Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

When More Ratings Make Shrinkage Less Useful: An AVA Annotation-Budget Experiment

William Warren[1] · Mark Palmer[2] · Gavin Bell[3]

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

Original publication: October 6, 2026 | Article 617

Abstract

Borrowing a common rating prior helps most when little image-specific evidence is retained. We split AVA images by identifier, estimated a prior from the first half, and simulated nested five- and twenty-vote budgets in the remaining 127,764 images. Shrinkage with ten fixed pseudo-votes reduced five-vote mean absolute error against unused ratings from 0.506 to 0.420. At twenty votes, error changed from 0.265 to 0.266. The benefit also declined as the unused-vote mean moved farther from the prior. These findings describe one finite-histogram subsampling experiment and support budget-sensitive evaluation of aggregate labels, rather than a universal shrinkage rule or a claim about an image-quality model.

Simulated subsampling of real public rating histograms. This is not a newly collected observer study; human scientific review is pending.

When More Ratings Make Shrinkage Less Useful: An AVA Annotation-Budget Experiment

Simulated subsampling of real public rating histograms. This is not a newly collected observer study; human scientific review is pending.

Abstract

Borrowing a common rating prior helps most when little image-specific evidence is retained. We split AVA images by identifier, estimated a prior from the first half, and simulated nested five- and twenty-vote budgets in the remaining 127,764 images. Shrinkage with ten fixed pseudo-votes reduced five-vote mean absolute error against unused ratings from 0.506 to 0.420. At twenty votes, error changed from 0.265 to 0.266. The benefit also declined as the unused-vote mean moved farther from the prior. These findings describe one finite-histogram subsampling experiment and support budget-sensitive evaluation of aggregate labels, rather than a universal shrinkage rule or a claim about an image-quality model.

1. A quality label has an annotation budget

An aggregate image rating is often treated as a stable target once it appears in a dataset. Its precision nevertheless depends on how many judgments are retained and on how strongly they disagree. A common prior can stabilize an estimate based on a few judgments, but can also pull unusual images toward an inappropriate average. We examine that tradeoff using real vote histograms and an explicitly simulated annotation-budget restriction.

AVA contains aesthetic judgments for a large collection of photographs (Murray et al., 2012). Its released category counts permit draws of a limited number of ratings without inventing a response distribution. We do not recruit observers or create new judgments. Instead, we ask how a simple empirically centered shrinkage rule agrees with ratings left unused after a declared draw, and whether the answer changes between five and twenty retained judgments.

The image-quality representation problem addressed by Meng, 2026 motivates attention to the supervision target as well as model architecture. Our evidence concerns aesthetic judgments on photographs, rather than technical distortion labels on generated images. It does not test that framework. The connection is that a quality predictor inherits decisions about how human responses are aggregated, and those decisions can have different consequences under different annotation budgets.

2. Stabilization has a bias cost

The empirical-Bayes perspective of Efron and Morris, 1973 provides a useful motivation for borrowing information across related estimation problems. Our rule is much simpler than a fully specified hierarchical model: it combines an image's sample mean with one independently estimated global image-average mean using a fixed pseudo-count. We make no claim that a classical optimality theorem applies to bounded aesthetic ratings, absolute-error loss, or this finite-population sampling scheme.

Rating-distribution modeling in Talebi and Milanfar, 2017 and subjective-data designs in Fang et al., 2020, Gu et al., 2020, Kirstain et al., 2023, and Wu et al., 2023 illustrate different ways human judgments enter visual evaluation. Their judgment tasks are not interchangeable with AVA aesthetics. We do not pool their labels or treat a preference score and an aesthetic category mean as common numerical units. They motivate examining aggregation choices, not expanding the sample with incompatible studies.

Visual-quality systems such as Ke et al., 2021, Wang et al., 2022, and Yang et al., 2022 can be evaluated against aggregate labels, but no such predictor is trained here. The experiment evaluates the label estimator before introducing image features. If a shrinkage rule changes the target, a later model comparison should identify that change. Otherwise, an apparent improvement in predicting the supplied target may reflect a different aggregation convention rather than better understanding of image content.

3. Independent images for the prior, unused votes for evaluation

The acquired maintainer CSV has 255,528 unique image names and ten integer rating-count fields. We check nonnegative counts and reconstruct exact means from those fields rather than treating a rounded MOS column as the complete source of truth. Images are sorted lexicographically by name. The first 127,764 supply the prior, calculated as an equal-image average of their reconstructed means. The remaining 127,764 are candidate evaluation images.

This identifier partition prevents the evaluation images from contributing to the prior. It is not random assignment of images to a study population and need not be representative of future photography. The source's image identifiers can relate to its collection process. We therefore describe held-image evaluation within the captured resource, without claiming a prospective external validation or independence of all image content from the broader dataset's curation.

Every evaluation image has at least forty recorded votes, satisfying the declared eligibility requirement. For each, we draw twenty ratings without replacement from its real histogram using a fixed random generator. A nested five-rating sample is then drawn from those twenty. The common evaluation target is the mean of all ratings remaining after removal of the twenty, so neither retained budget contributes its votes to the target used for that image.

The unused-vote target is still an aggregate from the same original rating process. It is not a latent true aesthetic value or a newly independent observer panel. Within a fixed finite histogram, retaining some votes changes the composition of what remains. The experiment therefore measures performance under this explicit subsampling design, including that finite-population relation, rather than assuming that every draw comes from an infinite independent population.

4. A fixed shrinkage policy

The training-image prior mean is 5.3346 on the one-to-ten rating scale. For a retained budget n, the rule combines n times the sample mean with ten times that prior and divides by n plus ten. The pseudo-count of ten is fixed for both budgets. It is not selected to minimize evaluation error. The five-vote estimate therefore gives the prior more weight than the twenty-vote estimate, making the bias-variance tradeoff explicit.

The first proposition asks whether the mean absolute-error gain from this rule differs between budgets. Gain is the raw sample mean's absolute error minus the shrinkage estimate's error against the common unused-vote target. Positive gain favors shrinkage. A second, retrospective proposition asks how the gain varies with the target's absolute distance from the training prior. That distance uses evaluation information and is a diagnostic variable, not a feature available for a deployable selection gate.

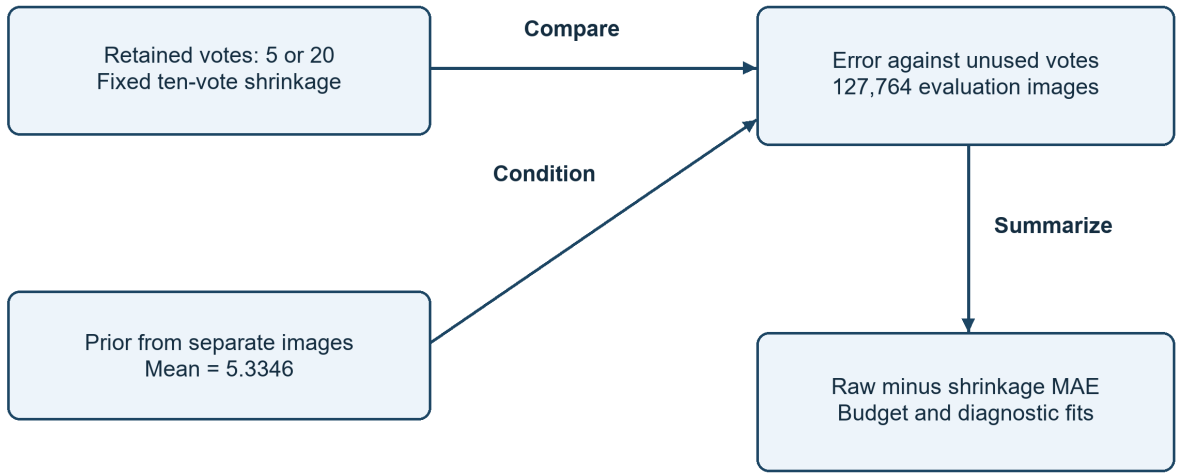

Figure 1. An independently estimated image-average prior combines with nested vote budgets. Common unused-vote targets define paired error gains; budget and retrospective target extremity describe those gains.

5. Models and uncertainty units

The primary regression contains an intercept and an indicator for the twenty-vote budget. Each evaluation image supplies two gain observations. The intercept is the five-vote mean gain; the indicator is the twenty-minus-five difference. The second model adds target extremity and its interaction with budget. These are descriptive regressions on simulated estimators applied to real response counts, not a new experiment assigning human observers to conditions.

Covariance clusters by the first supplied semantic-tag code, yielding sixty-six groups. This allows images within a broad content category and both observations of an image to be dependent. Cameron and Miller, 2015 motivates attention to grouped data, while White, 1980 supplies the general heteroskedastic foundation. The category codes do not identify shared observers or every related image series, so this grouping does not eliminate all potential dependence in the original aesthetic judgments.

The fixed seed is 20261107, and there is one nested draw per evaluation image. Intervals do not integrate over repeated simulated draws, re-estimation of the prior from a new training collection, or a new observer population. Structured evaluation arguments in Roberts et al., 2017 and selection concerns in Varma and Simon, 2006 support preserving those distinctions. A large number of images does not make a single subsampling design equivalent to many independent rating studies.

Absolute-error interpretation follows the outcome-focused reasoning in Hyndman and Koehler, 2006 and Gneiting and Raftery, 2007. Probability calibration in Guo et al., 2017 concerns another target; the shrinkage estimate here is a bounded-scale mean, not a probability that an image is good. Its error reduction cannot establish calibration of a visual model, and its performance should not be transferred automatically to a different loss such as squared error, ranking, or user-specific preference.

Table 1. Mean shrinkage gains and their budget and retrospective-extremity associations. Positive gain favors shrinkage. Semantic-tag-clustered intervals condition on the fixed source histograms and draw.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	0.0854 (0.0037)	[0.0780, 0.0928]	255528
M1	Twenty vs five votes	-0.0864 (0.0026)	[-0.0916, -0.0812]	255528
M1: fit	R-squared = 0.0143	-	-	255528
M2	Unused-mean extremity	-0.5538 (0.0031)	[-0.5599, -0.5476]	255528
M2	Budget x extremity	0.3266 (0.0023)	[0.3220, 0.3311]	255528
M2: fit	R-squared = 0.2853	-	-	255528

Model labels: M1 = Shrinkage gain by retained vote budget; M2 = Retrospective gain versus validation extremity.

6. A substantial five-vote benefit and little twenty-vote gain

At five retained votes, raw-mean absolute error is 0.5057 and shrinkage error is 0.4203. The mean gain is 0.08536, with standard error 0.00371 and interval 0.07795 to 0.09277. Under this experiment, the common prior helps the small-budget estimate on average. This is a comparison against ratings not used in either estimator for that image, rather than a mechanical fit to the same votes from which the raw mean was calculated.

At twenty retained votes, raw error is 0.26486 and shrinkage error is 0.26589. The mean gain is -0.00103. The primary budget coefficient is -0.08639, with interval -0.09161 to -0.08116, showing a clear decline in average gain relative to five votes. We do not infer a separately significant harm at twenty votes from that difference interval; an interval for the twenty-vote mean would need the covariance of the fitted terms.

In the retrospective model, the five-vote extremity slope is -0.5538, with interval -0.5599 to -0.5476. Its budget interaction is 0.3266, with interval 0.3220 to 0.3311. The fitted gain therefore falls as the unused-vote mean lies farther from the prior, with a less negative slope at twenty votes. This is consistent with shrinkage imposing a larger bias on atypical images, but the diagnostic uses the evaluation target and is not an operational rule validated on unseen targets.

The primary R-squared is 0.0143 and the retrospective model's is 0.2853. These fit measures describe variation in estimator gains. They are not image-model quality correlations or the fraction of human aesthetic preference explained by a semantic representation. The source collection contains no generated images or target-model outputs in this analysis. Keeping the unit of the response explicit prevents a label-estimation result from being misreported as a visual-understanding improvement.

7. Reliability and usefulness are different questions

Reliability concepts in Shrout and Fleiss, 1979 and Koo and Li, 2016 concern how measurements vary across appropriate units and raters. Agreement approaches in Bland and Altman, 1986 and Lin, 1989 address still other aspects of comparison. Our experiment does not estimate an intraclass correlation or new inter-rater reliability. It examines whether a selected aggregate estimator better predicts a reserved part of an already observed response histogram. That narrower question is useful even while the deeper validity of aesthetic consensus remains open.

Measurement-error discussion in Carroll et al., 2006 also applies to the unused-vote mean, which is estimated from a finite set of responses. Missing-data principles in Little and Rubin, 2002 caution that images absent from the collection and observers absent from its records may differ systematically. No imputation of missing images or individual rater histories is performed. The complete vote counts define the available evidence, and the source's public accessibility does not establish representativeness of every aesthetic population.

Influence analysis following Cook, 1977 could locate images with extreme gain or loss, but selectively removing them would favor a prior whose main weakness is precisely atypical images. We retain all eligible evaluation images. The guidance in Wasserstein and Lazar, 2016, concerns in Ioannidis, 2005, and multiplicity considerations in Benjamini and Hochberg, 1995 support treating these models as exploratory rather than presenting a tuned pseudo-count as a universally optimal annotation policy. No alternative pseudo-count was selected on the final evaluation errors.

8. What an annotation team could learn

The experiment suggests that the aggregation rule should be evaluated together with the retained vote budget. A prior that materially helps at five votes can be unnecessary once more image-specific evidence is retained. An annotation team could test several policies on a separate calibration set and then lock one before collecting its final evaluation outcomes. Our fixed-ten rule supplies an accessible example of that design issue, not a prescription that ten pseudo-votes is the correct choice elsewhere.

Documentation follows Gebru et al., 2021 by recording the image split, exact histogram use, nested sampling, target construction, prior weight, seed, and semantic clustering. Wilkinson et al., 2016 motivates preservation of the public metadata and processed estimator outputs. The final conclusion remains conditional on those choices: real rating histograms show a budget-dependent benefit from simple shrinkage, and the benefit weakens for image means farther from the prior. General claims about aesthetic truth or image-quality architecture require additional evidence.

9. Decisions the vote experiment leaves open

The common unused-vote target is an intentional design choice. Had five retained votes been evaluated against all remaining votes while twenty retained votes were evaluated against a different remainder, the two estimators would face targets with different compositions. Here the five-vote estimator forgoes fifteen ratings that are available inside the nested twenty-vote draw, but neither uses any vote entering the common target. That makes the budget contrast easier to interpret. It also means the experiment estimates the value of retaining information under this construction, rather than the value of collecting twenty genuinely new judgments after collecting five.

A fixed pseudo-count puts more relative weight on the prior when the retained sample is small. Its reduced influence at twenty votes follows mechanically from the estimator, while the change in actual error depends on the observed rating distributions. Those two facts should remain separate. The negligible average twenty-vote gain could conceal gains for some images and losses for others; it does not establish that the estimators coincide image by image. The retrospective extremity model describes part of that heterogeneity, using information an annotation system would not possess before observing the reserved judgments.

Another unresolved decision is whether the purpose is ranking images, estimating their numerical mean rating, or identifying rare polarized responses. The reported absolute errors address only numerical agreement with a reserved mean. A global prior can reduce that error while compressing distinctions that matter for ranking or concealing disagreement that matters for curation. Neither possibility was evaluated

by the present loss function. A future annotation policy should select its task first, reserve separate data for policy selection, and then evaluate the locked estimator against a target appropriate to that task. The observed five-vote advantage is a reason to conduct that evaluation, rather than evidence that every purpose benefits from the same pooling rule.

References

1. 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.
2. 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.
3. 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.
4. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
5. 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.
6. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
7. 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.
8. Wu, Xiaoshi, et al. "Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis" *arXiv preprint arXiv:2306.09341*, 2023. https://arxiv.org/abs/2306.09341.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
17. Gu, Jinjin; Cai, Haoming; Chen, Haoyu; Ye, Xiaoxing; Ren, Jimmy; Dong, Chao. (2020). PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. https://arxiv.org/abs/2007.12142.
18. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.
19. 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.
20. 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.

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

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

23. 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.

24. 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.

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

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

27. 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.

28. 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.

29. 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.

30. 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.