Studies in Knowledge and Innovation
Vol. 2 No. 4A (2026)

Confidence Weights Slightly Raise Excluded-Peer Recommendation Error in a Historical Review Release

Cameron Porter[1] · Jack Brooks[2] · James Harrison[3]

[1] Canada · [2] Germany · [3] United Kingdom

Published: April 30, 2026 | Article 841

Abstract

Does a reviewer's reported confidence help combine other reviewers' recommendations? Fixed weighting rules were compared in 416 ICLR2017 paper records, using 1,287 complete reviews after collapsing identical source objects. Each review score was excluded before its peers predicted it. Equal weighting gave an absolute error of 0.9271 score points; linear confidence weights raised it to 0.9355, and squared confidence weights raised it further. Every single-paper deletion retained the direction of the overall linear-weight difference. Nevertheless, 124 papers improved, while 158 worsened and 134 were unchanged. The reference is another recommendation, so these results concern agreement among released scores. They neither identify better reviewers nor measure scientific correctness.

Confidence Weights Slightly Raise Excluded-Peer Recommendation Error in a Historical Review Release

Abstract

Does a reviewer's reported confidence help combine other reviewers' recommendations? Fixed weighting rules were compared in 416 ICLR2017 paper records, using 1,287 complete reviews after collapsing identical source objects. Each review score was excluded before its peers predicted it. Equal weighting gave an absolute error of 0.9271 score points; linear confidence weights raised it to 0.9355, and squared confidence weights raised it further. Every single-paper deletion retained the direction of the overall linear-weight difference. Nevertheless, 124 papers improved, while 158 worsened and 134 were unchanged. The reference is another recommendation, so these results concern agreement among released scores. They neither identify better reviewers nor measure scientific correctness.

1. An additional field is not automatically a better weight

A recommendation score and a confidence score answer different questions. One expresses an assessment of a submission; the other records a reviewer’s stated assurance about that assessment. A numerical aggregation rule can give the second field considerable influence without demonstrating that its values predict another reviewer’s judgment. The practical question is therefore narrower than whether confident reviewers know more. When their recommendations receive greater arithmetic weight, does the resulting combination more closely reproduce a recommendation that was kept outside that calculation?

Historical peer-review releases make that question observable without collecting new judgments. PeerRead supplies review records alongside submission information, including numerical recommendation and confidence fields in its ICLR2017 section (Kang et al., 2018). The present analysis uses those numeric fields and their within-paper organization. It does not read submission content to adjudicate its contribution. Each recorded score can instead serve as an excluded opinion against which a fixed combination of the remaining scores is compared. The resulting error is measured in recommendation points, with no external correctness interpretation.

Research on review processes has long separated institutional decisions from the observations that precede them. Bornmann’s review describes the broader evaluation setting, while the analysis of the NIPS2016 process illustrates why conference-specific procedures require empirical examination (Bornmann, 2011; Shah et al., 2018). Computational work on peer review likewise treats the design of evaluation as a collection of distinct problems (Shah, 2022). Here the research question is one specific combination problem: does a linear mapping from recorded confidence to positive peer weights reduce excluded-review recommendation error on the same papers?

An evaluator-quality framework can motivate this question without supplying its answer. Zhao and colleagues construct professional, authenticity and controversy indices for scoring in students’ contests (Zhao et al., 2022). Those are different constructs from a self-reported confidence value. None of those indices is reconstructed here, and an excluded conference recommendation is not substituted for contest-quality truth. The comparison asks whether one available field earns a limited operational role in score combination, rather than treating its favorable name as evidence that it measures competence.

2. Why the combination might help, and why it might not

Combining judgments is an established problem in forecast research. Bates and Granger study forecast combination, and Clemen reviews approaches and applications (Bates and Granger, 1969; Clemen, 1989). These provide a reason to examine how inputs are combined, but they do not imply that every additional weighting field improves every target. Equal weights are a meaningful comparison rule precisely because they require no conversion of a confidence category into a relative contribution. The study fixes that comparison before examining its errors.

There are at least two plausible relationships between assurance and an excluded score. If greater assurance accompanies information shared by other reviewers, increasing that review's weight could improve agreement. If assurance instead reflects a reviewer's distinctive interpretation, reweighting could move the estimate away from another opinion. Koriat's self-consistency model explicitly connects confidence with consistency of accessible representations (Koriat, 2012). Work on extreme confidence, determinants of confidence and multiple meanings of overconfidence also cautions against treating assurance as a single universal accuracy quantity (Fischhoff et al., 1977; Griffin and Tversky, 1992; Moore and Healy, 2008). These possibilities motivate an empirical comparison without imposing an expected sign.

The numerical mapping requires its own declaration. Confidence categories from one to five are ordered self-reports, not stated inverse error variances. A weight of four gives a score twice the contribution of a weight of two under the linear rule. Squaring the categories makes that ratio four. Neither interpretation follows merely from their order. Measurement theory distinguishes the assignment of numbers from the operations those numbers support (Stevens, 1946; Hand, 1996). Consequently, linear confidence is an operational policy under examination, and squared confidence is a fixed sensitivity policy, not an alternative psychological measurement scale.

Advice-taking and judgment-aggregation research also gives the equal-weight reference substantive importance. Larrick and Soll examine intuitions about averaging, and Soll and Larrick analyze how people revise judgments using others' opinions (Larrick and Soll, 2006; Soll and Larrick, 2009). Dawes discusses the usefulness of simple linear combinations (Dawes, 1979). The present calculation addresses none of their individual experimental effects. It uses their broader combination problem to justify comparing a transparent simple rule with an intuitively attractive unequal-weight rule, holding the available opinions fixed.

3. Source objects, prediction rules and analysis units

The official PeerRead repository was pinned to commit 9bb37751781a900cee9e74ec3105997732c8e8e5. Only the root documentation and normalized ICLR2017 review JSON files were checked out. Each downloaded review file was compared with its Git blob identifier and separately hashed. The source contains 427 paper files across its released train, development and test directories. These directories are pooled for this finite-record analysis; they are not used as model-fitting and external-testing partitions. No manuscript PDF, parsed manuscript or repository model was executed.

The normalized files contain 7,270 review objects. Within each paper, every complete object occurs exactly twice, leaving 3,635 distinct whole-object representations. A representative is retained for each exact canonical JSON identity, including its original comment and metadata in the identity calculation. This step does not merge reviewers across papers or declare similar comments equivalent. It prevents an identical source copy from remaining among the peers when its counterpart is excluded. The analysis concerns these source representations, without reconstructing the historical reason for the duplication.

Among the distinct non-meta objects, 1,321 contain an eligible recommendation between one and ten. Twelve lack a complete confidence value between one and five. Requiring at least three complete score-confidence pairs per paper leaves 1,287 reviews in 416 papers; eleven paper files are outside that final support. The released counts are used directly. In particular, the 1,321 scored-object count differs from the 1,304 reviews reported in the dataset paper's summary table. That difference is documented rather than repaired by deleting otherwise eligible records to reproduce a historical headline.

For a paper with at least three eligible reviews, one recommendation is excluded at a time. The other recommendations are combined using weights equal to one, their recorded confidence, or confidence squared. A weighted mean is the sum of peer score times peer weight divided by the sum of peer weights. The excluded review's score and confidence do not enter its observed-policy prediction. All weights are positive, so these predictions remain within the range of the included peer recommendations. No weight parameter is fitted or selected.

The main loss is the absolute difference between the predicted and excluded recommendation. Squared error is a declared secondary loss. Gneiting emphasizes that point forecasts must be evaluated in relation to their loss and target functional (Gneiting, 2011); the broader scoring-rule literature makes the evaluation target equally central (Gneiting and Raftery, 2007). A weighted arithmetic mean is not asserted to be optimal for absolute loss. Comparing the fixed rules under both losses shows what these particular combinations do, without choosing a favorable measure after observing their ranking.

Errors are first averaged over the excluded reviews within each paper, then averaged equally over the 416 papers. A paper with five reviews therefore receives the same final weight as a paper with three. This defines the estimand as the mean paper-level comparison in the retained release. It does not make reviews independent. Predictions within a paper share inputs, and the archive does not provide a verified cross-paper reviewer-person structure for this study. The paper is the aggregation unit, not a claimed independent draw from a conference population.

For each unequal-weight policy, paired paper-level differences from equal weighting are calculated. Deleting each paper in turn gives a deterministic sensitivity range for their overall mean. Although deletion procedures also appear in the jackknife literature (Efron and Stein, 1981), no jackknife variance estimator, standard error or confidence interval is reported here. Those would require a sampling interpretation that this finite historical comparison does not establish. The range instead answers whether removal of any single observed paper reverses the recorded aggregate direction.

An additional fixed calculation reallocates each paper's observed confidence multiset over its review positions. Every distinct assignment is enumerated, and its paper-average loss is calculated by excluding each score as before. The assignment reference uses the full within-paper confidence multiset, including values that can move between excluded and included positions; it is an arithmetic perturbation exercise, not a deployable prediction procedure. Assignments receive equal weight within paper. No assumption of random original assignment is made, so the resulting averages and extrema are not a randomization test or a significance calculation.

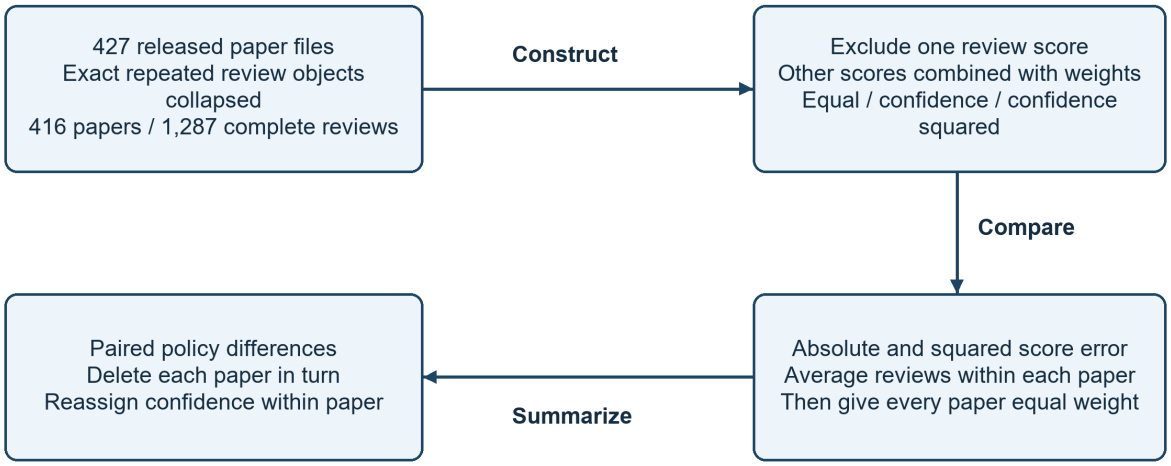


Figure 1. The analysis constructs several predictions from identical retained score objects. Its endpoint is disagreement with an excluded opinion. Arrows indicate calculation order, not a causal model of reviewer expertise.

The local protocol was saved after schema and support inspection and before error comparisons. This ordering reduces opportunities to select a favorable confidence transformation; it is not an externally registered confirmatory study. The distinction follows the concerns about advance specification discussed by Wagenmakers and colleagues (Wagenmakers et al., 2012). The two declared confidence policies remain a fixed, limited comparison after seeing the results.

4. A small adverse average with differences across papers

The principal comparison gives equal-weight absolute error of 0.927083 points and linear-confidence error of 0.935535. Their paired difference is positive, at 0.008451 points. Thus the extra weighting does not improve this recorded target on average. The magnitude is small relative to the one-point spacing of individual recommendations, but it is not rounded to zero. Squared confidence gives 0.943044 points, a larger adverse difference of 0.015960. Both comparisons use the same 416 papers and 1,287 excluded-review targets.

Table 1. Errors and paired differences for the full retained support. Absolute errors use score points; squared errors use squared score points. Each row averages reviews within paper and papers equally. Ranges delete one of the 416 whole papers and are deterministic sensitivity ranges, not confidence intervals. Positive differences indicate higher error than equal weighting.

Loss	Peer weights	Equal-paper error	Difference from equal	Delete-one-paper range
Absolute	Equal weights	0.927083	+0.000000	[+0.000000, +0.000000]
Absolute	Linear confidence	0.935535	+0.008451	[+0.007133, +0.009038]
Absolute	Squared confidence	0.943044	+0.015960	[+0.013859, +0.016933]
Squared	Equal weights	1.485276	+0.000000	[+0.000000, +0.000000]
Squared	Linear confidence	1.513798	+0.028522	[+0.021901, +0.030651]
Squared	Squared confidence	1.553447	+0.068171	[+0.054472, +0.071749]

All single-paper deletions preserve the positive direction of the linear-confidence absolute-error difference: the range is 0.007133 to 0.009038 points. A particularly unfavorable paper therefore does not by itself determine the sign of this mean. That stability is limited to the specified deletion exercise. It is neither evidence that a future conference would give the same result nor a test that a larger population effect differs from zero. The secondary squared-error comparison is also adverse, with a difference of 0.028522 squared points.

The paper-level comparisons are not uniformly unfavorable. Linear confidence lowers absolute error in 124 papers, raises it in 158 and leaves it exactly unchanged in 134. Ninety-five papers have no confidence variation, so reweighting cannot change their predictions. Some additional papers have unchanged average absolute loss even though their confidences vary. Counting improvements alone would discard both magnitudes and these ties. The primary average retains all three groups, including papers where the policy helps.

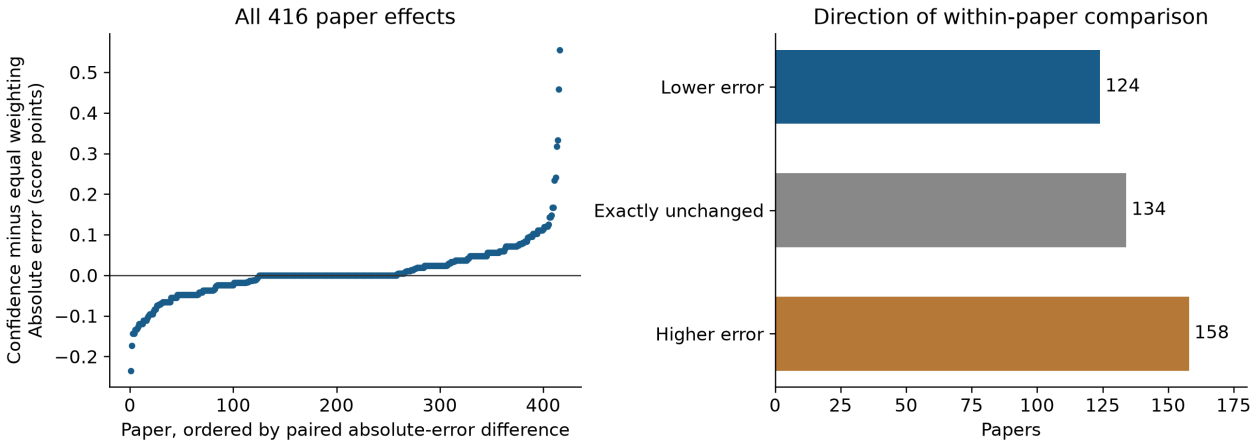


Figure 2. Individual paper differences range from approximately -0.235 to +0.556 recommendation points. The left panel includes every retained paper without trimming. The right panel counts exact rational signs, checked independently against the saved floating-point results. No population uncertainty is displayed.

Enumerating confidence assignments gives 1,414 distinct paper-assignment combinations per unequal-weight policy. Under linear weights, the mean assignment error is 0.933943, compared with the observed 0.935535. The recorded pairing of confidence with recommendation therefore performs 0.001592 points worse than this assignment-average reference. Its single-paper-deletion difference remains positive, from 0.000491 to 0.002162 points. This comparison does not establish that assignments were exchangeable in the actual review process; it describes how the observed arrangement compares with the specified finite rearrangements.

Table 2. Absolute-error comparisons with within-paper confidence reassignment. Joint assignment bounds independently choose each paper's lowest or highest attainable error and then average papers equally. They describe the finite product of assignments, rather than a confidence interval or a prediction of possible future reviewer behavior.

Peer weights	Observed error	Assignment-average error	Observed minus assignment mean	Joint assignment bounds
Linear confidence	0.935535	0.933943	+0.001592	[0.884991, 0.977478]
Squared confidence	0.943044	0.940887	+0.002156	[0.847596, 1.023219]

Assignment-average error itself exceeds equal-weight error. That fact should not be interpreted as an independent empirical discovery that random assurance is harmful. Symmetry makes the mean reallocated prediction equal to the unweighted peer mean; convexity of absolute or squared loss then prevents its

assignment-average loss from falling below the loss of that mean. The observed error's location relative to the assignment average contains information about the recorded pairing. The reference's nonnegative excess over equal weighting also has a mathematical component, which is kept separate in interpretation.

Review-count groups give a useful counterexample to a universal negative reading. The 381 three-review papers have a linear-confidence difference of +0.009554 points. The corresponding means are -0.000610 among 31 four-review papers and -0.026338 among four five-review papers. These small groups are reported as observed support rather than treated as independently established moderators. Their signs show why an adverse overall mean does not authorize a claim that confidence weighting worsens every paper or every review-panel size.

Table 3. Paper-level absolute-error comparison by the number of complete reviews. These groups partition the main support. The last group contains four papers and does not support a general conclusion about five-review panels.

Reviews per paper	Papers	Equal-weight absolute error	Linear-confidence difference	Squared-confidence difference
3	381	0.923885	+0.009554	+0.018211
4	31	0.924731	-0.000610	-0.004278
5	4	1.250000	-0.026338	-0.041592

5. What this comparison says about evaluation

The excluded score is an opinion produced in the same review process as its peers. Agreement with it can be useful to characterize a combination rule, yet it provides no independent adjudication of scientific validity. Latent-annotator approaches explicitly introduce models for unobserved labels and observer error (Dawid and Skene, 1979; Raykar et al., 2010). Such assumptions are not silently supplied here. Without an external criterion or a defended latent-label model, lower peer-score error cannot be renamed higher reviewer competence, better scientific selection or greater fairness.

Feedback aggregation can also target differences between groups rather than agreement with another individual response. Yuan and colleagues describe scale validation and hierarchical partial pooling for repeated anonymous-feedback comparisons (Yuan et al., 2026). Those operations are distinct from assigning fixed numeric weights to review confidence. This archive supplies paper blocks but no identified reviewer histories or validated confidence-to-reliability mapping. Its small adverse confidence-weight difference therefore does not establish that group pooling fails, or identify expertise or privacy-related disparities. It describes the declared arithmetic policies on the available recommendation target.

This boundary also matters when discussing bias. Lee and colleagues analyze the methodological and normative complexity of bias research, while Tomkins and colleagues examine a specific single- versus double-blind reviewing design (Lee et al., 2013; Tomkins et al., 2017). The present numeric comparison has no identified author attributes, reviewer identities or experimental masking assignment. It consequently cannot estimate any of those effects. Even a stable confidence-weight difference would not distinguish expertise, common interpretation, shared information, discussion effects or systematic preference as its explanation.

Historical records also limit temporal interpretation. Excluding a number from the arithmetic does not show that its source review was written before, or without seeing, every other review. This analysis does not reconstruct initial versus revised recommendations or pre-discussion information sets. It is a

comparison of the retained snapshot, not a prospective forecast experiment. Similarly, pooling the released directory splits for a record-level census is different from learning a confidence policy on training papers and evaluating it on previously unseen papers.

The failure of these two fixed mappings is therefore a reason to validate a proposed weighting rule, not a reason to discard confidence information generally. A different elicitation, a defensible mapping to reliability, or an explicitly trained and independently evaluated model might behave differently. Choosing among many such mappings on this archive would introduce model-selection concerns of the kind discussed by Cawley and Talbot (Cawley and Talbot, 2010). The present analysis avoids that search; it supplies no claim about the best achievable confidence-aware predictor.

Source documentation is part of this interpretation. Exact-object repetition, missing confidence and disagreement between current scored-object counts and the historical summary change the reader's understanding of the evidence. Dataset documentation frameworks emphasize making collection and intended uses visible (Gebru et al., 2021). Here the numeric projection, immutable file identities, inclusion reasons and separate paper-level calculations preserve those decisions. They do not guarantee that every historical observation is an independent judgment or that the archive covers every event in the review process.

6. Conclusion and reproducibility

On the retained ICLR2017 records, giving recommendations linear confidence weights slightly increases excluded-peer score error, and squaring the same confidence values increases it further. The overall direction survives removal of any one paper, while a substantial subset of individual papers improves. This is a bounded conclusion about fixed arithmetic policies and a finite set of opinions. Treating confidence as a useful contribution weight requires evidence for the intended target; its numeric presence alone supplies neither a correctness standard nor a reviewer-quality label.

All calculations were executed locally from the pinned public release. The saved analysis contains every policy prediction, per-paper loss, complete confidence-assignment enumeration and deterministic deletion summary. An independent implementation using exact rational arithmetic reconstructs the predictions and primary summaries directly from the original JSON objects. This separation of executable evidence from prose follows principles of computational reproducibility (Peng, 2011; Sandve et al., 2013). Raw review text and manuscript content are not included in the derived analysis table, and no reviewer identities are inferred.

Procedural preparation used the local Scientific Agent Skills documentation (Kassis et al., 2026); that resource is not an empirical validation of the present findings. Public research access does not establish an unrestricted redistribution license for every source component. The deliverable therefore retains source pointers and numerical derivatives while making no blanket licensing or anonymization claim. Human source-claim review, authorship decisions and submission approval remain pending. No new reviews, policy experiment, trained language model or paid external analysis service was used.

References

1. 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.

2. 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

3. Bates, J. M.; Granger, C. W. J.. (1969). The Combination of Forecasts. Journal of the Operational Research Society. https://doi.org/10.1057/jors.1969.103

4. Dawes, Robyn M.. (1979). The robust beauty of improper linear models in decision making.. American Psychologist. https://doi.org/10.1037/0003-066x.34.7.571

5. 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

6. Fischhoff, Baruch; Slovic, Paul; Lichtenstein, Sarah. (1977). Knowing with certainty: The appropriateness of extreme confidence.. Journal of Experimental Psychology: Human Perception and Performance. https://doi.org/10.1037/0096-1523.3.4.552

7. 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

8. Clemen, Robert T.. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting. https://doi.org/10.1016/0169-2070(89)90012-5

9. Raykar, Vikas C.; Yu, Shipeng; Zhao, Linda H.; Valadez, Gerardo Hermosillo; Florin, Charles; Bogoni, Luca; Moy, Linda. (2010). Learning From Crowds. Journal of Machine Learning Research. https://www.jmlr.org/papers/v11/raykar10a.html

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

11. 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

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

13. 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

14. Dawid, A. P.; Skene, A. M.. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Applied Statistics. https://doi.org/10.2307/2346806

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

16. Sandve, Geir Kjetil; Nekrutenko, Anton; Taylor, James; Hovig, Eivind. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1003285

17. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

18. Kang, Dongyeop; Ammar, Waleed; Dalvi, Bhavana; 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

19. Griffin, Dale; Tversky, Amos. (1992). The weighing of evidence and the determinants of confidence. Cognitive Psychology. https://doi.org/10.1016/0010-0285(92)90013-r

20. 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

21. Moore, Don A.; Healy, Paul J.. (2008). The trouble with overconfidence.. Psychological Review. https://doi.org/10.1037/0033-295x.115.2.502

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. 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

24. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847

25. Soll, Jack B.; Larrick, Richard P.. (2009). Strategies for revising judgment: How (and how well) people use others' opinions.. Journal of Experimental Psychology: Learning, Memory, and Cognition. https://doi.org/10.1037/a0015145

26. Koriat, Asher. (2012). The self-consistency model of subjective confidence.. Psychological Review. https://doi.org/10.1037/a0025648

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

28. Larrick, Richard P.; Soll, Jack B.. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. Management Science. https://doi.org/10.1287/mnsc.1050.0459

29. Efron, B.; Stein, C.. (1981). The Jackknife Estimate of Variance. The Annals of Statistics. https://doi.org/10.1214/aos/1176345462

30. 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.