Uncertainty and Assessment
Volume 1 · Number 1 · 2026

Which Aesthetic Images Count as Disputed? Variance and a Midscale Split

Joseph Hayes[1] · David Clark[2] · Matthew Fisher[3]

[1] Croatia · [2] Estonia · [3] Latvia

Original publication: October 9, 2026 | Article 652

Abstract

Shortlisting the most disputed images depends strongly on how disagreement is defined. Among 211,576 eligible AVA images, the highest-variance tenth and the highest midscale-split-disagreement tenth each contained 21,157 images but shared only 2,612. Their Jaccard overlap was 0.065790. Split disagreement was positively associated with variance after mean-rating, vote-count, and semantic-tag adjustment, yet that association did not make the shortlists interchangeable. Both summaries came from the same vote histograms and therefore share mathematical information. This analysis concerns aesthetic-rating distributions, not uncertainty of a visual assessment model or evidence that one disagreement definition better captures human judgment.

Two explicitly defined summaries of the same public vote histograms, with mean ratings restricted to a common middle range.

Which Aesthetic Images Count as Disputed? Variance and a Midscale Split

Two explicitly defined summaries of the same public vote histograms, with mean ratings restricted to a common middle range.

Abstract

Shortlisting the most disputed images depends strongly on how disagreement is defined. Among 211,576 eligible AVA images, the highest-variance tenth and the highest midscale-split-disagreement tenth each contained 21,157 images but shared only 2,612. Their Jaccard overlap was 0.065790. Split disagreement was positively associated with variance after mean-rating, vote-count, and semantic-tag adjustment, yet that association did not make the shortlists interchangeable. Both summaries came from the same vote histograms and therefore share mathematical information. This analysis concerns aesthetic-rating distributions, not uncertainty of a visual assessment model or evidence that one disagreement definition better captures human judgment.

1. A disputed-image label needs a decision rule

An image with a wide spread of aesthetic ratings can be called disputed. So can an image that receives nearly half its votes on each side of a chosen dividing score. These descriptions sound similar, but they ask different questions. Variance uses distances among all numeric score levels. A binary split discards distances within each side and asks how evenly the votes cross one boundary. Their resulting image lists need not identify the same cases.

We ask whether these two operational definitions select the same high-disagreement images after restricting average scores to a middle band. We also describe their association while accounting for the mean, number of votes, and released semantic tag. The first question concerns finite shortlist overlap. The second is a descriptive regression of one histogram summary on others. Neither assumes that disagreement is a single underlying construct revealed perfectly by one convenient statistic.

The distinction between image semantics and distortion in Meng, 2026 motivates clarity about what an evaluation target contains. Our source comprises aesthetic opinions on photographs, rather than separate semantic-alignment and technical-distortion labels. No visual-language model is run. The analysis therefore cannot validate the cited model or determine its uncertainty. It examines how a benchmark label can change before any predictive assessment method is introduced.

2. Public histogram data and the middle-score cohort

We use the AVA metadata associated with Murray et al., 2012, containing 255,528 unique image identifiers and counts for rating bins one through ten. We verify nonnegative counts and positive totals. Every image's mean is reconstructed directly from those counts, rather than substituted from a separately normalized MOS column. The finite vote variance uses the total number of captured votes as its denominator. It is a distributional summary of the supplied histogram, not an estimate of the variance of an image's mean score.

Eligibility requires at least one hundred votes and a reconstructed mean between 4.5 and 6.5 inclusive. The resulting cohort contains 211,576 images. These are declared restrictions designed to reduce the obvious contrast between strongly low- and high-scoring images. They do not force identical means, vote

totals, or histogram shapes. The results apply to this cohort and should not be relabeled a complete analysis of every AVA photograph.

Eligible images average 208.399 captured votes, mean rating 5.444427, and vote variance 2.038734. These values are averages over images, not pooled quantities weighted by all votes. The released first semantic tag defines sixty-six grouping categories, retaining an untagged category where it occurs. We do not infer the identities of photographers or individual judges from these codes. A semantic tag is neither a rater group nor evidence of independent observers.

Measurement-scale issues associated with Stevens, 1946 matter because variance assigns numerical meaning to distances among the ten rating bins. We follow the supplied numeric scoring convention while retaining that interpretation as a choice. The binary split makes a different choice by treating all ratings within a side identically. Neither transformation can be said to preserve every feature of the original opinions. The choice should follow the research question a later evaluation intends to answer.

3. Two disagreement definitions and equal-sized shortlists

For variance, each score's squared distance from its image mean is multiplied by that bin's count, summed, and divided by total votes. This uses all ten bin locations. An image with a small fraction of very distant opinions can have substantial variance even if a clear majority falls on one side of the midscale boundary. Conversely, many votes concentrated near adjacent central bins may straddle that boundary without producing an exceptionally wide numerical spread.

The second measure divides votes into bins one through five and bins six through ten. Split disagreement is twice the smaller side's share. It equals zero when all votes lie on one side and one when the two sides are equal. The boundary between five and six is a declared middle-scale dichotomy. It is not a validated threshold for image acceptability, professional aesthetic quality, or a viewer's decision to use an image.

For each measure we select exactly the highest integer tenth of eligible images, giving 21,157 entries. Scores sort in descending order and source filenames resolve ties. The variance cutoff has 21,156 strictly higher images and one image at the boundary. The split cutoff has 21,123 strictly higher images and thirty-four at the boundary, all of which enter the selected list. Thus the limited overlap in this realization is not produced by selecting only an arbitrary fraction of a large tied boundary block.

Agreement research in Fleiss, 1971, Shrout and Fleiss, 1979, and Koo and Li, 2016 addresses related but distinct questions about judgments and reliability. We calculate no kappa, intraclass coefficient, or observer-specific reliability estimate. Marginal vote histograms can support some aggregate agreement calculations under appropriate assumptions, but they do not restore missing individual judge identities and covariance. Calling the two current scores reliability coefficients would misstate both their definitions and what was measured.

4. The image lists overlap little

The two lists share 2,612 images. Each has 18,545 entries absent from the other, and their union contains 39,702 images. Intersection divided by union gives a Jaccard overlap of 0.065790. The shared fraction relative to either single list is approximately 12.3 percent. We state both denominators because an overlap ratio is incomplete without identifying whether its reference is a list or their union.

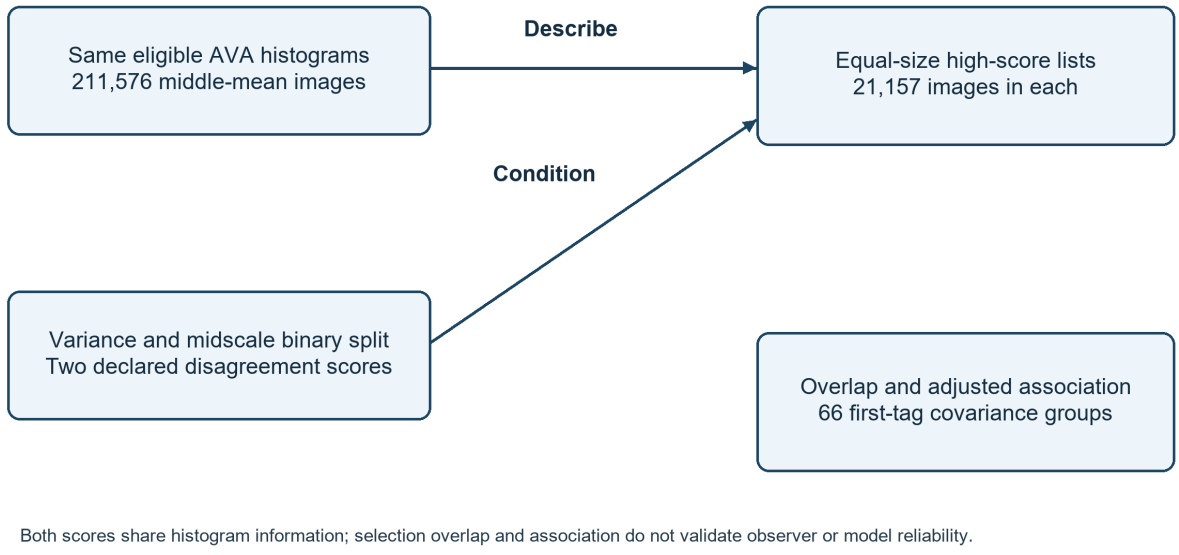

Figure 1. The same eligible histograms produce variance and midscale-split scores, equal-sized lists, and a descriptive association adjusted for other histogram properties.

This is an exact finite-cohort comparison after the declared restrictions, score formulas, and tie convention. We do not attach a sampling confidence interval to the overlap as though the data had been drawn through an independently specified random photograph-sampling design. A new corpus or a changed eligibility band could produce another overlap. The observed mismatch remains useful without claiming that its precise percentage is a universal property of aesthetic disagreement.

The agreement perspectives in Bland and Altman, 1986 and Lin, 1989 reinforce that a positive association between scores does not establish interchangeable rankings or decisions. We do not estimate their particular agreement statistics here. Instead, the equal-list comparison measures the selection consequence directly. The two summaries can move together over broad parts of the data while prioritizing very different images at the upper end of their respective scales.

5. Association with controls for the shared histogram information

The first ordinary least-squares fit contains an intercept and summarizes mean split disagreement. The second models split disagreement using standardized variance, mean rating, its square, log vote count, and first-semantic-tag indicators. Variance is standardized using its population standard deviation within the eligible cohort. The mean polynomial allows a simple nonlinear relationship with the location of the histogram. Both the predictors and outcome are computed from the same released votes.

Covariance clusters by the sixty-six first-tag categories, with a finite-sample correction and a t reference using sixty-five degrees of freedom. Dependence considerations in Cameron and Miller, 2015 motivate acknowledging repeated content groupings. Heteroskedasticity correction associated with White, 1980 alone would not represent this structure. First-tag grouping still omits other shared content, photographers, and assessors. It is an available approximation rather than a certification that all images or votes are independent.

Table 1. Mean split disagreement and descriptive histogram associations in the eligible AVA cohort.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.7243 (0.0015)	[0.7213, 0.7272]	211576
M1: fit	R-squared = 0.0000	-	-	211576
M2	Vote variance, per SD	0.0347 (0.0004)	[0.0339, 0.0356]	211576
M2	Mean rating	6.1754 (0.0561)	[6.0634, 6.2875]	211576
M2	Squared mean rating	-0.5541 (0.0050)	[-0.5640, -0.5441]	211576
M2	Log vote count	0.0074 (0.0010)	[0.0055, 0.0093]	211576
M2: fit	R-squared = 0.8255	-	-	211576

Model labels: M1 = Mean midscale split disagreement; M2 = Variance association adjusted for mean and vote count.

The mean-only regression has zero R-squared. The adjusted model has R-squared 0.825466, which summarizes all its included mean, count, tag, and variance terms. That fit is not validation of a predictive model on independent labels. The summaries share their underlying histograms and can be mathematically constrained. A high in-sample fit can therefore coexist with limited evidence about an external construct, a different observer population, or future model uncertainty.

6. Positive association and decision disagreement coexist

Mean split disagreement is 0.724253790, with clustered standard error 0.001461905 and interval 0.721334163 to 0.727173416. The coefficient on standardized variance in the adjusted model is 0.034718738, with standard error 0.000423576 and interval 0.033872800 to 0.035564677. Thus higher variance is associated with somewhat greater split disagreement after the specified adjustment. The positive coefficient does not contradict the small shortlist overlap because it answers a different question.

The fitted mean terms are strongly nonlinear within the selected band: the linear coefficient is 6.175449 and the quadratic coefficient -0.554060. They should be interpreted jointly in the included score range, rather than as independent linear effects at impossible mean ratings. Log vote count has coefficient 0.007393578, with interval 0.005477240 to 0.009309915. These descriptive terms reflect relationships among finite histogram summaries; they do not identify effects of recruiting more observers or changing an image's mean rating.

The causal-design distinctions in Greenland et al., 1999 and Stuart, 2010 are pertinent here. Variance is not an independently assigned exposure, and an intervention described only as increasing variance leaves many possible vote-distribution changes unspecified. The regression cannot tell which change would occur. We report the association and selection mismatch without claiming that variance causes a particular binary split or that the controls isolate a unique mechanism of human disagreement.

Measurement reasoning in Carroll et al., 2006 also distinguishes numerical summaries from an error-free latent opinion. The same mean can accompany many distributions, and aggregation loses judge-specific information. Missing-data principles in Little and Rubin, 2002 do not justify inventing that information from the marginals. A different disagreement target might require raw responses, repeated ratings, or an explicit observer model, none of which is supplied by this histogram analysis.

7. Aesthetic dispersion is a different target from quality-model uncertainty

Assessment methods represented by Talebi and Milanfar, 2017 and Ke et al., 2021 supply broader context for image-score prediction. We run neither method and compare no model outputs. Preference-based generation evaluation such as Xu et al., 2023 concerns yet another target. The current disputed-image labels arise from transformations of existing human-rating histograms, before a model is scored. A visual assessor's error or confidence cannot be read directly from a high variance or an evenly split histogram.

The natural-quality resources in Hosu et al., 2019 and Fang et al., 2020 also differ from an aesthetic voting collection. Their labels and acquisition conditions do not enter this cohort. Predictive-scoring principles in Gneiting and Raftery, 2007 and error-measure considerations in Hyndman and Koehler, 2006 emphasize specifying the evaluation target before comparing methods. We do not recommend one disputed-image shortlist as a universal test set merely because it produces a larger numerical disagreement score.

A hierarchical approach motivated by Laird and Ware, 1982 could model repeated observers if suitable identities and response data were available. Shrinkage ideas associated with Efron and Morris, 1973 could address another question about finite-vote estimates. Neither is fitted in this report. Our score uses the captured vote distribution directly, without a model of a population consensus or a correction that turns each histogram into a high-confidence estimate of how every new audience would respond.

8. Restrictions, influence, and validation

The middle-mean band and vote minimum are part of the estimand. Expanding the range could make location differences dominate more strongly; changing the binary boundary would change split disagreement itself. We do not select another boundary after inspecting the overlap or choose a cohort that maximizes a preferred association. Validation concerns in Roberts et al., 2017 and selection bias in Varma and Simon, 2006 require independent assessment if these data are later used to design a model-evaluation protocol.

Influence reasoning in Cook, 1977 motivates preserving unusual histograms rather than deleting disputed cases as inconvenient outliers. Multiplicity considerations in Benjamini and Hochberg, 1995 limit broad inference from many possible score summaries, thresholds, and mean bands. Interpretation following Wasserstein and Lazar, 2016 keeps the exact overlap and the conditional regression separate from a claim of universal measurement superiority. The current results support the need to define disagreement clearly, without establishing which definition best serves an application.

The semantic-tag covariance model is another restriction. An image may have multiple content associations, while this analysis uses only the first recorded tag. Judges may have rated many images, and unknown photographer or acquisition similarities can persist across tags. The sixty-six groups do not eliminate those relationships. Additional raw-response or provenance data would be needed to model them, rather than treating the large image count as a substitute for observer-level information.

9. Publish the disagreement definition with the selected images

The completed analysis finds a positive adjusted relationship between vote variance and midscale-split disagreement, alongside markedly different high-disagreement lists. These observations illustrate two distinct uses of a score: describing broad variation and deciding which records enter an evaluation. A score name alone does not specify either use. A disputed-image benchmark should therefore preserve its exact histogram transformation and selection rule, including where its dividing boundary originates.

Documentation under Gebru et al., 2021 includes the numeric score bins, finite-variance denominator, binary split, eligibility band, minimum votes, source filename order, and first-tag grouping. Preservation following Wilkinson et al., 2016 retains the eligible histogram summaries and both list-membership indicators. They make the result reproducible without inventing judge identities or a model-confidence target. Establishing an externally valid notion of aesthetic disagreement or quality-model uncertainty requires additional scientific evidence beyond these two operational definitions.

References

1. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
2. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
3. Efron, Bradley; Morris, Carl. (1973). Stein's Estimation Rule and Its Competitors--An Empirical Bayes Approach. Journal of the American Statistical Association. https://doi.org/10.2307/2284155.
4. 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.
5. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
6. Murray, Naila; Marchesotti, Luca; Perronnin, Florent. (2012). AVA: A large-scale database for aesthetic visual analysis. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6247954.
7. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
8. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
9. Xu, Jiazheng, et al. "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation." *arXiv preprint arXiv:2304.05977*, 2023. https://arxiv.org/abs/2304.05977.
10. 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.
11. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
12. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
13. 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.
14. 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.
15. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
16. 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.
17. 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.

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. 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.

20. 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.

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

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

23. Benjamini, Yoav and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

24. Shrout, Patrick E.; Fleiss, Joseph L.. (1979). Intraclass correlations: Uses in assessing rater reliability.. Psychological Bulletin. https://doi.org/10.1037/0033-2909.86.2.420.

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

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

27. Fleiss, Joseph L.. (1971). Measuring nominal scale agreement among many raters.. Psychological Bulletin. https://doi.org/10.1037/h0031619.

28. 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.

29. Fang, Yuming; Zhu, Hanwei; Zeng, Yan; Ma, Kede; Wang, Zhou. (2020). Perceptual Quality Assessment of Smartphone Photography. CVPR 2020, 3677-3686. https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_Perceptual_Quality_Assessment_of_Smartphone_Photography_CVPR_2020_paper.html.

30. 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.