Population Variation and Equity
Volume 1 · Number 1 · 2026

Which Images Enter the Shortlist? Raw Votes and Released MOS in KonIQ-10k

Jeffrey Franklin[1] · Nicholas Murphy[2] · Evan Anderson[3]

[1] Denmark · [2] Norway · [3] Italy

Original publication: November 3, 2026 | Article 658

Abstract

Quality filtering often ends with a list rather than a correlation coefficient. In 10,073 KonIQ-10k records, we compared the top thousand images ranked by reconstructed raw-vote mean with those ranked by the released MOS field. The lists shared 803 images: 197 left and 197 entered when the scoring field changed. Within a fixed raw top-two-thousand pool, rating entropy had an uncertain association with membership switching after score curvature and vote-count adjustment. The result identifies a consequential label choice without proving which scoring procedure better reflects human quality. Source-field definitions and the intended selection task should accompany image-quality shortlists.

Exploratory public-metadata manuscript; no new quality model or human-rating study is claimed. Human publication review remains outstanding.

Which Images Enter the Shortlist? Raw Votes and Released MOS in KonIQ-10k

Exploratory public-metadata manuscript; no new quality model or human-rating study is claimed. Human publication review remains outstanding.

Abstract

Quality filtering often ends with a list rather than a correlation coefficient. In 10,073 KonIQ-10k records, we compared the top thousand images ranked by reconstructed raw-vote mean with those ranked by the released MOS field. The lists shared 803 images: 197 left and 197 entered when the scoring field changed. Within a fixed raw top-two-thousand pool, rating entropy had an uncertain association with membership switching after score curvature and vote-count adjustment. The result identifies a consequential label choice without proving which scoring procedure better reflects human quality. Source-field definitions and the intended selection task should accompany image-quality shortlists.

1. The operational object is a set

An image-quality workflow may rank thousands of candidates and retain only a fixed number. In that setting, the practical output is membership in a shortlist. Two scoring fields can be highly related yet disagree at the selection boundary. A global correlation therefore cannot answer how many actual images change. We examine this issue by reconstructing a raw-vote mean and comparing its selected set with the set obtained from a separately released MOS column.

KonIQ-10k provides quality-rating metadata for images with authentic distortions (Hosu et al., 2019). The author repository contains five rating-distribution columns, a total response count, and a MOS field. We analyze that public table as supplied. Its distribution columns sum to one and reconstruct nearly integer response counts after multiplication by the total. This permits a directly checkable raw mean without access to individual observer identities or a new annotation exercise.

The motivation connects to the distinction between semantic and perceptual information in Meng, 2026. A quality model can be carefully designed yet still inherit the operational meaning of its supervision labels. Our analysis asks a prior question about label-based selection. It does not implement the target model, evaluate generated images, or establish that a particular representation can resolve discrepancies between two aggregate rating fields.

2. Related measurement and decision questions

NIMA treats image assessment through rating information (Talebi and Milanfar, 2017), while MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022) develop different visual prediction architectures. Language-image features in Wang et al., 2022 supply another route to quality assessment. These works are relevant because their evaluation ultimately depends on a chosen quality target. Our result concerns that target's consequences for a shortlist, rather than an accuracy comparison among the architectures.

AGIQA-3K separates perceptual quality from alignment for generated imagery (Li et al., 2023), and Kirstain et al., 2023 collects human preferences in another generation setting. These examples show that quality-related labels can arise from different judgment tasks. We do not pool their scores with KonIQ or treat their observer procedures as interchangeable. The present study remains within one authentic-distortion release, where even two fields attached to the same image can produce different retained sets.

The natural-scene-statistics approach of Mittal et al., 2013 is an example of a visual-quality score not reconstructed from these votes. It could later be compared with shortlist membership, but it was not computed here. Similarly, probability calibration in Guo et al., 2017 addresses agreement between probabilistic predictions and observed outcomes. Neither a raw score rescaling nor a membership-switch regression is automatically a calibration assessment of a neural quality predictor.

3. Research questions and definitions

The first research question is a finite-release audit: how many members of a top-thousand set change when ranking by the reconstructed raw-vote mean instead of the released MOS field? We choose one thousand as a declared operational budget and use the same number for both lists. This does not optimize a threshold against an outcome. Stable image-name ordering breaks ties, ensuring that the membership comparison is reproducible even at an equal-score boundary.

The second question asks whether rating dispersion, represented by normalized entropy, accompanies switching within a prespecified candidate pool. We define that pool as the top two thousand by raw mean. It concentrates attention near the upper part of the quality ranking without fitting a model over images that have virtually no chance of entering either list. This conditional sample must remain visible in the interpretation: its coefficients do not describe all 10,073 images equally.

Agreement concepts in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 help distinguish the objects involved. Two fields can agree on broad ordering while disagreeing about which images cross a cutoff. A high reliability statistic does not guarantee identical decisions. We therefore report set intersection and union explicitly. No intraclass correlation is estimated, and the main set result does not require an assumption that either numerical scale represents a universally correct latent quality unit.

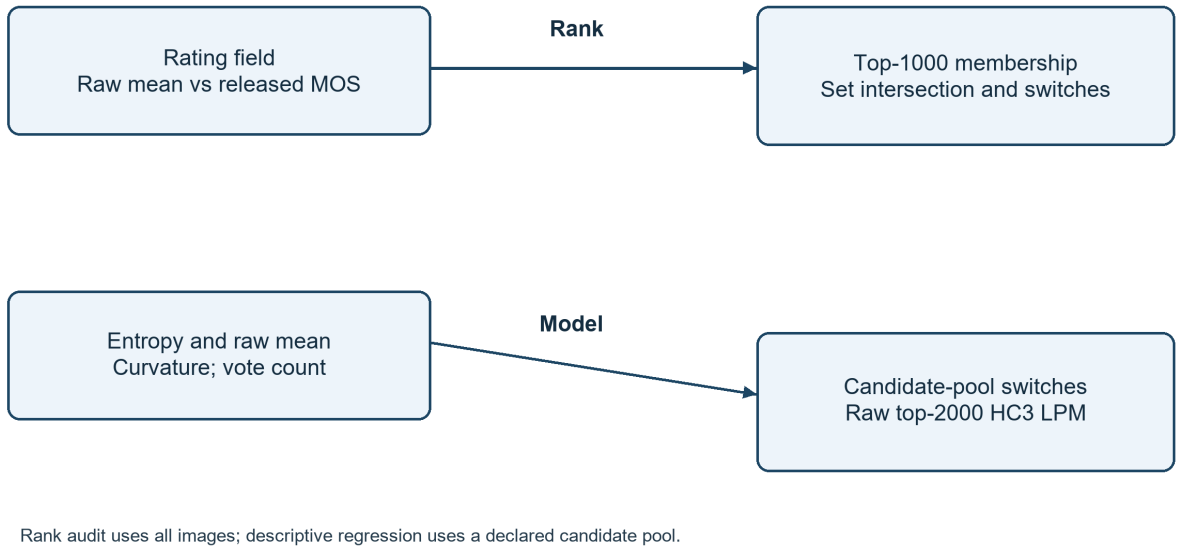


Figure 1. The chosen rating field determines top-thousand membership; within a fixed raw top-two-thousand pool, entropy, raw-score curvature, and response count describe membership switching.

4. Reconstructing what the file supports

The table contains 10,073 unique image names. For each row, the five c columns are interpreted as proportions of ratings one through five because they sum to one and yield integer-like counts after multiplication by c_total. The largest fractional-count residual is about eight times ten to the minus eleven. We compute the raw mean as the sum of category value times proportion. We do not mistakenly treat the proportions themselves as raw counts.

The reconstructed raw means range from 1.0962 to 4.3100. The supplied MOS field ranges from 3.9118 to 88.3889. Their units plainly differ. We compare rankings, which are unaffected by an increasing affine rescaling, rather than subtracting one field from the other. The exact observer-level transformation behind the released MOS is not established by this analysis. We therefore do not label every difference a causal consequence of worker normalization or claim to have reproduced the original processing pipeline.

Normalized entropy is calculated from the five proportions as negative summed proportion times log proportion, divided by log five; zero-proportion contributions are defined as zero. This yields a bounded descriptive measure of response spread. It does not distinguish multimodal preferences from inconsistent individual observers. Measurement-error concerns in Carroll et al., 2006 and missing-data distinctions in Little and Rubin, 2002 caution against using one aggregate column as a complete account of the rating process.

The response-count covariate is the natural logarithm of c_total. Counts may be associated with rating precision, source scheduling, or filtering procedures. A coefficient on this variable cannot identify which explanation applies. The dataset lacks the full assignment history needed for that causal question. The descriptive model therefore conditions on the released counts while retaining uncertainty about their origin and their relationship to image difficulty.

5. Modeling a boundary-sensitive outcome

The binary outcome is one when an image belongs to exactly one of the two top-thousand lists. Within the raw top-two-thousand pool, a linear probability model includes entropy, raw mean, raw mean squared, and log response count. The quadratic score terms allow switching to vary nonlinearly with proximity to the selection region. They are adjustment terms, not a claim that membership has a globally quadratic causal response to perceived quality.

We use HC3 standard errors and t-based intervals for the 2,000 candidate images. White, 1980 motivates heteroskedastic covariance for a binary-outcome linear model. Cameron and Miller, 2015 highlights another limitation: shared observers or collection groups could produce dependence across images, but those clusters are not supplied in the processed table. HC3 does not correct that unobserved dependence. We do not manufacture observer identifiers or treat every recorded vote as a separate independent regression row.

A richer repeated-measure model of the kind motivated by Laird and Ware, 1982 would require observer-level information. Robust-estimation perspectives from Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 would change the influence assigned to unusual observations, but would not reveal the missing assignment process. We fit the stated ordinary linear probability model once. Its fitted probabilities are used only for descriptive association, and no claim of an operationally calibrated switch predictor is made.

Table 1. Candidate-pool associations with membership switching. Raw-mean curvature and log response count are adjustment variables; intervals are HC3 and do not account for unidentified shared raters.

Model	Term	Estimate (SE)	95% CI	N
M1	Normalized rating entropy	-0.0915 (0.1369)	[-0.3600, 0.1771]	2000
M1	Raw vote mean	33.7693 (5.2281)	[23.5162, 44.0225]	2000
M1	Raw vote mean squared	-4.4511 (0.6826)	[-5.7897, -3.1125]	2000
M1	Log response count	0.7330 (0.1948)	[0.3510, 1.1150]	2000
M1: fit	R-squared = 0.0567	-	-	2000

Model labels: M1 = Shortlist-switch associations in candidate pool.

6. Shortlist changes and uncertain entropy evidence

The two top-thousand sets share 803 images. There are 197 raw-only members and 197 released-MOS-only members. The union contains 1,197 images, giving Jaccard overlap 0.6708. In the full table, 394 images switch membership. These are exact counts for the acquired release and the declared tie rule; no standard error is needed to describe the finite sets themselves.

The raw top-two-thousand pool contains 391 of the 394 switching images. Three released-MOS-only entrants lie outside that candidate pool. This matters because the regression is not a complete model of every switch. We report the global count separately and retain the three exceptions in the all-image processed table. Expanding the pool after observing them would answer a different conditional question and is not silently substituted for the planned analysis.

The entropy coefficient is -0.0915, with standard error 0.1369 and interval -0.3600 to 0.1771. The data therefore do not clearly establish the direction of its adjusted association in the candidate pool. One should not infer that disagreement is irrelevant, nor claim it decisively predicts instability. The coefficient conditions on raw-score curvature and response count, and entropy itself summarizes only the supplied category proportions.

The log-count coefficient is 0.7330, with interval 0.3510 to 1.1150. Raw-score terms show marked curvature, and the model R-squared is 0.0567. The large raw linear and squared coefficients should be interpreted jointly over the observed score range. Extrapolating either term alone, or interpreting the intercept at a raw mean of zero, would be meaningless for this upper-quality candidate sample. The archived table retains all terms so that adjustment is transparent.

7. Implications for a quality-filtering pipeline

The switch count shows that the choice of rating field can materially alter a fixed-size retained collection. It does not establish which collection is preferable. A user might value agreement with a normalized observer consensus, preservation of raw-category meaning, or a downstream task measured by a separate outcome. Those goals require different evidence. The present audit identifies a decision that must be documented before comparing models trained on different target columns.

Class-imbalance and threshold discussions in Saito and Rehmsmeier, 2015, ROC analysis in Fawcett, 2006, and cost-sensitive decision theory in Elkan, 2001 all emphasize that an operating decision is not fully described by one overall fit statistic. Our top-thousand budget is not a binary truth label, however.

Switching is a property of two procedures, not proof that one produced an error. No precision, recall, or harm rate for an allegedly correct quality set is inferred from these counts.

Influence reasoning following Cook, 1977 could guide inspection of boundary cases, and causal diagrams in Greenland et al., 1999 could help organize a future observer-normalization study. Neither resolves the current identification limit without new information. The guidance in Wasserstein and Lazar, 2016 and concerns about selective findings in Ioannidis, 2005 support retaining the uncertain entropy result. It would be misleading to replace it with a more favorable specification merely to make a simple narrative about disagreement causing list instability.

8. Reproducible boundaries

The shortlist budget deserves explicit interpretation. Retaining one thousand images from this release is a convenient audit scenario, not a scientifically validated quality threshold. A different collection size or selection budget could move the boundary into a denser or sparser score region. We therefore retain the rank inputs and membership indicators rather than suggesting that the observed replacement percentage is a fixed property of KonIQ under every possible filter. A new budget would be a new finite-set calculation, not an additional independent study.

Nor does an image leaving the list imply that its appearance changed. The images are identical across the two ranking procedures; only the field used to order them changes. This makes the comparison especially useful for data-pipeline audits. An analyst can discover a consequential metadata choice before training a model or spending resources collecting downstream outcomes. It also limits the claim: the audit cannot diagnose which visual features are misrepresented, because it never observes a disagreement between the image and an independently validated notion of quality.

The conditional regression should be read with the same care. Entropy and the raw mean arise from the same vote proportions, so their association is structurally constrained. Adjusting for score curvature changes the question from a simple entropy-switch correlation to a comparison among images with similar raw means and response counts. That is a defensible descriptive question, but it does not isolate a manipulable treatment. Collecting additional independent ratings would be necessary to test whether reducing uncertainty in the rating distribution changes the stability of the selected set.

The package preserves the author-repository metadata, reconstructed means, entropy values, list indicators, and conditional analysis rows. It follows Gebru et al., 2021 by documenting source fields and follows Wilkinson et al., 2016 by retaining accessible provenance and local checksums. Public code-repository availability is distinguished from image-level licensing. No image pixels, observer identities, target-model outputs, or inferred personal attributes are required to recompute the reported memberships.

The result is a practical measurement warning with a limited empirical scope. Under one fixed selection budget, 19.7 percent of either thousand-image list is replaced when the ranking field changes. The conditional model offers little decisive evidence that entropy explains those switches. Future work should verify the complete MOS transformation and connect retained images to an independent application outcome. Until then, the honest conclusion concerns shortlist sensitivity to supplied rating fields, not superiority of one aggregation rule.

References

1. Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.

2. Meng, Zijie. (2026). Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment. arXiv:2606.16799. https://arxiv.org/abs/2606.16799.

3. Hosu, Vlad, et al. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.

4. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

5. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

6. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

7. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

8. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

9. Li, Chunyi, et al. "AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment" *arXiv preprint arXiv:2306.04717*, 2023. https://arxiv.org/abs/2306.04717.

10. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

11. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

12. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

13. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

14. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

15. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

16. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

17. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

18. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

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

20. Kirstain, Yuval, et al. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation" *arXiv preprint arXiv:2305.01569*, 2023. https://arxiv.org/abs/2305.01569.

21. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

22. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.

23. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

24. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

25. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

26. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

27. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

28. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

29. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

30. Wang, Jianyi, et al. "Exploring CLIP for Assessing the Look and Feel of Images." *arXiv preprint arXiv:2207.12396*, 2022. https://arxiv.org/abs/2207.12396.