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

Can an Observer's Rating Offset Transfer Across Crowdsourcing Rows? A Held-Row Study of Public Video-Quality Opinions

Jeremy Mitchell[1] · Simon Wells[2] · Ryan King[3]

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

Published: July 08, 2026 | Article 864

Abstract

Lower average rating error did not mean that calibration helped every observer. Using 233,168 public visual-quality ratings, we evaluated additive observer offsets across five partitions of complete source presentation rows. Repeated ratings formed 233,048 observer–video means from 1,257 observer codes and 1,596 stimuli. Against a reference excluding the evaluated observer, equal-observer squared error fell from 0.787228 to 0.552414, but 211 observers incurred greater error and 438 predictions fell outside the 1–5 scale. All 499 crossed resamples were defined, although every resample invoked the planned no-calibration fallback and the original absolute-error change lay outside its percentile interval. The findings support limited transfer of individual scale offsets conditional on peer ratings. They establish neither evaluator skill nor objective video quality.

Can an Observer's Rating Offset Transfer Across Crowdsourcing Rows? A Held-Row Study of Public Video-Quality Opinions

Abstract

Lower average rating error did not mean that calibration helped every observer. Using 233,168 public visual-quality ratings, we evaluated additive observer offsets across five partitions of complete source presentation rows. Repeated ratings formed 233,048 observer-video means from 1,257 observer codes and 1,596 stimuli. Against a reference excluding the evaluated observer, equal-observer squared error fell from 0.787228 to 0.552414, but 211 observers incurred greater error and 438 predictions fell outside the 1-5 scale. All 499 crossed resamples were defined, although every resample invoked the planned no-calibration fallback and the original absolute-error change lay outside its percentile interval. The findings support limited transfer of individual scale offsets conditional on peer ratings. They establish neither evaluator skill nor objective video quality.

The question concerns an observer's scale, not a video's true quality

Two people can use the same five response categories differently. One may repeatedly choose lower categories than other viewers of the same material, while another may use higher categories. A mean opinion score hides those individual departures. Whether a departure can be estimated from one collection of stimuli and carried into another is an empirical question. The KonVid-150k study provides individual ratings of ordinary videos embedded in source presentation rows, making that question inspectable without downloading or processing the videos themselves (Götz-Hahn et al., 2021a).

Subject-scoring research distinguishes several possible sources of variation. Janowski and Pinson (2015) discuss subject bias, scoring inaccuracy and stimulus difficulty. Li et al. (2020) separately model systematic bias and inconsistency. Work on discrete quality scales further examines how individual scoring behavior interacts with the available response categories (Tiotsop et al., 2024). These approaches motivate an individual-level analysis, but an estimated additive offset is only one possible summary. We do not infer that every departure from a crowd mean is erroneous, or implement the cited latent-quality models.

Subjective assessment also separates different meanings of reliability and validity. The historical July 2022 edition of Recommendation P.910 distinguishes within-person repeatability, between-person agreement and agreement with the intended true value (ITU-T, 2022). These are different criteria even when they use the same numerical responses. Personalized streaming-quality modeling addresses another distinction, between a model for an aggregate viewer and a model for an individual viewer (Peroni et al., 2023). Our task is narrower: predict a known observer's held-row ratings when other observers' ratings of those test videos are already available.

The research question is therefore whether a fixed additive observer calibration transfers across complete source rows and lowers the observer-weighted loss, and how often it instead worsens an individual's loss. Proposition P1 compares the calibrated and uncalibrated held-row predictions. Proposition P2 preserves individual heterogeneity and every fold rather than treating an average improvement as universal. No direction was required by the frozen design. The comparison does not test a pixel-based quality model, identify a playback anomaly or evaluate a new observer without calibration data.

This distinction matters for measurement. Hand (2004) emphasizes the connection between an operation on numbers and the meaning assigned to its result. Treating ordered categories as equally spaced adds an operational assumption of the kind discussed by Stevens (1946). Here squared and absolute differences concern the released score codes; they are not distances on an independently calibrated perceptual scale. Validity also concerns the inference made from a response rather than a numerical property alone (Messick, 1995). A lower prediction error does not by itself support an inference about evaluator competence.

Research on subjective annotations gives an additional reason to retain individual outcomes. Aggregating judgments can conceal systematic differences among annotators (Mostafazadeh Davani et al., 2022), while human label variation need not always represent annotation error (Plank, 2022). Those studies concern other annotation tasks and do not identify the causes of disagreement here. They support a concrete reporting choice: every worsening observer remains in the evaluation, and neither calibration success nor disagreement is translated into a personal quality label.

The acquired raw annotation record

We acquired the official dataset page's linked annotation mirror on 17 November 2026 local time, corresponding to 16 November 2026 UTC (Götz-Hahn et al., 2021b). The metadata archive contains 24,900,011 bytes and has SHA256 69e83ae8627f517fff6fe9924c08aa906efd7b0039df597fcc976e975a739977. Two successful downloads matched that digest. Its original B-vote member contains 233,168 rows, and the archive's internal CRC32 was verified during extraction. Those checks establish reproducible acquired bytes, rather than a signed authenticity certificate or independent validation of the ratings.

We retained only the source-row locator, stimulus filename, observer code, presentation-row code and numeric score from the full B-vote table. Country, submission times, channel and trust fields were discarded during extraction, and the complete original archive was not persisted. The A tables and videos do not enter this study. The provider describes research availability but supplies no explicit open annotation license in the acquired documentation. We preserve that scope, attribute the source and do not infer that the paper's publication license permits unrestricted redistribution of the underlying ratings.

Every raw score is a finite integer from one to five. The file contains 1,257 observer codes and 1,596 stimulus keys. It has 233,048 distinct observer-video pairs, including 120 pairs with two raw rows each. We average both constituent ratings within each repeated pair, retaining their original locators and multiplicities. This defines a pair-level response without asserting that the repeated entries are duplicate exports or independently repeated experimental judgments. The resulting response is the observed pair mean, which can include half-category values.

The public aggregate-score table lists 1,576 stimuli, all found in the raw votes. Twenty additional raw stimuli, accounting for 2,165 rows, lack an aggregate-table entry. Public descriptions also give different headline counts across versions. We retain all 1,596 raw stimuli because numeric rating analysis does not require successful video decoding or a supplied aggregate score. Neither aggregate-score column is used as a reference or calibration input. The study consequently concerns the complete acquired raw B annotation record, rather than claiming to reproduce the final aggregate-MOS benchmark.

The source protocol explains why the row identifier matters. Each relevant source row mixes three quality-control videos with twelve ordinary dataset stimuli; the acquired B table contains the ordinary stimuli. All 1,596 stimuli occupy 133 source-row groups of twelve, with each stimulus appearing in exactly one group. We construct connected components of the row-stimulus graph so that a shared

stimulus could not cross a partition through another row. In this realized record the components coincide with those 133 rows. Raw observations per stimulus range from 89 to 175, while raw observations per observer range from 48 to 1,596.

Source controls and their limits remain separate from the local computation. The acquired author preprint describes visual-defect instructions, post-playback category rating and hidden quality-control questions. We read that actual preprint and verified the final publication's bibliographic identity; an inaccessible publisher PDF was not represented as read. The record does not independently verify observer identities, the screening history of every raw row or the chronology needed for prospective prediction. A source row is also smaller than a multirow page, so the chosen holdout does not remove every possible shared page or viewing-context influence.

Constructing peer references and training offsets

Let y_iv be the mean of observer i's released ratings on video v. For each pair, the peer reference q_iv is the mean of the other distinct observers' pair means for that video. The evaluated observer has no contribution to q_iv, including through a repeated raw rating. This deliberately avoids using a self-inclusive supplied MOS as the test reference. Leakage analysis requires stating what information is actually available to a computation (Kaufman et al., 2012). Here the other observers' test-video labels are available by design, while the evaluated observer's own test response is unavailable to their offset fit.

We fixed five partitions using IDs alone before calculating rating means or losses. Entire row-stimulus components were sorted, shuffled with seed 86420261117 and assigned round-robin to folds. The resulting test folds contain 27, 27, 27, 26 and 26 components. Every pair appears once as a test response. Repeated occurrences of an observer-video pair and all ordinary stimuli from its source row stay together. For every observer-fold combination with test observations, at least twelve distinct calibration pairs remain in other folds; none lacks main-data calibration support.

For test fold k, the observer's offset b_i,-k is the mean of y_is-q_is over that observer's pairs outside fold k. Baseline prediction is q_iv; calibrated prediction is q_iv+b_i,-k. This is a fixed one-parameter additive adjustment for each observer and fold. There is no fitted slope, shrinkage, nonlinear mapping or outcome-selected feature. Predictions outside the response scale are retained without clipping, and the full pair support is identical in both models. Figure 1 links the reference, training information and held-row comparison.

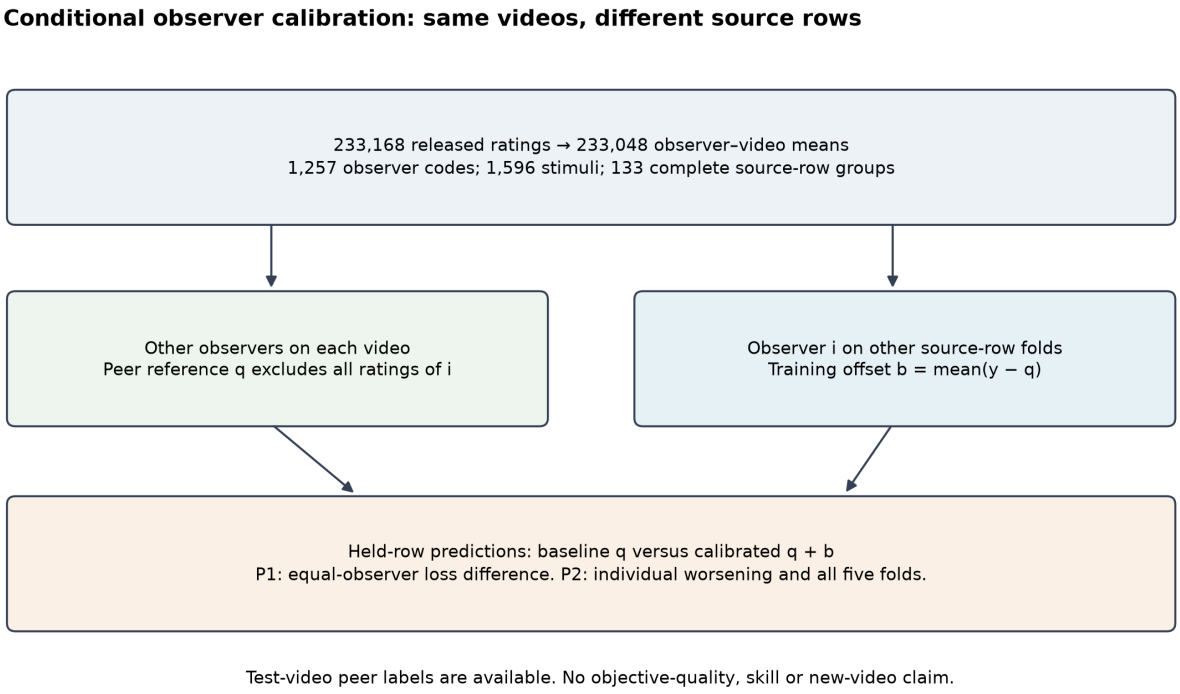

Figure 1. Arrows indicate computation, not a causal mechanism. P1 evaluates the same pair means under two predictions. P2 preserves observer heterogeneity and every fold. Other observers' test-video ratings remain available, so this is conditional calibration rather than prediction for an unrated video.

A directly relevant evaluation boundary comes from the contest-judge framework of Zhao et al. (2022). Its actual methods distinguish deviation from an estimated work level, score differentiation and reconsideration-related controversy. Our data contain neither established true scores nor contest reconsideration rounds. Repeated scale offsets are not those professionalism, authenticity or controversy indices. The source motivates distinguishing evaluator-level criteria; it does not license interpreting crowd conformity as skill or applying its complete index system to these ratings.

The model comparison is predictive only in its explicitly conditional sense. Yuan et al. (2026), in the registered abstract of their working paper, connect scale validation and hierarchical Bayesian partial pooling with uncertainty-aware group differences in anonymous community-experience feedback. The group-comparison objective described there differs from our prediction of a known observer's held-row rating. Here a recurring observer code supplies that observer's calibration ratings from other source rows, while other observers' ratings supply the test-video reference. An anonymous-survey description alone does not determine whether person-level histories can be linked, so we do not infer the linkage architecture of their non-public data. We neither implement their Bayesian framework nor claim a privacy guarantee for our observer codes. The reported gains cannot be transported to an unknown observer or an unrated video merely by calling the five partitions cross-validation.

Loss and crossed resampling

We first calculate each observer's mean squared held-out residual across all their pairs, pooling the folds by pair count within that observer. The primary model loss is then the equal-weight average of the 1,257 observer losses. Its effect is calibrated minus baseline MSE, in squared score-code units. This weight answers an average-observer question rather than allowing observers with more ratings to dominate. Weighting follows the intended estimand, as emphasized by Solon et al. (2015). We do not obtain the primary result by simply averaging the five fold losses.

Absolute error is a prespecified secondary loss, calculated with the same observer weights. Point-forecast evaluation depends on the functional and loss being assessed (Gneiting, 2011), and forecast-accuracy measures can favor different aspects of errors (Hyndman and Koehler, 2006). The two losses therefore remain separately reported. Each fold's secondary summary first averages pairs within its present observers and then averages those observers equally. Fold results are not independent replications: their calibration sets overlap, and observers recur. General variance difficulties for fold-based performance estimates are discussed by Bengio and Grandvalet (2004).

Bootstrap methods provide a framework for resampling-based uncertainty calculations (Efron, 1979), but a crossed observer-stimulus record requires its dependence structure to be respected. McCullagh (2000) examines resampling for exchangeable arrays, and Owen (2007) develops separate row-and-column resampling for crossed, unbalanced data. Owen and Eckles (2012) investigate independent factor reweighting under stated conditions. We use that crossed-resampling idea as a working analysis, without assuming that its results automatically provide exact coverage for this nonlinear, sparse calibration pipeline or a population of viewers.

The frozen calculation generates 499 replicates with independent multinomial multiplicities for observer identities and source-row components. A fixed PCG64 stream uses seed 86420261118. Each replicate recomputes peer references, training offsets, held-out predictions and both losses. The reference excludes all copies of the evaluated original observer, and source-component multiplicities weight both calibration and evaluation. Nonempty observer losses are averaged with observer multiplicities. If an active observer-fold has no selected calibration component, its offset is zero, the planned uncalibrated fallback; the test observations remain included.

We retain every replicate and every fallback flag. An absent peer reference or nonfinite calculation would remain undefined without redrawing, and any undefined replicate would make the designated uncertainty summaries unavailable. When all replicates are defined, the reported working SE is their sample standard deviation and the interval is the 2.5 th to 97.5 th percentile with linear interpolation. Bootstrap standard errors, interval construction and diagnostics are distinct choices (Davison and Hinkley, 1997). No interval is recentered, replaced or selected after observing its relation to the original estimate.

Average improvement and individual worsening

Equal-observer MSE decreased from 0.787228 for the peer baseline to 0.552414 after additive calibration, giving a signed change of -0.234813 score² (Table 1). MAE decreased from 0.706446 to 0.565226, a change of -0.141220 score. These are conditional predictions of individual pair means, not errors against independent quality truth. Each estimate uses 233,048 pairs, 1,257 observer codes, 1,596 stimuli and 133 source components, with all 233,168 raw ratings represented. There was no main-data fallback or missing peer reference.

Table 1. Actual same-support held-row losses and full-pipeline crossed resampling. Every row uses the same 1,257 observers and 233,048 pairs. Intervals are conditional working percentile summaries, with unestablished population coverage. Their displacement from some original estimates is retained and examined below; no centering adjustment was applied.

Quantity	Original estimate	Working resample SE	Unmodified percentile interval
Baseline MSE, score²	0.787228	0.019955	[0.756756,0.836028]
Calibrated MSE, score²	0.552414	0.015933	[0.555945,0.618894]
ΔMSE, score²	-0.234813	0.012903	[-0.235003,-0.184903]
Baseline MAE, score	0.706446	0.009818	[0.689769,0.729084]
Calibrated MAE, score	0.565226	0.008491	[0.568365,0.602475]
ΔMAE, score	-0.141220	0.006984	[-0.137270,-0.111133]

The average improvement concealed 211 observers with greater MSE. Another 1,046 had lower MSE, and none was tied under the prespecified absolute 1e-12 score² numerical tolerance. Individual changes ranged from -3.259467 to 0.574557, with median -0.113798. These differences describe observed prediction errors under the two fixed rules. They do not identify good and bad judges, establish why a particular observer benefited, or provide individual confidence intervals. Figure 2 shows the complete individual distribution rather than selecting favorable examples.

Table 2. All five prescribed folds. Folds 1-3 contain 27 components and 324 stimuli each; folds 4-5 contain 26 components and 312 stimuli each. Fold losses use their present observers equally and are not averaged to reconstruct Table 1.

Held-row fold	Pairs	Observers	Baseline MSE	Calibrated MSE	ΔMSE
1	47,906	1,029	0.806645	0.564511	-0.242133
2	47,059	1,022	0.792017	0.564651	-0.227367
3	47,902	1,056	0.779024	0.554311	-0.224714
4	45,403	1,022	0.789104	0.548825	-0.240280
5	44,778	998	0.775982	0.542678	-0.233304

Every fold's MSE difference was negative, from -0.242133 to -0.224714 (Table 2). This consistency concerns the fixed partition, not five independent experiments. The additive model also produced 124 predictions below one and 314 above five; its prediction range was 0.352076-5.575703. The peer baseline remained within 1.838462-4.020690. No out-of-range prediction was truncated or excluded from evaluation. The unbounded additive rule can improve its chosen loss while remaining unsuitable as a model constrained to the response scale.

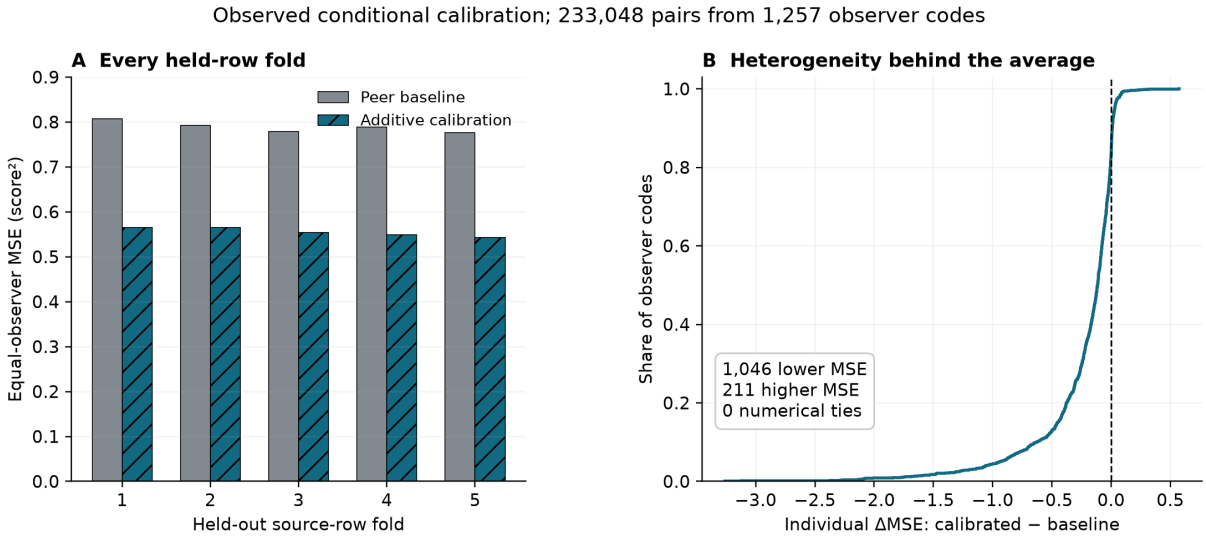

Figure 2. Panel A retains every prescribed fold with a zero-based loss axis. Panel B plots the empirical cumulative distribution of all 1,257 individual changes; the dashed zero line separates lower and higher MSE. Bar hatching and labels supplement color. Individual changes are not estimates of skill.

The duplication sensitivity used each pair's first source-row rating instead of its mean and rebuilt the full pipeline on identical folds. MSE then changed from 0.787298 to 0.552482, with ΔMSE -0.234816; MAE changed from 0.706497 to 0.565266, with ΔMAE -0.141231. Thus the two declared pair conventions gave similar aggregate comparisons in this record. That does not adjudicate whether the 120 repeated pairs were duplicate exports or repeated judgments. Both conventions and their full outputs are retained, and the mean-pair result remains primary.

The resampling distribution changes support

All 499 replicates were defined and none was redrawn. However, every replicate used at least one planned zero-offset fallback. Across all replicates there were 20,025 fallback observer-fold combinations, with a maximum of 79 in one replicate. The original data had none. The smallest number of nonempty observer identities in a replicate was 757. These are consequences of resampling a sparse crossed record with multiplicities: some identities and components disappear while others receive repeated mass. They are material properties of the executed uncertainty calculation, not implementation failures to be silently repaired.

The resampling distribution was displaced relative to several original statistics. Original calibrated MSE, 0.552414, was below its percentile interval [0.555945, 0.618894]. Original calibrated MAE, 0.565226, was also below [0.568365, 0.602475]. The original ΔMAE, -0.141220, lay below [-0.137270, -0.111133], while ΔMSE, -0.234813, remained near the lower end of [-0.235003, -0.184903]. Figure 3 displays both signed changes and their original estimates, preserving the difference between squared and absolute error.

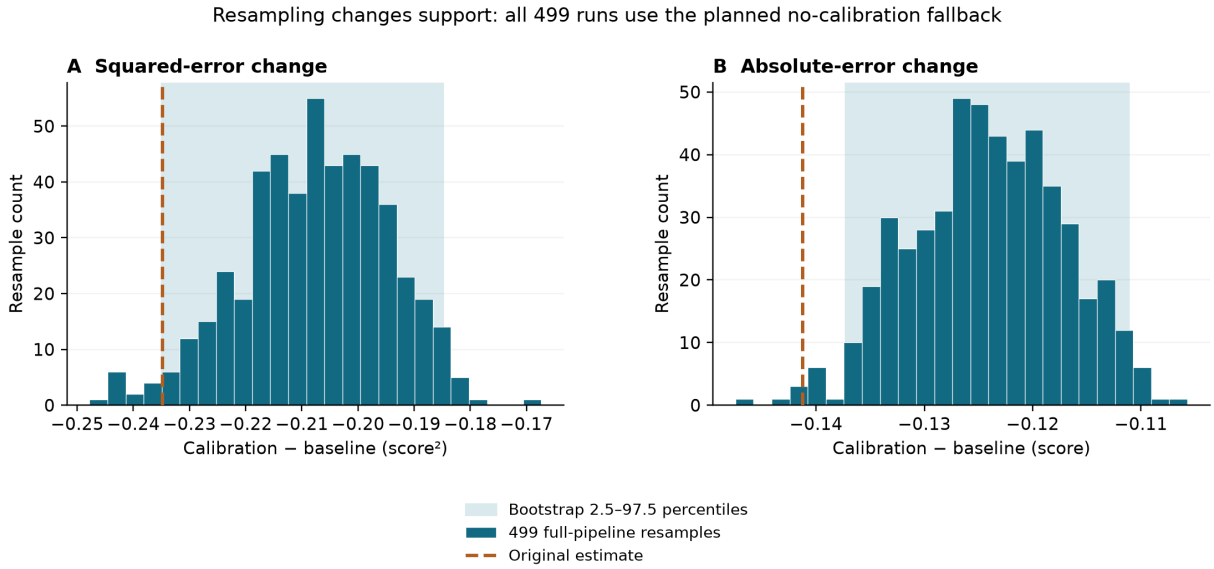

Figure 3. Histograms include all 499 full-pipeline replicates. Shading marks the unmodified percentile bounds, and the dashed line marks the original estimate. Support and calibration availability change under resampling. An original estimate outside a percentile interval does not by itself prove invalid coverage; coverage for this task remains unestablished.

The shifted intervals should be interpreted with their algorithm and support, rather than as simple symmetric perturbations around the original fit. Peer references and offsets are re-estimated under different identity and component weights, including the declared fallback where calibration disappears. This changes the realized prediction task within a replicate. We do not infer from interval displacement alone that a mathematical coverage theorem has been disproved. Equally, successful computation does not establish such coverage. No bias correction, alternative split or more favorable interval replaces the frozen outputs.

Evidence boundaries and computational checks

The study identifies a transfer property of additive offsets within one released quality-rating design. It does not separate perceptual differences, response style, common presentation context and genuine disagreement. Other observers' labels remain available on the test videos, and observers recur in calibration. Unknown timing prevents a prospective claim, while the missing page structure leaves possible context shared across source rows. The ordinal-code spacing assumption, unequal sparse incidence and 20 stimuli absent from the supplied aggregate table further constrain interpretation. No observer is removed for worsening, scale use or lack of conformity.

Reproducibility requires preserving how the responses become estimates, rather than retaining only a final model (Peng, 2011). We retain the minimal raw projection locally, every pair's constituents, fold assignments, predictions, loss summaries, resampling weights and fallback flags. Source and derived-file hashes accompany the code. The computational practices advocated by Sandve et al. (2013) motivate those retained links; they do not constitute independent scientific approval or unrestricted permission to redistribute the original annotation archive.

An independent calculation reread the CSV with standard-library parsing, reconstructed graph components and folds, and computed original peer references by explicitly summing other observers. Scalar summation independently reproduced both pair conventions. A different sorted-group reducer then refitted all 499 weighted pipelines and verified the random-weight stream, losses, fallbacks and intervals. The verification passed 11,460 scalar comparisons, 4,894,008 pair-table cell comparisons and 1,632,176

source-provenance cell comparisons. This establishes numerical agreement for the retained data and procedures, without verifying the source observers or latent visual quality.

The primary implementation used Python 3.12.14, NumPy 2.5.4 for arrays (Harris et al., 2020) and pandas 3.0.6 for numeric tables (McKinney, 2010). Matplotlib 3.11.2 generated the model and empirical figures (Hunter, 2007). Scientific-writing, statistical-analysis and scientific-visualization procedural skills guided evidence reporting and figure review, with their actual library credited to Kassis et al. (2026). These tools were used for this local analysis; no specialized visual-quality network, contest index system, reinforcement-learning policy or new participant experiment was executed.

Conclusion

An observer's additive offset estimated on other source rows lowered average conditional rating error in this public record. The result coexisted with 211 worsening observers, 438 out-of-scale predictions and substantial support changes during crossed resampling. Reporting those outcomes together answers a narrower question than whether observers are accurate or videos have objectively measured quality. The evidence supports within-cohort transfer of a simple numeric scale adjustment, while its unbounded predictions, reference dependence and unestablished interval coverage limit broader use.

Declarations

This is an AI-assisted empirical manuscript draft. Accountable human source, scientific-claim, authorship, funding, competing-interest and submission review remains PENDING. Only public numeric ratings and necessary pseudonymous grouping keys were processed locally; no raw observer IDs or personal context fields are displayed in this manuscript. The source's research-availability wording is preserved without inventing an open annotation license. No full original raw archive is redistributed. The accompanying delivery includes separate PDF and originality reports; these checks do not constitute human scientific approval.

References

1. Peng, Roger D. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
2. 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
3. Gneiting, Tilmann. (2011). Making and Evaluating Point Forecasts. Journal of the American Statistical Association. https://doi.org/10.1198/jasa.2011.r10138
4. Bengio, Yoshua; Grandvalet, Yves. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Journal of Machine Learning Research, 5, 1089-1105. https://jmlr.org/papers/v5/grandvalet04a.html
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. https://doi.org/10.48550/arXiv.2609.00065
6. Hyndman, Rob J.; Koehler, Anne B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2006.03.001
7. Janowski, Lucjan; Pinson, Margaret H. (2015). The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2015.2484963
8. International Telecommunication Union, Telecommunication Standardization Sector. (2022). Subjective video quality assessment methods for multimedia applications. Recommendation ITU-T P.910 (07/2022), historical edition. https://www.itu.int/rec/T-REC-P.910-202207-S/en
9. Hand, David J. (2004). Measurement Theory and Practice: The World Through Quantification. Wiley. https://www.wiley-vch.de/de/fachgebiete/mathematik-und-statistik/measurement-theory-and-practice-978-0-470-68567-9

10. 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

11. Davison, A. C.; Hinkley, D. V. (1997). Bootstrap Methods and their Application. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

12. Efron, Bradley. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. https://doi.org/10.1214/aos/1176344552

13. Davani, Aida Mostafazadeh; Díaz, Mark; Prabhakaran, Vinodkumar. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00449

14. 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.

15. Götz-Hahn, Franz; Hosu, Vlad; Lin, Hanhe; Saupe, Dietmar. (2021a). KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3077642

16. Tiotsop, Lohic Fotio; Servetti, Antonio; Barkowsky, Marcus; Masala, Enrico. (2024). Modeling Subject Scoring Behaviors in Subjective Experiments Based on a Discrete Quality Scale. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2024.3382483

17. McCullagh, Peter. (2000). Resampling and Exchangeable Arrays. Bernoulli. https://doi.org/10.2307/3318577

18. Li, Zhi; Bampis, Christos G.; Janowski, Lucjan; Katsavounidis, Ioannis. (2020). A Simple Model for Subject Behavior in Subjective Experiments. Electronic Imaging. https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-131

19. Götz-Hahn, Franz; Hosu, Vlad; Lin, Hanhe; Saupe, Dietmar. (2021b). The Konstanz 150k in-the-Wild Video Database (KonVid-150k). Dataset documentation and subjective annotation release, acquired 17 November 2026. https://database.mmsp-kn.de/konvid-150k-vqa-database.html

20. Owen, Art B. (2007). The pigeonhole bootstrap. The Annals of Applied Statistics. https://doi.org/10.1214/07-AOAS122

21. 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; Fernández del Río, Jaime; 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.

22. Owen, Art B.; Eckles, Dean. (2012). Bootstrapping data arrays of arbitrary order. The Annals of Applied Statistics. https://doi.org/10.1214/12-AOAS547

23. 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

24. Hunter, John D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55

25. 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

26. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301

27. Peroni, Leonardo; Gorinsky, Sergey; Tashtarian, Farzad; Timmerer, Christian. (2023). Empowerment of Atypical Viewers via Low-Effort Personalized Modeling of Video Streaming Quality. Proceedings of the ACM on Networking. https://doi.org/10.1145/3629139

28. 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

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

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.