**Journal of Integrated Knowledge**
Vol. 2 No. 7 (2026)

# Self-Inclusion in a Rating Reference: Small Error Contraction and Pairwise Rank Changes in Jester

Logan Franklin[1] · Jacob Knight[2] · Gavin Bell[3]

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

Published: July 08, 2026 | Article 861

## Abstract

Every released rater row appeared closer to its item reference when its own score helped define that reference. Across 73,421 rows and 4,136,360 observed Jester ratings, removing that contribution increased mean absolute deviation by 0.00009375 rating points. The change reversed 10,970 of 2,695,284,910 strictly ordered row pairs, or 0.000407%. All 100 item-deletion summaries retained a positive mean difference. A simple identity explains the direction; the complete matrix establishes its magnitude and ordering consequences. The result concerns conformity to an internally constructed reference. Neither agreement nor the small arithmetic advantage establishes evaluator competence or an externally correct judgment.

# Self-Inclusion in a Rating Reference: Small Error Contraction and Pairwise Rank Changes in Jester


## Abstract

Every released rater row appeared closer to its item reference when its own score helped define that reference. Across 73,421 rows and 4,136,360 observed Jester ratings, removing that contribution increased mean absolute deviation by 0.00009375 rating points. The change reversed 10,970 of 2,695,284,910 strictly ordered row pairs, or 0.000407%. All 100 item-deletion summaries retained a positive mean difference. A simple identity explains the direction; the complete matrix establishes its magnitude and ordering consequences. The result concerns conformity to an internally constructed reference. Neither agreement nor the small arithmetic advantage establishes evaluator competence or an externally correct judgment.


## Who contributes to the reference used to assess them?

A common way to summarize a person's numerical judgments is to compare them with a group reference. Such a calculation appears straightforward when many people have rated the same objects. Yet the reference can include the very judgment being assessed. That contribution pulls the group value toward the observed score and mechanically reduces its apparent discrepancy. The issue is not whether the mean was calculated correctly. It is whether the comparison answers the intended question about agreement with others.

Our research question is how much that inclusion changes row-level conformity errors and their ordering in a complete released rating matrix. We compare two item references constructed from identical observed values: the mean including the evaluated row and the mean after subtracting that row's value. No observed rating is hidden, predicted, or imputed. The first analytical proposition is the exact relation between the two absolute deviations. The second asks whether the support-dependent correction preserves the empirical ordering of rows. Its answer requires the actual rating pattern as well as the identity.

The distinction has a concrete evaluation context. Zhao et al. (2022) compare judges' scores with an adjusted estimate of work level and discuss the complications of unequal reviewing workloads. Their professional, authenticity, and controversy framework concerns contest evaluators. Here, an explicit reference-construction question is examined on a different numerical matrix. We call the result conformity because the released preferences provide no external assessment of professional quality. Keeping the reference rule explicit is useful before attaching a stronger interpretation to any discrepancy index.

Evaluation studies in collaborative filtering also distinguish the user task, data design, and metric being measured (Herlocker et al., 2004). Shani and Gunawardana (2011) organize recommendation evaluation around such choices. These considerations apply even to a calculation without a learned recommender: an internally defined deviation and prediction of an unseen rating are different tasks. The Jester source paper uses its own training and testing design for recommendation algorithms (Goldberg et al., 2001). Our finite-matrix calculation neither repeats that evaluation nor borrows its reported performance as a new result.

Numeric error and ordering deserve separate attention. Steck (2013) distinguishes rating prediction from ranking in recommendation evaluation. Our ranking is specifically the ordering of released rows by mean conformity error, rather than an item recommendation list. A correction can change each numeric error

without changing every ordering, or alter a close pair despite a small average difference. Reporting both quantities avoids treating a favorable error number as a complete description of the comparison.

## The complete older three-file release

We acquired all three numerical archives linked from the owner's older Jester dataset page (Goldberg, n.d.). The first two contain 24,983 and 23,500 rows with at least 36 recorded ratings. The third contains 24,938 rows with between 15 and 35. Together they provide 73,421 released rows and 100 item columns. The source describes anonymous ratings collected in the older release and permits research use with acknowledgment of Eigentaste. We preserve that statement rather than infer a blanket Creative Commons license. No joke text, user profile, or identity attribute was acquired.

Each ZIP passed its internal CRC check. The manifest records the downloaded object hashes and the hash of each extracted XLS. These are local version identifiers, not provider-supplied cryptographic checksums. The first workbook column gives the row's recorded rating count; the next 100 columns contain values from -10 to 10 or the missing-value code 99. Reconstructing the count from every row exactly matches the first column in all three files. There are no invalid or nonfinite selected values. The full numerical scope is retained instead of choosing particularly dense rows or favorable items.

The analysis contains 4,136,360 observed ratings and 3,205,740 cells coded 99. The latter are absent values, not neutral scores. Item support ranges from 18,505 to 73,413 observed rows. Every item therefore permits a leave-row-out mean, and every row has at least 15 ratings. The matrix's row unit follows the provider's format; we do not independently authenticate distinct people across the files or merge equal-looking preference patterns. Source file and row number are retained only as technical correspondence keys.

These choices matter because ratings are selectively observed. Marlin et al. (2007) examine the missing-at-random assumption in collaborative filtering, while Steck (2010) addresses training and evaluation with nonrandom missingness. Their findings do not identify the missingness process of each cell here. They do make clear why an unrated item cannot be silently supplied with a population-average opinion. We compare reference rules only on the observed matrix. Exposure probabilities are unavailable, so this is not a debiased population-preference estimate of the kind of evaluation problem considered by Schnabel et al. (2016).

Documentation should retain both the source meaning and the resulting operational unit (Gebru et al., 2021). Jester uses a numerical response scale, but numerical continuity does not by itself establish equal psychological distances across respondents. The distinction among measurement scales remains relevant (Stevens, 1946), as does the relationship between a statistical operation and the quantity it is intended to represent (Hand, 1996). Our absolute deviations are explicitly in the released rating units. They are not calibrated distances between people's underlying opinions.

## A fixed comparison with a known algebraic direction

For an observed rating r on item j, let n be that item's observed count and let m be its mean over all observed rows. The included-reference deviation is the absolute value of r minus m. Removing the evaluated row gives a reference equal to the original item sum minus r, divided by n minus one. Elementary rearrangement yields the leave-row-out deviation as n divided by n minus one times the included deviation. This relation holds cell by cell whenever n exceeds one.

The direction is therefore determined before observing an effect estimate: subtracting a row's contribution cannot reduce its absolute deviation under these two mean-reference rules. The empirical questions

concern size and ordering. If every item had the same support, the same positive multiplier would apply to every row's mean error and preserve its order. With unequal item support and different observed item sets, the row-level correction need not be a common multiplier. Close row pairs can then change order even when every individual correction is small.

For each released row, we average absolute deviations over its observed items under each reference. The primary effect is the row-equal mean of leave-row-out error minus included-reference error. This weighting gives one contribution to each released row; the 4.1 million ratings do not become that many independent evaluators. Dividing an error by the 20-point scale width gives an explicitly labeled full-scale percentage. It does not change the reference or identify the meaning of a preferred rating. Both rules remain comparisons with an internal consensus.

All source rows, observed cells, reference definitions, and summaries were fixed before the effects and rankings were computed. The protocol was saved after source and support inspection, so it is a planned secondary analysis rather than externally preregistered or outcome-blind research. Advance specification helps distinguish intended calculations from outcome-driven changes (Wagenmakers et al., 2012). Recording the complete comparison also follows the experimental-reporting concern of Dodge et al. (2019). We do not search an item subset, activity threshold, or selected evaluator cutoff to obtain a larger difference.

This is a complete-release calculation, with no fitted predictor to tune. Nevertheless, the more general warning about selection on an evaluation criterion is relevant to what could follow from it (Cawley and Talbot, 2010). Choosing a particularly discordant subset after inspecting all corrections would answer a different question. We retain the full matrix and provide the entire item-deletion family, including the smallest effect, instead of presenting an illustrative extreme as the principal result.

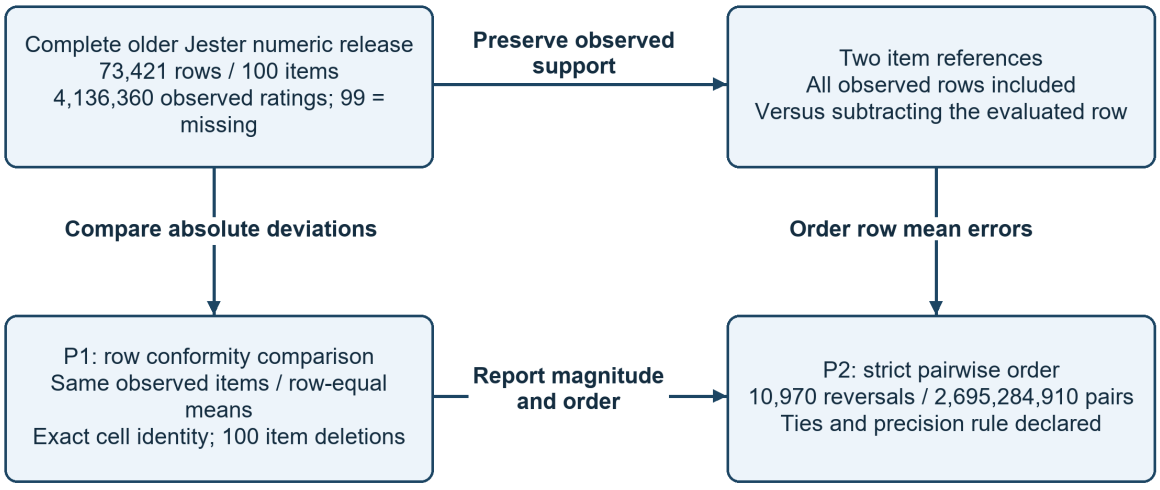


*Figure 1. P1 reconciles the two reference rules through an exact cell-level identity and measures their row-level effect. P2 compares the resulting strict pairwise orders. Arrows describe calculation, not a causal mechanism, evaluator expertise, or prediction of unobserved ratings.*

## Ordering and finite-composition uncertainty

We count a strict reversal when two rows have opposite strict orders under their two mean errors. Ties are reported separately rather than broken by an inferred user characteristic. The comparator uses the stored double-precision row summaries with exact equality as its declared tie rule. Kendall (1938) provides the rank-concordance background; the treatment of tied rankings requires an explicit convention (Kendall, 1945). Neither the pair count nor a rank coefficient will be used as an independent-sample significance test.

The number of possible unordered pairs is large, but enumeration of every pair is unnecessary. The primary implementation sorts by one error and accumulates reverse-order counts in the other using a binary indexed tree, following the cumulative-frequency data structure described by Fenwick (1994). A separate mergesort-based implementation checks the inversion count. Comparable pairs exclude ties in either score, with jointly tied pairs restored once to avoid double subtraction. In this acquired matrix, both error summaries happen to have no ties under the declared precision rule.

For a composition diagnostic, we delete each of the 100 items in turn and recompute every remaining row mean. All rows remain eligible because at least 14 ratings survive. The remaining items' references are unchanged: each is still defined by its own observed column. We retain all 100 mean paired effects and report their minimum and maximum. This deterministic range concerns the finite item collection; it is not a confidence interval for later users, different jokes, or repeated sampling from an underlying population.

The distinction is consequential with billions of pair comparisons. Those pairs repeatedly reuse the same rows, item means, and observed values. They cannot manufacture billions of independent judgments. Similarly, deleting an item is a controlled change in the computational support, rather than a new trial or a random sample of jokes. The resulting range describes a specific sensitivity. It does not absorb uncertainty about selective rating, individual identity, source history, or the reference's substantive validity.

## Small contraction, with a measurable but sparse change in pair orders

Table 1 shows the actual magnitude of the reference effect. The row-equal mean absolute deviation is 4.19647044 points with self-inclusion and 4.19656419 after removing the row. Their difference is 0.00009374736 points, or 0.000468737% of the full 20-point scale. Every released row has a positive correction. The cell identity predicts its direction, while the complete-data calculation establishes that it is extremely small on this scale.

***Table 1. Full-release conformity errors and finite reference-construction effects.***

| Quantity | Value | Unit or scope |
|---|---|---|
| Released rows | 73,421 | Full three-file release |
| Observed ratings | 4,136,360 | Cells across 100 items |
| Unrated code-99 cells | 3,205,740 | Excluded from means, retained in audit |
| Observed ratings per item | 18,505-73,413 | Actual reference support |
| Mean included-reference error | 4.19647044 | Rating points; row-equal mean |
| Mean leave-row-out error | 4.19656419 | Same rows and observed items |
| Mean paired increase | 0.00009374736 | Rating points |
| Median paired increase | 0.00008794533 | Rating points |
| Row-level increase range | 0.00001198371 to 0.00030520380 | Finite observed rows |
| Mean increase / 20-point scale | 0.000468737% | Full-scale percentage |
| Delete-one-item mean-effect range | 0.00009314957 to 0.00009468031 | All 100 omissions; not a CI |

*Table 1 note: Deviations use original numerical rating points. The primary estimator weights released rows equally and uses each row's observed items. The item-deletion range is deterministic finite-composition variation, not a sampling confidence interval. No external correct-score label or latent-quality model is fitted.*

The row corrections range from 0.00001198371 to 0.00030520380 points, with median 0.00008794533. Different observed item sets and column support produce different multipliers. However, even the least-supported item has 18,505 ratings, which makes the algebraic correction close to one. The large source collection limits the possible self-inclusion effect under this particular mean-reference construction. A smaller panel could behave differently; it is not represented by these observed values.

***Table 2. Strict pairwise order comparison under the two references.***

| Quantity | Value | Interpretation |
|---|---|---|
| All unordered row pairs | 2,695,284,910 | Repeatedly share rows/items |
| Ties in included / leave-out / both | 0 / 0 / 0 | Exact stored-double summaries |
| Strictly comparable pairs | 2,695,284,910 | Denominator after tie accounting |
| Strict reversals | 10,970 | Opposite strict orders |
| Concordant pairs | 2,695,273,940 | Same strict order |
| Reversed share | 0.000407007% | Pairs, not a fraction of people |
| Kendall tau-b | 0.99999186 | Descriptive; no p-value |

*Table 2 note: These are dependent comparisons of the same released row summaries. Strict reversals use exact stored-double ordering and have no significance-test interpretation. A row's conformity order is not a professional-quality order or an item recommendation list.*

There are 10,970 reversed strict pairs among 2,695,284,910 comparable pairs, a fraction of 0.000407007%. The corresponding Kendall coefficient is 0.99999186. Thus numeric contraction occurs in every row, while the overwhelming majority of pair orders are preserved. The reversal fraction

describes pairs, not the fraction of people affected. No most-conforming group is selected, and a small pair fraction cannot establish that every consequential selection rule would produce the same set.

Removing each item leaves the average correction between 0.00009314957 and 0.00009468031 points. All 100 values remain positive, as expected from the same identity on the remaining support. This check adds a finite measure of magnitude sensitivity; it is not additional empirical evidence for the algebraic sign. Figure 2 shows the full row correction distribution and every item omission so that the range can be judged alongside the principal effect.

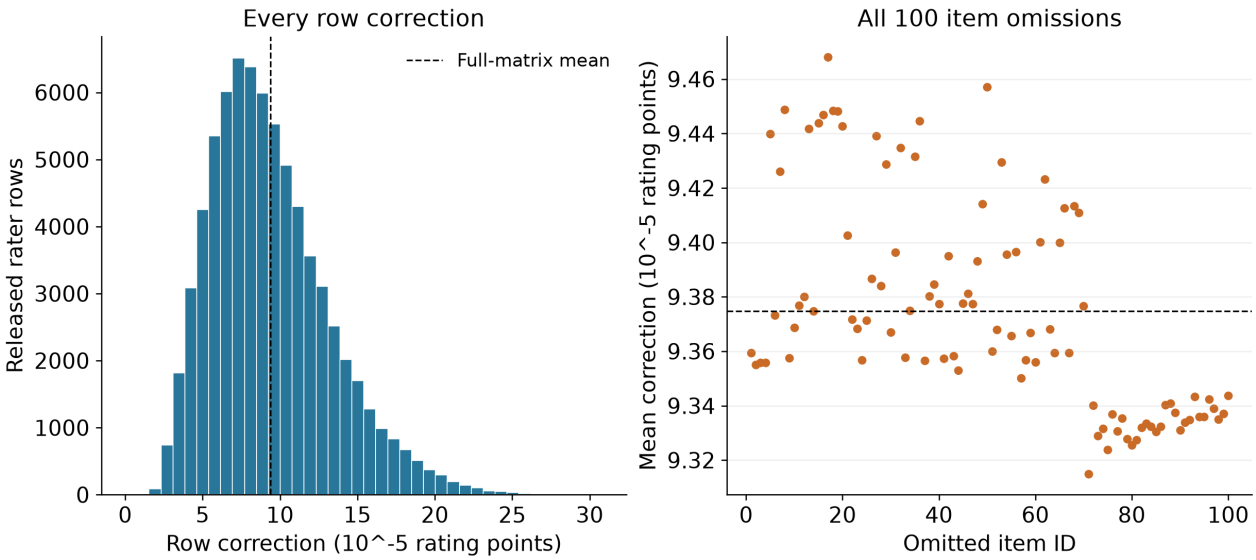


*Figure 2. Left: all 73,421 row corrections, with no outcome-based deletion. Right: mean effects after each item omission, with the full-matrix mean as a reference. The point-panel axis is expressed in rating units times 100,000 and does not start at zero; it displays a small magnitude range rather than error bars or independent item replications.*

## Conformity remains a limited construct

The result separates two issues that can otherwise be bundled into an appealing evaluator score. Self-inclusion creates a known arithmetic advantage under an item mean. Its observed magnitude here is tiny. Neither fact establishes that closeness to the resulting mean is a measure of expertise. A person may differ from consensus because of taste, a different interpretation, or a scale-use pattern. The released matrix contains no adjudicated label that distinguishes those possibilities.

Models of observer error and learning from multiple annotators make assumptions about how judgments relate to an underlying target (Dawid and Skene, 1979; Raykar et al., 2010). We estimate no such latent target. Aroyo and Welty (2015) discuss the role of disagreement in annotation, and Plank (2022) examines human label variation. Their perspective helps locate the present limit: variation is an observation requiring interpretation, not evidence that any particular rater is defective. A conformity statistic alone supplies no basis for identifying a problematic person.

The intended interpretation matters as much as the calculation. Messick (1995) places validity in the inference made from an assessment, while Jacobs and Wallach (2021) examine the consequences of measurement choices. In this study, reference inclusion is made explicit and independently checked. That strengthens computational transparency, not the validity of a professional-quality label. The distinction remains even when the internally defined score is reproducible to many decimal places.

Yuan et al. (2026) describe scale validation and uncertainty-aware group evidence in an anonymous-feedback framework supporting management decisions. Only the primary abstract and

metadata of that work were available here; its private multi-year survey is not analyzed. Our complete numeric matrix permits a narrower reference-construction check. It provides no validation of an experience scale, privacy guarantee, or management decision. What it contributes is an auditable demonstration of how one declared internal convention changes a score and its pairwise order.

## Reproducibility and conclusion

Primary calculation uses NumPy arrays and pandas workbook/table handling (Harris et al., 2020; McKinney, 2010); Matplotlib produces the quantitative display (Hunter, 2007). The numerical record preserves the three original archives and workbooks, schema checks, protocol, row results, all item omissions, and source hashes. A separately written checker reads the raw XLS cells directly, uses compensated sums, reconstructs every row error, and checks the ordering with mergesort. It agrees on all 10,970 reversals and all 100 deletions; maximum row-error differences are below 0.000000000001 points.

This is computational reproducibility of one finite comparison, rather than independent evidence about people or preferences (Peng, 2011). Scientific Agent Skills supported evidence organization and consistency review (Kassis et al., 2026); it is a procedural software citation rather than support for the empirical conclusion. Accountable-author source verification, declarations, and submission approval remain pending. Original rights notices are retained, and the local source bundle is not presented as a new unrestricted public dataset.

Self-inclusion contracts every observed row's deviation under these two item-mean rules. In the complete older Jester release, the contraction is minute and strict pair reversals occupy only a small fraction of comparisons. The practical lesson of this calculation is to state who contributes to an assessment reference and what the resulting score means. The demonstrated quantity is agreement with that reference; a stronger judgment about evaluator competence would require additional evidence.

## References


1. Plank, Barbara. (2022). The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2022.emnlp-main.731
2. 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.
3. Goldberg, Ken. (n.d.). Jester Collaborative Filtering Dataset: anonymous ratings data, older three-file release. University of California, Berkeley. Retrieved 17 November 2026. https://goldberg.berkeley.edu/jester-data/
4. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
5. 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
6. 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.
7. Fenwick, Peter M.. (1994). A new data structure for cumulative frequency tables. Software: Practice and Experience. https://doi.org/10.1002/spe.4380240306
8. 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.
9. 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

10. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901
11. KENDALL, M. G.. (1938). A NEW MEASURE OF RANK CORRELATION. Biometrika. https://doi.org/10.1093/biomet/30.1-2.81
12. Marlin, Benjamin; Zemel, Richard S.; Roweis, Sam; Slaney, Malcolm. (2007). Collaborative Filtering and the Missing at Random Assumption. Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence. https://arxiv.org/abs/1206.5267
13. Schnabel, Tobias; Swaminathan, Adith; Singh, Ashudeep; Chandak, Navin; Joachims, Thorsten. (2016). Recommendations as Treatments: Debiasing Learning and Evaluation. International Conference on Machine Learning. https://proceedings.mlr.press/v48/schnabel16.html
14. KENDALL, M. G.. (1945). THE TREATMENT OF TIES IN RANKING PROBLEMS. Biometrika. https://doi.org/10.1093/biomet/33.3.239
15. Aroyo, Lora; Welty, Chris. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AI Magazine. https://doi.org/10.1609/aimag.v36i1.2564
16. Goldberg, Ken; Roeder, Theresa; Gupta, Dhruv; Perkins, Chris. (2001). Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval. https://doi.org/10.1023/a:1011419012209
17. Steck, Harald. (2013). Evaluation of recommendations: rating-prediction and ranking. Proceedings of the 7th ACM conference on Recommender systems. https://doi.org/10.1145/2507157.2507160
18. Shani, Guy; Gunawardana, Asela. (2011). Evaluating Recommendation Systems. Recommender Systems Handbook. https://doi.org/10.1007/978-0-387-85820-3_8
19. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677
20. McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-00a
21. 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
22. 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
23. 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
24. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723.
25. Dodge, Jesse; Gururangan, Suchin; Card, Dallas; Schwartz, Roy; Smith, Noah A.. (2019). Show Your Work: Improved Reporting of Experimental Results. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1224
26. 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
27. Herlocker, Jonathan L.; Konstan, Joseph A.; Terveen, Loren G.; Riedl, John T.. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems. https://doi.org/10.1145/963770.963772
28. Hunter, John D.. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55
29. Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.. American Psychologist. https://doi.org/10.1037/0003-066x.50.9.741
30. Steck, Harald. (2010). Training and testing of recommender systems on data missing not at random. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. https://doi.org/10.1145/1835804.1835895