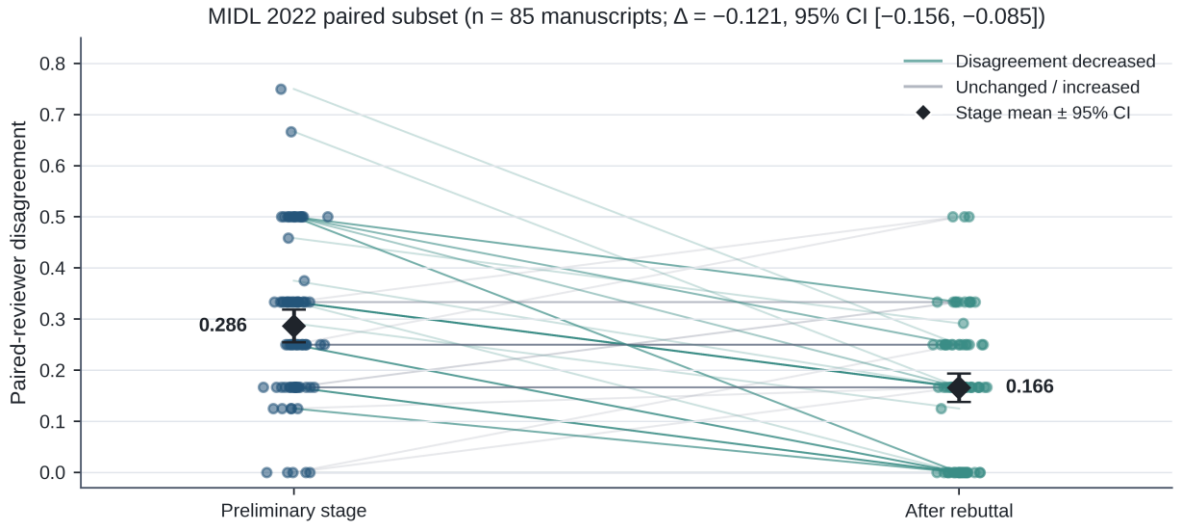


Fig. 8. Paired-review disagreement narrows after the rebuttal stage in MIDL 2022. The exact paired reviewer set is retained within each of 85 manuscripts. Distances are mean pairwise category-step differences divided by four. Lines connect manuscript values; diamonds and bars show stage means and approximate 95% intervals. No causal rebuttal effect is identified.

Cross-Dataset Evidence: Large-Scale Patterns and Small-Sample Mechanisms

As a historical methodological audit of the prior small-sample project (detailed in Supplementary Information Section S6, Table S10, and Figure S4), Table 5 contrasts the primary ORB evidence with the supplied F1000Research case without pooling scales or

outcomes. Because the supplied F1000Research rows remain authenticity-unverified, this comparison serves solely to diagnose identification differences rather than as an independent empirical replication. The latter's 21 articles and 71 reports reproduce a mean-severity comparison with no significant excess for label two, an unadjusted median-centered Levene statistic of 4.290 (P=0.042073), and very low agreement. That variance test treats 71 report rows as independent and assumes numerical distances on a three-category scale. It is insufficient evidence for a robust polarization mechanism. The replicated position-plus-round model uses the original weak ridge penalty and gives an adverse label-three-plus coefficient; that result belongs to a tiny, authenticity-unverified case rather than a validated counterexample to ORB.

Table 5. Cross-dataset evidence and identification limits

Dimension	ORB primary evidence	Supplied F1000 audit	Interpretation
Mean harshness	8,885 papers; rank contrast near zero	No excess in supplied rows	Compatible estimates; authenticity and precision differ
Polarization	Endpoint and distance intervals span zero	Unadjusted Levene P=0.042	Small spread test is not confirmed general polarization
Label three-plus	ICLR 2021 joint model more favorable	Adverse ordinal association	Selection and institutions may differ; recruitment unknown
Disagreement	Matched removal does not isolate R2	Ordinal alpha about 0.012	Marginals and scales differ; no pooled reliability estimate
Editorial association	Explicit decisions; standard sample 5,964	Indexing reconstructed exactly from statuses	Threshold dependence cannot identify editor weight
Dynamics	85 paired MIDL manuscripts; lower disagreement	11/13 supplied transitions soften	Different stages and identity evidence; no causal pooling
Linguistic patterns	No report-text inference	Short excerpts without retrieval history	Criticism, civility and recommendation remain distinct

Eleven of thirteen adjacent-version transitions soften in the supplied rows, while two are unchanged. The apparent 84.6% softening reproduces arithmetically, but continuity of the same reviewer person is not independently established by repeated numerical position. More decisively, every one of the 21 supplied indexing-eligibility outcomes exactly reconstructs from the stated recommendation threshold. Its association with a negative second review is consequently entangled with outcome construction. The reported 44.4-percentage-point adverse-outcome gap cannot establish independent editor behavior. Linguistic claims based on short excerpts lack retrieval provenance and are not replicated as empirical text findings.

Three explanations for differences remain distinguishable. Precision is testable: in 1,000 subsamples of 21 ORB manuscripts, endpoint-contrast estimates ranged from -0.111 to 0.143

over their central 95%, despite a full-sample contrast near zero. Institutional mechanisms are partly observable through different outcome definitions, but there is no common editorial target to meta-analyze. Selection and assignment explanations are plausible because public label, timing, and review count differ, yet invitation reasons are absent. The small case supplies analytical questions; it does not independently verify a cultural mechanism.

Robustness and mechanism simulations

Leave-one-venue-out and leave-one-year-out results, removal of low-volume strata, exact-label comparisons, bootstrap intervals, alternative chronology definitions, and numerical-coding sensitivities are retained in the supplement. Selection on missing observations remains a limitation rather than an imputed solution. The original 10,000-replication simulation suite is retained because its generating mechanisms, ordinal estimands, and convergence accounting are explicit. Informative missingness continues to show that valid-looking complete-case inference can have poor coverage.

The four new scenarios add 4,000 simulation replications. Under symmetric polarization without a location shift, the harshness test rejected zero in 5.3% of replications while the tail test detected the specified dispersion effect in 99.8%. Swapping the second-reviewer label in half of manuscripts attenuated the tail target from 0.127 to 0.032. In the exchangeable null, a symmetric two-of-three endorsement rule generated a second-reviewer adverse-status risk difference of approximately 0.494, while the R2-minus-R1 difference averaged only 0.002. Strong association with a constructed outcome therefore does not require special reviewer weight. The selection scenario likewise demonstrates a possible recruitment artifact; it does not show that ORB editors selected tie breakers that way.

Discussion

The enriched evidence does not support universal Reviewer 2 harshness or a general hidden polarization pattern in the observed public-label population. The two claims were tested separately. An ordinary mean could have concealed more extreme recommendations, but the endpoint, broader-tail, and median-distance contrasts remain close to zero. This makes the empirical conclusion more specific than a binary declaration that the stereotype is true or false. It also prevents an attractive small-sample variance result from becoming the premise of a large-scale narrative before the corresponding large-scale test is performed.

That conclusion does not establish that reviewer-number conventions are harmless or that author experience is mistaken. Public labels are administrative traces, and their disagreement with creation order is substantive. A second-labelled reviewer may have submitted first or last. If a stereotype originates in report arrival, invitation sequence, display layout, or an editor's description, a public-signature analysis targets only part of that experience. Figure 2 demonstrates the mismatch rather than correcting it away through arbitrary relabelling. Future claims about assignment mechanisms require invitation and recruitment records with their actual meanings preserved.

The editorial results identify an association that is expected to matter substantively without attributing special authority to label two. Severe second-reviewer ratings remain associated with rejection after accounting for the mean of peers, but the data do not clearly support the tested disagreement interaction. Comparing one rating coefficient with the coefficient on an average of several ratings would further confound unit size with editorial weight. The complete-profile comparison and the same-paper predictive evaluation address different questions: the former describes conditional score associations, while the latter asks whether separating one labelled review improves retrospective prediction. Neither identifies a causal decision policy.

The equal-weight model's better held-out loss is particularly informative because it separates an explanatory-looking coefficient from predictive usefulness. The improvement is modest in absolute loss, and the intervals describe the fixed training fit on one held-out sample rather than repeated retraining across every possible split. A richer nonlinear or institutional model could perform differently. Even so, the observed result provides no justification for assigning automatic extra decision weight to Reviewer 2. Editors should examine reasons, uncertainty, expertise, and disagreements rather than infer review quality from a label or optimize a policy around one coefficient.

Review convergence deserves its own interpretation. Matched MIDL assessments become less dispersed after the rebuttal stage, while the chronological second review does not show a clearly different direction of change from peers. Collective convergence is a manuscript-level process, whereas a position-specific response concerns who changes more. Convergence could reflect a better manuscript, clarified objections, mutual information, selective retention of responsive reviewers, or shared pressure toward consensus. The current fields cannot separate these possibilities. The records should therefore motivate a design with timestamped score histories and independent measures of substantive revision, not a causal claim that rebuttal makes reviewers more lenient.

The smaller supplied case clarifies an identification problem even though it cannot provide independent empirical validation. Its indexing outcome is reconstructed exactly from the review-status threshold. A negative review can be associated with non-indexing because it enters the rule, without an editor favoring that reviewer. The symmetric-threshold simulation makes this point explicit under exchangeable reviewers. Likewise, the original decomposition's zero residual does not prove that harshness, confounding, and editorial sensitivity have been separately identified; it follows from regression group means. Removing these causal interpretations strengthens the paper by aligning its central claims with observed quantities.

Why, then, might the stereotype persist? Three accounts remain plausible but unproven here. First, an unusually adverse report is salient even when its numbered position is statistically ordinary. Second, authors may judge the quality of a report partly through whether it supports their desired outcome, a distinction highlighted by Goldberg et al. (2025). Third, institutional sequences can place unusual reports in memorable roles: a late specialist, a tie breaker, or a publicly displayed second slot. These explanations require perceptions or recruitment data. The present results constrain numerical accounts of the stereotype; they do not measure recall bias or establish its cultural origin.

The contrast between a dramatic small case and a stable large-scale null also concerns research practice. A three-category distribution on 21 selected articles is highly sensitive to which manuscripts and rounds enter the sample. The ORB subsampling exercise demonstrates imprecision, not that the F1000 records were drawn from the same population. Venue-specific score conventions and deterministic indexing rules add genuine structural differences. Triangulation is useful when those differences are made explicit. Calling repeated source descriptions independent datasets or averaging incompatible outcome contrasts would give an appearance of corroboration while weakening the evidence.

Theoretical and Methodological Contributions

The contribution is a measurement-centered account of reviewer position. It separates public numbering, creation chronology, ordinal location, tail behavior, collective disagreement, conditional decisions, and paired stage change. Each target has a documented population and observation unit. The paired comparison of position definitions shows that equal population means do not imply that two definitions classify the same reports. The joint ordinal model tests

whether apparent label patterns persist after timing adjustment without equating timing with assignment.

A second contribution is to use cross-source disagreement as an identification diagnostic. The supplied F1000 case is not promoted to an independent replication after its provenance fails that standard. Its outcome rule and repeated-position trajectories instead reveal which claims need stronger records. The new simulation scenarios show how polarization, misclassification, selection, and symmetric thresholds can produce different signals. These mechanisms are separated from empirical inference and are not asserted to be the true ORB generating process. The rank contrast, multilevel model, sandwich covariance, and bootstrap remain established methods; integration and traceability, rather than renamed indices, carry the methodological contribution.

Practical Implications

Editorial systems should retain distinct fields for reviewer assignment, public display label, invitation date, report creation, score revisions, and independent decisions. A single overwritten score or enumerated export list cannot answer a temporal or positional question. Public releases can preserve anonymous local labels and precise event semantics without linking reviewer identities across manuscripts. For post-publication platforms, releases should distinguish a deterministic indexing threshold from a discretionary editorial decision.

The results also discourage position-based aggregation rules. A negative recommendation can warrant serious attention without acquiring extra authority from its label. When recommendations conflict, documenting the reasoning behind the decision is more informative than reporting only the final vote count. Evaluation of new aggregation or review protocols should use genuinely prospective snapshots, prespecified outcomes, calibration and loss metrics, and external or later-period validation. Existing metadata support a retrospective audit; they do not justify deployment of a reviewer-weighting algorithm.

Limitations

The primary sample is a selected OpenReview population dominated by ICLR and covers source venue years rather than verified calendar submission dates. Missing or nonnumeric labels, ratings, and hidden outcomes restrict eligibility. Formal invitation order, recruitment reasons, stable reviewer identities, and complete revision histories are unavailable. Meta-review recommendations are not separately recovered as an independent editorial layer. Recorded agreement cannot reveal whether criticism is correct, constructive, or abusive.

The modelling assumptions also delimit interpretation. Manuscript random intercepts and cluster covariance do not eliminate unobserved reviewer selection or cross-manuscript dependence from repeated anonymous persons. Ordered-step distances, original-score scaling, logistic links, and the limited disagreement interaction are explicit working choices. Lack of evidence for that one interaction is not proof that every nonlinear form of editorial moderation is absent. Timing filters reduce one risk of retrospective leakage but do not establish which version of a rating an editor used. The fixed holdout evaluates internal retrospective performance, not transportability to new venues.

The MIDL sample is small and selected, and paired-score availability may depend on response behavior. F1000 source records remain authenticity-unverified; their arithmetic cannot establish cross-platform replication. The literature review is verified and focused rather than a registered systematic review. The project is exploratory, with exposure to earlier results and additional analyses selected after source inspection. No causal mechanism, equivalence margin, universal journal conclusion, or methodological priority is claimed.

Conclusion

Within the observed OpenReview-dominated public-label sample, Reviewer 2 shows no clear general excess in ordinal harshness or polarization. Recommendation severity is associated with rejection, but the tested disagreement interaction is inconclusive and separating Reviewer 2 does not improve retrospective held-out prediction of final archived snapshots over an equal-weight mean. Paired review disagreement narrows after rebuttal in one selected venue without a demonstrated position-specific shift. Because the metadata exclude report text and concentrate on machine-learning conferences, these findings constrain numerical-score claims rather than linguistic hostility or traditional journal workflows. The stereotype is therefore better investigated as a problem of measurement, selection, institutional rules, and author experience than as a universal property of the second-labelled reviewer.

Declarations

Data availability: The fixed ORB release is available at https://doi.org/10.5281/zenodo.13628746 under CC BY 4.0. The accompanying package redistributes compact metadata derivatives with source provenance. It excludes report text and personal reviewer identities; the supplied small-study rows are explicitly marked unverified and separated from primary analyses.

Code availability: The package includes executable source-reconstruction, validation, statistical analysis, simulation, figure, table, and native-Word generation scripts, pinned Python requirements, and a source-to-result inventory. The default analytical run uses compact Parquet files and does not require a multi-gigabyte download. Exact Excel re-authoring requires the documented artifact runtime; the delivered workbooks and their CSV/JSON inputs are independently readable.

Ethics and responsible use: This study analyzes publicly released, de-identified peer-review metadata (ORB v1.2.0) and does not involve human-subject intervention or attempt to identify anonymous reviewers; institutional ethics review is therefore not applicable.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: The author declares no competing interests.

Author contributions: Muh. Akbar Idris: Conceptualization, Methodology, Data Curation, Formal Analysis, Software, Visualization, Writing – Original Draft, Writing – Review & Editing.

Generative AI disclosure: During the preparation of this work the author(s) used Google Gemini in order to improve language fluency, ensure academic formatting, and assist with LaTeX coding. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Supplementary Materials

Supplementary Information provides the mathematical audit, source hierarchy, original and enriched model diagnostics, endpoint and editorial sensitivities, paired dynamics, simulation mechanisms, calibration curves, and cross-source comparison. Machine-readable tables retain exact probabilities, sample sizes, model terms, and all operating-characteristic results.

References

Aczel, B., Barwich, A., Diekman, A. B., Fishbach, A., Goldstone, R. L., Gomez, P., Gundersen, O. E., von Hippel, P. T., Holcombe, A. O., Lewandowsky, S., Nozari, N., Pestilli, F., & Ioannidis, J. P. A. (2025). The present and future of peer review: Ideas, interventions, and evidence. Proceedings of the National Academy of Sciences, 122(5), e2401232121. https://doi.org/10.1073/pnas.2401232121

Agresti, A. (2010). Analysis of Ordinal Categorical Data. Wiley. https://doi.org/10.1002/9780470594001

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. BMJ, 311(7003), 485. https://doi.org/10.1136/bmj.311.7003.485

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bornmann, L., Mutz, R., & Daniel, H. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. PLoS ONE, 5(12), e14331. https://doi.org/10.1371/journal.pone.0014331

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. Biometrics, 46(4), 1171–1178. https://doi.org/10.2307/2532457

Caner, M., & Asma, B. (2026). Gatekeepers or Nightmares? Exploring Academics' Experiences With the Reviewer 2 Phenomenon. Learned Publishing, 39(3), e2075. https://doi.org/10.1002/leap.2075

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213-220. https://doi.org/10.1037/h0026256

De las Cuevas, C. (2026). Reviewer recommendations and editorial outcomes in peer review: a longitudinal analysis of agreement and disagreement across review rounds. Research Integrity and Peer Review, 11(1), 15. https://doi.org/10.1186/s41073-026-00200-7

Efron, B., & Tibshirani, R. (1994). An Introduction to the Bootstrap. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. PLOS ONE, 17(1), e0259238. https://doi.org/10.1371/journal.pone.0259238

Goldberg, A., Stelmakh, I., Cho, K., Oh, A., Agarwal, A., Belgrave, D., & Shah, N. B. (2025). Peer reviews of peer reviews: A randomized controlled trial and other experiments. PLOS ONE, 20(4), e0320444. https://doi.org/10.1371/journal.pone.0320444

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29-48. https://doi.org/10.1348/000711006x126600

Hedeker, D., & Gibbons, R. D. (2006). Longitudinal Data Analysis. Wiley. https://doi.org/10.1002/0470036486

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. Biometrics, 64(2), 627-634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Lakens, D. (2017). Equivalence Tests. Social Psychological and Personality Science, 8(4), 355-362. https://doi.org/10.1177/1948550617697177

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2-17. https://doi.org/10.1002/asi.22784

Malički, M., & Mehmani, B. (2024). Structured peer review: pilot results from 23 Elsevier journals. PeerJ, 12, e17514. https://doi.org/10.7717/peerj.17514

McCullagh, P. (1980). Regression Models for Ordinal Data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 42(2), 109-127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074-2102. https://doi.org/10.1002/sim.8086

Mutz, R., Bornmann, L., & Daniel, H. (2025). Scientific publishing without gatekeeping: an empirical investigation of eLife's new peer review process. Scientometrics, 130(9), 5067-5092. https://doi.org/10.1007/s11192-025-05422-y

Peterson, D. A. M. (2020). Dear Reviewer 2: Go F' Yourself. Social Science Quarterly, 101(4), 1648-1652. https://doi.org/10.1111/ssqu.12824

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. Proceedings of the National Academy of Sciences, 115(12), 2952-2957. https://doi.org/10.1073/pnas.1714379115

Pollitt, A., Taylor, C., Sreenan, N., & Grant, J. (2026). The costs and benefits of research grant funding peer review. F1000Research, 15, 534. https://doi.org/10.12688/f1000research.176989.2

Ragone, A., Mirylenka, K., Casati, F., & Marchese, M. (2013). On peer review in computer science: analysis of its effectiveness and suggestions for improvement. Scientometrics, 97(2), 317-356. https://doi.org/10.1007/s11192-013-1002-z

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. F1000Research, 6, 588. https://doi.org/10.12688/f1000research.11369.2

Severin, A., Strinzel, M., Egger, M., Barros, T., Sokolov, A., Mouatt, J. V., & Müller, S. (2023). Relationship between journal impact factor and the thoroughness and helpfulness of peer reviews. PLOS Biology, 21(8), e3002238. https://doi.org/10.1371/journal.pbio.3002238

Shah, N. B. (2022). Challenges, experiments, and computational solutions in peer review. Communications of the ACM, 65(6), 76-87. https://doi.org/10.1145/3528086

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. Proceedings of the National Academy of Sciences, 112(2), 360-365. https://doi.org/10.1073/pnas.1418218112

Simonsohn, U. (2015). Small Telescopes. Psychological Science, 26(5), 559-569. https://doi.org/10.1177/0956797614567341

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P., & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. Science Advances, 7(2), eabd0299. https://doi.org/10.1126/sciadv.abd0299

Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. Research Integrity and Peer Review, 5(1), 6. https://doi.org/10.1186/s41073-020-00092-1

Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K., & Evans, J. (2018). The sociology of scientific validity: How professional networks shape judgement in peer review. Research Policy, 47(9), 1825-1841. https://doi.org/10.1016/j.respol.2018.06.014

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. Proceedings of the National Academy of Sciences, 114(48), 12708-12713. https://doi.org/10.1073/pnas.1707323114

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Waltman, L., Kaltenbrunner, W., Pinfield, S., & Woods, H. B. (2023). How to improve scientific peer review: Four schools of thought. Learned Publishing, 36(3), 334-347. https://doi.org/10.1002/leap.1544

Watling, C., Ginsburg, S., & Lingard, L. (2021). Don't be reviewer 2! Reflections on writing effective peer review comments. Perspectives on Medical Education, 10(5), 299-303. https://doi.org/10.1007/s40037-021-00670-z

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48(4), 817–838. https://doi.org/10.2307/1912934

Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: promoting transparency in open science. Scientometrics, 125(2), 1033-1051. https://doi.org/10.1007/s11192-020-03488-4

Worsham, C., Woo, J., Zimerman, A., Bray, C. F., & Jena, A. B. (2022). An Empirical Assessment of Reviewer 2. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 59, 00469580221090393. https://doi.org/10.1177/00469580221090393

Szumega, J., Bougueroua, L., Gkotse, B., Jouvelot, P., & Ravotti, F. (2024). The Open Review-Based (ORB) dataset, version 1.2.0. Zenodo. https://doi.org/10.5281/zenodo.13628746

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificManuscriptSupplementaryInformation.docx