Annals of Science and Inquiry
Vol. 2 No. 6 (2026)

Review-Score Dispersion Adds a Small Held-Out Gain in Reconstructing a Conference Decision Label

Adam Johnston[1] · Eric Fisher[2] · Paul Lewis[3]

[1] Switzerland · [2] Ireland · [3] Finland

Published: June 17, 2026 | Article 854

Abstract

Adding review-score dispersion lowered squared probability error from 0.04047 to 0.03720 for 38 held-out papers in a historical PeerRead release. Both logistic models were fitted on 349 other papers and already included mean recommendation and panel size. The added dispersion coefficient was 0.551, with a working 95% interval of −0.512 to 1.613. Its favorable aggregate test difference survived every single-paper deletion, although nine test papers incurred greater loss. A separate development partition showed the same aggregate direction. These results distinguish a small conditional scoring gain from precise estimation of a disagreement effect. They reconstruct a stored institutional acceptance label; unverified review timing and decision validity prevent a prospective or scientific-quality interpretation.

Review-Score Dispersion Adds a Small Held-Out Gain in Reconstructing a Conference Decision Label

Abstract

Adding review-score dispersion lowered squared probability error from 0.04047 to 0.03720 for 38 held-out papers in a historical PeerRead release. Both logistic models were fitted on 349 other papers and already included mean recommendation and panel size. The added dispersion coefficient was 0.551, with a working 95% interval of -0.512 to 1.613. Its favorable aggregate test difference survived every single-paper deletion, although nine test papers incurred greater loss. A separate development partition showed the same aggregate direction. These results distinguish a small conditional scoring gain from precise estimation of a disagreement effect. They reconstruct a stored institutional acceptance label; unverified review timing and decision validity prevent a prospective or scientific-quality interpretation.

A decision label asks a different question from reviewer agreement

Two papers can receive the same average recommendation while their reviewers disagree to different degrees. A panel giving uniformly moderate scores and a panel split between low and high scores can therefore look identical after averaging. Whether the additional disagreement matters depends on the endpoint. Agreement with another reviewer, an institutional acceptance decision, and the eventual scientific value of a paper are different quantities. This study asks about the second, while keeping that distinction visible throughout the analysis.

The research question is whether within-paper recommendation dispersion adds information about a released acceptance label beyond mean recommendation and the number of valid recommendations. We examine both a fitted association and scoring on papers withheld under the source's original partitions. Proposition P1 concerns the dispersion coefficient conditional on the other two predictors. Proposition P2 concerns the change in held-test probability loss when that predictor is added. Neither proposition requires the direction to be positive, and neither treats disagreement as an intervention assigned to a paper.

Peer review is a process with several purposes and institutional arrangements, rather than one numerical measurement instrument (Bornmann, 2011). A computational assessment must therefore state which part of that process it observes. Shah (2022) surveys challenges and experiments in peer review; that broader literature motivates examining a narrow, inspectable component here. Our contribution is a finite, reproducible comparison of two simple numerical summaries. We do not claim that a small historical partition reveals the best conference policy or supplies a benchmark for automated acceptance.

The process also matters when interpreting a label. Lee et al. (2013) review bias in peer review, and Tomkins et al. (2017) study single- versus double-blind reviewing in a different setting. Neither study establishes the source of an association in this release. They explain why a recorded decision cannot automatically serve as an unbiased measure of merit. The design and analysis of the NIPS 2016 review process likewise illustrates the specificity of conference procedures (Shah et al., 2018). We do not transfer that procedure or its findings to the ICLR records analyzed below.

There is a useful adjacent distinction in evaluating judges. Zhao et al. (2022) separate professional, authenticity, and controversy dimensions in a data-driven contest-judge framework. The present outcome is none of those indices: it is the archive's acceptance field. Numerical score spread alone cannot recover professional competence, authenticity, or the source framework's complete evaluation. Making the endpoint explicit prevents a classifier that reconstructs an institutional label from being described as an

estimator of judge quality. This is a measurement boundary, rather than a replication of the contest method.

The difference between explaining a process and predicting its recorded output is central to statistical modeling (Shmueli, 2010). Our two fitted equations have predictive scores, but the available chronology supports only retrospective label reconstruction. A held-out paper is withheld from estimation; that does not establish that its reviews were available before the historical decision. These are separate information requirements, and a favorable result for one cannot repair missing evidence for the other.

A numerical reconstruction of one public release

PeerRead provides a public research resource containing reviews and acceptance decisions, with data organized into venue-specific train, development, and test partitions (Kang et al., 2018). We reuse the complete acquired ICLR 2017 review-JSON section at repository revision 9bb37751781a900cee9e74ec3105997732c8e8e5. The source README identifies the partitions and notes licensing constraints in parts of the broader collection. We preserve the particular source notices and public-research context, without declaring a blanket license for all review text. No manuscript PDFs, author attributes, or review prose enter this analysis.

All 427 acquired source files have their original byte hashes rechecked. A local inventory projects only the source split, the original Boolean acceptance field, recommendation values, meta-review indicators, and identifiers needed to preserve record correspondence. There are 7,270 embedded review objects, of which 3,635 are exact repeated full objects within their respective source files. We retain one representative of each identical object. This rule does not merge two reviews merely because they contain the same score; equal recommendations can remain separate observations.

Of the 3,635 distinct objects, 385 have a meta-review indicator that is true or not explicitly false. Another 1,929 lack a finite integer recommendation in the permitted range from one to ten. The remaining 1,321 recommendations belong to 427 papers, each with at least two valid values. Confidence is neither required nor used. No paper is excluded because of its decision, mean, spread, or model prediction. This complete accounting makes the dependence unit inspectable: a recommendation contributes to a paper summary, while a paper contributes once to evaluation.

The official training partition contains 349 papers and 139 positive acceptance labels. Development contains 40 and 18, respectively; test contains 38 and 15. Development is a secondary score report, not a tuning sample. Source IDs, normalized-title digests, and whole-file digests have no duplicates across the acquired paper records. These checks support record separation within this release. They do not demonstrate independent reviewers, independent scientific topics, or an absence of unrecorded relationships between papers. Dataset documentation should preserve these distinctions rather than reduce provenance to a dataset name (Gebru et al., 2021).

We define the mean as the arithmetic average of valid recommendations and dispersion as their population standard deviation, with divisor equal to the number of recommendations. The latter is a summary of the observed panel, not an unbiased estimate of an unobserved population variance. The calculations treat the released score codes as equally spaced. Stevens (1946) makes the distinction between ordinal and interval scales consequential; Hand (1996) connects statistical operations to measurement meaning. Here equal spacing is an explicit operational assumption. A monotone recoding could alter both summaries, so the fitted coefficients are not representation-invariant measures of disagreement.

Two fixed logistic specifications

The first model, M1, uses an intercept, mean recommendation centered at 5.5, and the number of valid recommendations centered at three. M2 adds the uncentered within-paper standard deviation. For each model, probability is the logistic transformation of its linear predictor. Modeling a binary response through covariates has a long statistical lineage (Cox, 1958). In this application, the binary response remains the stored acceptance field, and its numerical coding does not add a substantive definition of quality.

Both equations are estimated by unpenalized maximum likelihood on the training papers only. We require a full-rank design, convergence within 200 Newton iterations, finite coefficients and covariance, a small terminal score gradient, and no separation or convergence warning. A failed condition would remain a reported failure; the procedure has no silent penalized fallback. The actual fits satisfy these checks. Odds ratios are exponentiated logit coefficients, with intervals transformed on that same scale, following the distinction between coefficient and odds interpretations in logistic regression (Hosmer et al., 2013).

Coefficient uncertainty uses the uncorrected empirical-score sandwich, HC0. White (1982) distinguishes inference under a working likelihood from the stronger requirement that its distributional specification be true. Our interval additionally depends on an independent-paper working model. Unknown shared reviewers and common institutional processes are not repaired by a sandwich label. The resulting normal Wald intervals describe that declared calculation; they are not causal intervals or validated coverage for future conferences. We report the full coefficients and their uncertainty instead of turning a threshold crossing into the sole finding.

The design was frozen after source and structural inspection but before fitting the dispersion comparison. It was not externally preregistered or outcome blind. A saved protocol identifies the models, official partitions, scoring rule, and deletion sensitivity. Advance specification makes the relationship between intended and performed analysis inspectable (Wagenmakers et al., 2012). We preserve both fitted specifications without selecting between penalties, transformations, panel cutoffs, or alternative test subsets. The concern that model selection can overfit an evaluation criterion is particularly relevant with only 38 primary test papers (Cawley and Talbot, 2010).

Two information checks remain distinct. Fitting never reads development or test labels, and no preprocessing constant is learned from them. This addresses one familiar form of leakage (Kaufman et al., 2012). However, the original record does not authenticate each review's availability before the decision. Kapoor and Narayanan (2023) emphasize the consequences of leakage for scientific claims; simply retaining a train/test split would not justify a prospective prediction claim here. We therefore describe what was executed as retrospective scoring on source-partitioned records.

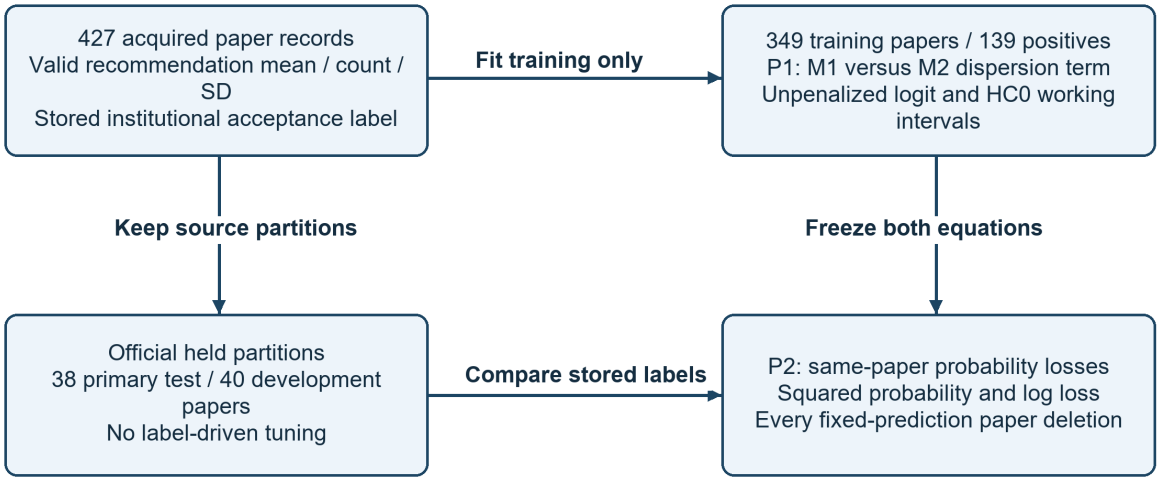

Figure 1. P1 compares the conditional dispersion term in the training model. P2 compares frozen probabilities on the same held-test papers, with development reported separately. Arrows indicate computation and information flow, not causal effects or a historically verified decision timeline.

Loss, discrimination, and the scope of deletion sensitivity

The primary probability loss is the mean squared difference between predicted probability and the stored zero-or-one label, using the binary event-component convention of the Brier score (Brier, 1950). Lower values are better for this declared target. M0 is the constant training acceptance frequency, 139 divided by 349. M1 and M2 use the same papers in every comparison. A difference therefore concerns their probability assignments, not a change in eligibility or an operational decision threshold.

Proper scoring rules make probability assessment different from a count of correct thresholded classifications (Gneiting and Raftery, 2007). We also report mean logarithmic loss, calculated from the linear predictor with a stable expression rather than clipping probabilities. Both losses use the full probability. The choice of a point forecast or functional likewise belongs to the definition of the evaluation target (Gneiting, 2011). No acceptance cutoff, rejection cost, capacity constraint, or utility function is estimated from these data.

The primary effect is M2 loss minus M1 loss on the 38 test papers. We then remove each test paper in turn while holding both fitted models and every remaining prediction fixed. The minimum and maximum resulting means form a deterministic composition range. They are not a confidence interval, a bootstrap distribution, an influence analysis of refitted training coefficients, or evidence from 38 independent repetitions. Development is evaluated analogously. Training scores remain explicitly in sample and cannot provide an additional independent validation set.

We also retain a separation diagnostic already saved by the analysis: the mean predicted probability among positive labels minus that among negative labels, as described by Tjur (2009). It is an auxiliary description, not a prespecified selection criterion. It must not replace a proper probability loss. The probability-score partition of Murphy (1973) further shows why an aggregate loss difference cannot by itself identify a change in calibration or resolution. No such decomposition or calibration curve is fitted here. The distinction between forecast quality and decision value is broader still (Murphy, 1993); a smaller numerical loss does not establish a better institutional policy.

A favorable aggregate score with an imprecise dispersion coefficient

Table 1 reports every coefficient in both fits. The dispersion coefficient in M2 is 0.5505 with HC0 standard error 0.5420 and a working interval from -0.5119 to 1.6129. Its odds ratio is 1.734, with interval 0.599-5.017. The interval includes values corresponding to either direction of association. It does not establish absence of a relationship, and the positive estimate does not establish that controversy causes acceptance. Mean recommendation has a large fitted relationship with the institutional label in both equations; that is unsurprising when the recorded process can itself use reviewers' recommendations.

Table 1. Training coefficients for the two nested logistic models.

Model	Term	Logit coefficient [working 95% interval]	HC0 SE	Odds ratio [working 95% interval]
M1	Intercept	-2.6821 [-3.4504, -1.9139]	0.3920	0.068 [0.032, 0.148]
M1	Mean score minus 5.5	4.4255 [3.4419, 5.4090]	0.5018	83.551 [31.248, 223.401]
M1	Valid recommendations minus 3	0.0876 [-1.6410, 1.8162]	0.8820	1.092 [0.194, 6.149]
M2	Intercept	-3.1655 [-4.5066, -1.8245]	0.6842	0.042 [0.011, 0.161]
M2	Mean score minus 5.5	4.5763 [3.4524, 5.7001]	0.5734	97.150 [31.576, 298.902]
M2	Valid recommendations minus 3	-0.0272 [-1.5795, 1.5252]	0.7920	0.973 [0.206, 4.596]
M2	Within-paper population SD	0.5505 [-0.5119, 1.6129]	0.5420	1.734 [0.599, 5.017]

Table 1 note: Both unpenalized logit fits use the same 349 training papers and 139 positive labels. Intervals use HC0 empirical-score covariance and normal critical values under the independent-paper working model. Odds ratios refer to one operational score point, one recommendation, or one standard-deviation point as named; centering changes the intercept interpretation. McFadden R2 and log likelihood are descriptive in-sample diagnostics.

The larger equation improves training log likelihood from -62.5053 to -61.9083, and McFadden R2 from 0.73361 to 0.73615. These are not held-out effects. The addition of one coefficient does not require an improvement on withheld papers. Table 2 therefore keeps training, development, and primary test scores separate, with the constant benchmark visible. The two review-summary models score much better than the constant on this acquired release, but the question of interest is the much smaller difference between them.

Table 2. Probability scores against the stored acceptance label.

Partition	Papers / positive labels	Model	Squared probability loss	Logarithmic loss
train	349 / 139	M0	0.239653	0.672308
train	349 / 139	M1	0.056198	0.179098
train	349 / 139	M2	0.055332	0.177388
dev	40 / 18	M0	0.250175	0.693650
dev	40 / 18	M1	0.080943	0.254608
dev	40 / 18	M2	0.078432	0.243825
test	38 / 15	M0	0.238932	0.670846
test	38 / 15	M1	0.040465	0.193543
test	38 / 15	M2	0.037198	0.188280

Table 2 note: Each row uses all eligible papers in the named source partition. Models are trained only on the 349 training papers; M0 is their positive-label frequency. Squared probability loss and logarithmic loss are unitless. Training rows are in sample; development was not used to tune the models.

For test, squared probability loss falls from 0.040465 to 0.037198, a difference of -0.003267. Logarithmic loss also falls, from 0.193543 to 0.188280. The direction is consistent across these two reported criteria, without making either a measure of scientific quality. The auxiliary coefficient of discrimination changes from 0.81973 to 0.82874 on test. This additional summary describes separation of the same stored labels, rather than an independent outcome or a reason to select another specification.

Table 3. Paired M2-minus-M1 loss differences and fixed-prediction deletion ranges.

Partition	Papers	Mean difference	Delete-one minimum	Delete-one maximum	Lower / equal / higher individual losses
train	349	-0.000865	-0.001214	-0.000169	235 / 0 / 114
dev	40	-0.002512	-0.005661	-0.000992	28 / 0 / 12
test	38	-0.003267	-0.003651	-0.001686	29 / 0 / 9

Table 3 note: A negative paired difference favors M2. Every deletion removes one scored paper without refitting. The ranges measure deterministic record composition only and are not confidence intervals. Training is shown as a labeled descriptive comparison, not additional validation.

All 38 one-paper test deletions retain a negative aggregate difference, ranging from -0.003651 to -0.001686. This rules out dependence of its sign on retaining any single test paper under the fixed-prediction calculation. It does not make the gain uniform: 29 papers have lower loss, nine have higher loss, and none has exactly equal loss. The development difference is -0.002512, also negative under every one-paper deletion, but those 40 records remain another small partition of the same conference release.

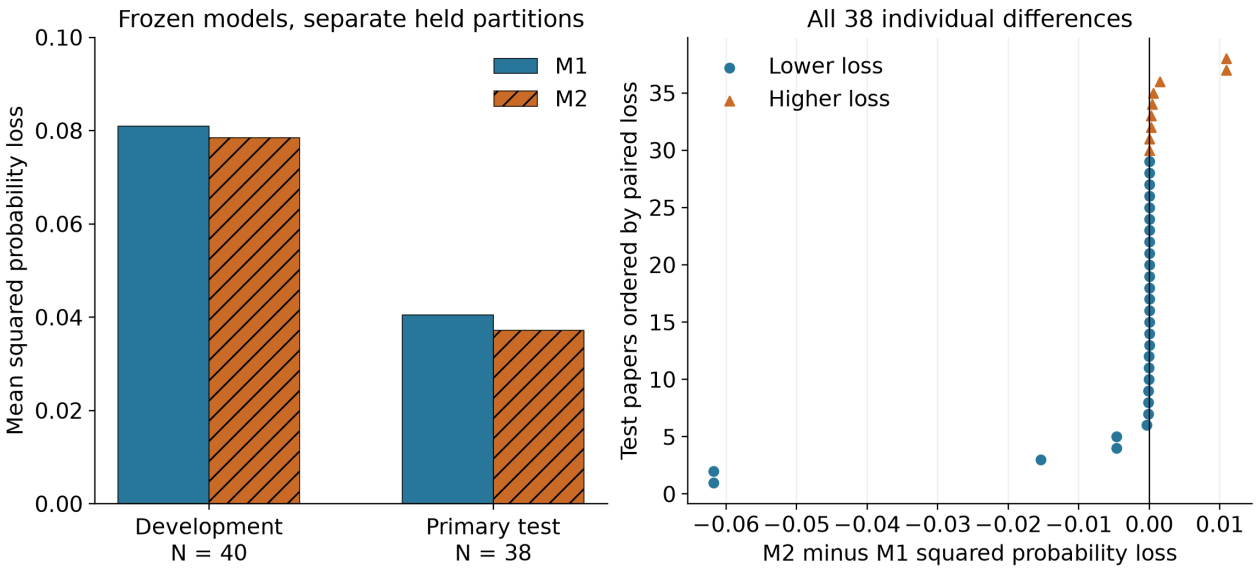

Figure 2. Left: M1 and M2 squared probability losses, with development and primary test separated. Right: all 38 primary test paper differences; values above zero show adverse changes. Zero is a comparison reference, not a significance threshold. There are no sampling uncertainty bars in this figure.

What the institutional label cannot validate

The coefficient and the held-out score answer different questions. The former is an estimated conditional parameter in a working equation. The latter averages losses of two fixed prediction rules on a particular paper set. A wide coefficient interval can coexist with a small favorable test difference. Treating one as a substitute for the other would obscure the distinction between parameter estimation and empirical evaluation. The deletion range adds information about this set's composition, while leaving uncertainty about another release unresolved.

Anonymous feedback also raises a distinction between summarizing responses and validating the decision for which they are used. Yuan et al. (2026) describe scale validation, uncertainty-aware group comparisons, and visual verification in an anonymous-experience framework intended to support management decisions. The primary abstract concerns a private multi-year learning-community source, which we do not acquire or analyze. Our acceptance field already records an institutional decision. Improved reconstruction of that field cannot demonstrate valid experience measurement, sound group decisions, or privacy guarantees. This specific contrast limits the conclusion that a favorable classifier score would otherwise invite.

Several limits are structural rather than computational. Recommendation codes need not have equal psychological distances. The recorded number of valid scores is not necessarily the number of assigned reviewers, and exact-object deduplication does not establish independent people. Review revisions, discussion, and decisions may occur in an unresolved order. The score summaries could reflect a process closely linked to acceptance itself. Consequently, neither the held-out partition nor the sandwich calculation separates reviewer information from the institutional mechanism that produced the target.

The comparison is also modest in scope: two low-dimensional equations, one historical venue section, and one realized official split. It does not rank modern acceptance-prediction systems or recommend using a dispersion coefficient to make editorial decisions. Stronger prospective claims would require authenticated information timing and a target appropriate to the intended use. A fresh conference cohort would answer a transport question that cannot be manufactured by treating the 427 source files as repeated independent conferences.

Computational record and conclusion

The preserved record contains source-file hashes, the minimal numeric paper table, the fixed protocol, all predictions, per-paper losses, coefficient covariance, and every deletion. NumPy and SciPy support the primary numerical calculations (Harris et al., 2020; Virtanen et al., 2020), and the fitted implementation uses Statsmodels (Seabold and Perktold, 2010). An independent standard-library implementation rereads the raw numeric fields and solves the likelihood equations separately. It agrees on all 427 paper summaries and probabilities, with coefficient differences below 0.000000001. Separate checks reconcile every saved loss summary and the interval-to-odds transformations.

That computational agreement addresses reproducibility of this calculation, consistent with the distinction emphasized by Peng (2011). It does not supply independent reviews or validate the institutional outcome. Scientific Agent Skills informed evidence organization and the manuscript consistency workflow (Kassis et al., 2026); the software citation is procedural disclosure, not empirical support for the model. Accountable-author source verification, declarations, and submission approval remain pending. The local research bundle is retained without claiming a public deposit or unrestricted redistribution of original review material.

Within this finite release, dispersion adds a small favorable held-out probability-score difference beyond mean recommendation and panel size. The estimated conditional dispersion term remains imprecise, and some individual losses increase. Reporting these facts together gives the result its useful meaning: a reproducible incremental comparison against a specified historical label, with no inference that disagreement improves science or that the learned rule should govern conference decisions.

References

1. Murphy, Allan H.. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting. https://doi.org/10.1175/1520-0434(1993)008<0281:wiagfa>2.0.co;2
2. White, Halbert. (1982). Maximum Likelihood Estimation of Misspecified Models. Econometrica. https://doi.org/10.2307/1912526
3. Hosmer, David W.; Lemeshow, Stanley; Sturdivant, Rodney X.. (2013). Applied Logistic Regression. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9781118548387
4. Yuan, Yuewei; Huang, Jialei; Yin, Jiayang; Xu, Tianyi. (2026). Visual Verification of user Experience Differences under Privacy Constraints in Anonymous Feedback Data. SSRN working paper 7393298. https://doi.org/10.2139/ssrn.7393298.
5. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.
6. Lee, Carole J.; Sugimoto, Cassidy R.; Zhang, Guo; Cronin, Blaise. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology. https://doi.org/10.1002/asi.22784
7. Kassis, Timothy; Agarwal, Vinayak; He, Yuhuan; Patel, Darshil; Brueckner, Aubrey M.. (2026). Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv preprint. https://arxiv.org/abs/2609.00065
8. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
9. Tomkins, Andrew; Zhang, Min; Heavlin, William D.. (2017). Reviewer bias in single- versus double-blind peer review. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1707323114
10. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
11. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437

12. Tjur, Tue. (2009). Coefficients of Determination in Logistic Regression Models-A New Proposal: The Coefficient of Discrimination. The American Statistician. https://doi.org/10.1198/tast.2009.08210

13. Kang, Dongyeop; Ammar, Waleed; Dalvi, Bhavana; van Zuylen, Madeleine; Kohlmeier, Sebastian; Hovy, Eduard; Schwartz, Roy. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1149

14. Brier, Glenn W.. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

15. Shmueli, Galit. (2010). To Explain or to Predict?. Statistical Science. https://doi.org/10.1214/10-sts330

16. Virtanen, Pauli; Gommers, Ralf; Oliphant, Travis E.; Haberland, Matt; Reddy, Tyler; Cournapeau, David; Burovski, Evgeni; Peterson, Pearu; Weckesser, Warren; Bright, Jonathan; van der Walt, Stéfan J.; Brett, Matthew; Wilson, Joshua; Millman, K. Jarrod; Mayorov, Nikolay; Nelson, Andrew R. J.; Jones, Eric; Kern, Robert; Larson, Eric; Carey, C J; Polat, İlhan; Feng, Yu; Moore, Eric W.; VanderPlas, Jake; Laxalde, Denis; Perktold, Josef; Cimrman, Robert; Henriksen, Ian; Quintero, E. A.; Harris, Charles R.; Archibald, Anne M.; Ribeiro, Antônio H.; Pedregosa, Fabian; van Mulbregt, Paul; SciPy 1.0 Contributors; Vijaykumar, Aditya; Bardelli, Alessandro Pietro; Rothberg, Alex; Hilboll, Andreas; Kloeckner, Andreas; Scopatz, Anthony; Lee, Antony; Rokem, Ariel; Woods, C. Nathan; Fulton, Chad; Masson, Charles; Häggström, Christian; Fitzgerald, Clark; Nicholson, David A.; Hagen, David R.; Pasechnik, Dmitrii V.; Olivetti, Emanuele; Martin, Eric; Wieser, Eric; Silva, Fabrice; Lenders, Felix; Wilhelm, Florian; Young, G.; Price, Gavin A.; Ingold, Gert-Ludwig; Allen, Gregory E.; Lee, Gregory R.; Audren, Hervé; Probst, Irvin; Dietrich, Jörg P.; Silterra, Jacob; Webber, James T; Slavič, Janko; Nothman, Joel; Buchner, Johannes; Kulick, Johannes; Schönberger, Johannes L.; de Miranda Cardoso, José Vinícius; Reimer, Joscha; Harrington, Joseph; Rodríguez, Juan Luis Cano; Nunez-Iglesias, Juan; Kuczynski, Justin; Tritz, Kevin; Thoma, Martin; Newville, Matthew; Kümmerer, Matthias; Bolingbroke, Maximilian; Tartre, Michael; Pak, Mikhail; Smith, Nathaniel J.; Nowaczyk, Nikolai; Shebanov, Nikolay; Pavlyk, Oleksandr; Brodtkorb, Per A.; Lee, Perry; McGibbon, Robert T.; Feldbauer, Roman; Lewis, Sam; Tygier, Sam; Sievert, Scott; Vigna, Sebastiano; Peterson, Stefan; More, Surhud; Pudlik, Tadeusz; Oshima, Takuya; Pingel, Thomas J.; Robitaille, Thomas P.; Spura, Thomas; Jones, Thouis R.; Cera, Tim; Leslie, Tim; Zito, Tiziano; Krauss, Tom; Upadhyay, Utkarsh; Halchenko, Yaroslav O.; Vázquez-Baeza, Yoshiki. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. https://doi.org/10.1038/s41592-019-0686-2.

17. Shah, Nihar B.. (2022). Challenges, experiments, and computational solutions in peer review. Communications of the ACM. https://doi.org/10.1145/3528086

18. Cox, D. R.. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

19. Hand, D. J.. (1996). Statistics and the Theory of Measurement. Journal of the Royal Statistical Society. Series A (Statistics in Society). https://doi.org/10.2307/2983326

20. Wagenmakers, Eric-Jan; Wetzels, Ruud; Borsboom, Denny; van der Maas, Han L. J.; Kievit, Rogier A.. (2012). An Agenda for Purely Confirmatory Research. Perspectives on Psychological Science. https://doi.org/10.1177/1745691612463078

21. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804

22. Cawley, Gavin C.; Talbot, Nicola L. C.. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research. https://jmlr.org/papers/v11/cawley10a.html

23. Shah, Nihar B.; Tabibian, Behzad; Muandet, Krikamol; Guyon, Isabelle; Luxburg, Ulrike von. (2018). Design and Analysis of the NIPS 2016 Review Process. Journal of Machine Learning Research. https://www.jmlr.org/papers/v19/17-511.html

24. Bornmann, Lutz. (2011). Scientific peer review. Annual Review of Information Science and Technology. https://doi.org/10.1002/aris.2011.1440450112

25. Murphy, Allan H.. (1973). A New Vector Partition of the Probability Score. Journal of Applied Meteorology. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

26. Kaufman, Shachar; Rosset, Saharon; Perlich, Claudia; Stitelman, Ori. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data. https://doi.org/10.1145/2382577.2382579

27. Zhao, Runhao; Guo, Yunxiao; Ma, Xiaying; Yin, Xinrong; Tang, Jiuyang. (2022). Research on Evaluation System of the Judge Quality in Students' Contest Based on Data Driven. 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM). https://doi.org/10.1109/icekim55072.2022.00233.

28. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

29. Seabold, Skipper; Perktold, Josef. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-011

30. Gneiting, Tilmann. (2011). Making and Evaluating Point Forecasts. Journal of the American Statistical Association. https://doi.org/10.1198/jasa.2011.r10138