Resource Allocation and Optimization
Volume 1 · Number 1 · 2026

When a Small Pilot Misallocates Rating Effort: Equal-Budget Estimation from Public Aesthetic Histograms

George Dawson[1] · Derek Carter[2] · Ryan King[3]

[1] United Kingdom · [2] France · [3] Australia

Original publication: October 31, 2026 | Article 705

Abstract

Should images with greater pilot disagreement receive more ratings? In a fixed sample of 4,000 AVA histograms, an eight-vote pilot guided allocation of 48,000 additional retained judgments. Its exact conditional mean squared error was 0.1554, versus 0.1411 for twelve additional judgments per image. An allocation allowed to inspect the remaining vote distributions achieved 0.1379. The small pilot therefore harmed this budget rule despite modest available headroom. These results concern estimating captured finite-record means, rather than recruiting future raters or evaluating an automated quality model.

Exploratory finite-record study. Simulated retention of existing judgments, without new annotators, measured annotation costs, or a completed publication review.

When a Small Pilot Misallocates Rating Effort: Equal-Budget Estimation from Public Aesthetic Histograms

> Exploratory finite-record study. Simulated retention of existing judgments, without new annotators, measured annotation costs, or a completed publication review.

Abstract

Should images with greater pilot disagreement receive more ratings? In a fixed sample of 4,000 AVA histograms, an eight-vote pilot guided allocation of 48,000 additional retained judgments. Its exact conditional mean squared error was 0.1554, versus 0.1411 for twelve additional judgments per image. An allocation allowed to inspect the remaining vote distributions achieved 0.1379. The small pilot therefore harmed this budget rule despite modest available headroom. These results concern estimating captured finite-record means, rather than recruiting future raters or evaluating an automated quality model.

1. Unequal difficulty does not guarantee useful unequal allocation

Annotation budgets are often distributed evenly because equal counts are easy to explain and administer. A natural alternative is to spend more on images whose initial ratings disagree. If some images genuinely have larger rating variance, they can contribute more estimation error at the same sample count. However, a short pilot measures that variance imprecisely. A rule that adapts aggressively to the pilot can move budget toward noisy estimates of difficulty rather than toward the images that would benefit most.

The research question asks whether pilot-estimated disagreement improves allocation across images when the total retained-judgment budget is held fixed. We use the public AVA aesthetic vote histograms Murray et al., 2012 to calculate a finite-record answer. The study is not a new annotation campaign. Existing histograms supply the actual distribution of released vote values; uniform random retention from those finite multisets creates an explicitly simulated budget problem. The target is each image's captured full-record mean, which is known for evaluation.

Meng, 2026 distinguishes semantics and distortions in generated-image quality assessment through a two-stream vision-language framework. That distinction motivates keeping the reference-label construction visible when evaluating an image-assessment method. We do not apply the framework to AVA or treat aesthetic disagreement as a technical-distortion label. The present contribution concerns the reliability of an upstream budget allocation rule. A downstream quality model would inherit the resulting label estimates, but its performance is outside the executed experiment.

The analytical proposition is conditional: allocating according to useful variance information may reduce mean-estimation error, while allocating according to a noisy pilot may fail to realize that advantage. Classical sampling work Neyman, 1934 motivates distinguishing equal allocation from variance-guided allocation. Our numerical rule includes bounds and integer rounding, so it is not presented as an exact implementation of an unconstrained optimal theorem. We compare its actual calculated consequences, including an information-rich reference that would be unavailable during genuine annotation.

2. What disagreement and the target mean represent

AVA records counts in ten ordered rating categories. We treat those bins as the numerical values one through ten and define an image's full-record mean from all its captured votes. This choice makes squared estimation error calculable, while inheriting the equal-spacing convention of the numerical scale. The measurement discussion of Stevens, 1946 helps clarify that convention. It does not establish that an increment from one category to the next has identical psychological meaning for every observer or every image.

Within-image vote dispersion is also different from inter-rater reliability. Intraclass-correlation methodology Shrout and Fleiss, 1979 and reporting guidance Koo and Li, 2016 require specifying the reliability object and data structure. Nominal agreement work Fleiss, 1971 addresses another construction. We possess aggregate vote histograms, without individual rater identities or the cross-image response matrix. The pilot sample variance is consequently a numerical disagreement summary. It is not an estimated reliability coefficient, and the analysis cannot reconstruct whether the same observers repeatedly gave high or low scores.

Agreement between an estimated mean and a captured full mean is itself a separate question. The distinction between agreement and correlation emphasized by Bland and Altman, 1986 is relevant because our objective is not a high rank correlation across images. We minimize average squared error in estimating numerical means. Scoring and estimation principles Gneiting and Raftery, 2007 make that target explicit. A policy that improves the overall average can still change which individual images are estimated well, and a different ranking or decision objective could favor another allocation.

We also distinguish budget allocation from shrinkage. Empirical Bayes estimation Efron and Morris, 1973 suggests using shared information to stabilize uncertain estimates, but our policies do not shrink the estimated image means toward a common prior. They change how many additional votes are retained for each image. A future method could regularize pilot variance estimates before allocating, or combine allocation with mean shrinkage. Those are plausible follow-ups, not procedures silently included in the current calculations.

3. The finite-record sampling design

The complete AVA metadata file contains 255,528 histograms. After sorting by image name, we restrict candidate images to the lexicographic second half and require at least one hundred recorded votes. This creates an eligible evaluation pool of 127,580 images. A generator initialized with seed 20261107 selects 4,000 images uniformly without replacement from that pool. The restriction and seed are recorded. The resulting cohort is a numerical evaluation sample, not a representative sample of all generated images or all photographic preferences.

For each selected image, the script draws eight pilot votes without replacement from its ten-bin count vector using a multivariate hypergeometric draw. The pilot therefore retains 32,000 existing judgments in total. The unused histogram defines the remaining finite pool. The three allocation policies each receive 48,000 additional votes, giving 80,000 retained judgments including the shared pilot. No additional rater is contacted, and no wall-clock annotation effort or monetary expenditure is measured by these counts.

Uniform allocation assigns twelve additional votes to every image. Pilot-guided allocation uses the pilot sample standard deviation, multiplied by the remaining-count-to-full-count ratio, to form allocation weights. Additional counts are constrained between four and thirty-two per image. A bounded continuous allocation is scaled to the required total, floored to integers, and completed by largest fractional

remainders with deterministic ties. Every policy is checked to sum to exactly 48,000 additional votes and to remain within the available histogram counts.

The third policy uses the actual sample standard deviation of each remaining histogram in the same bounded allocation procedure. It is labeled an oracle-guided reference because it sees information that an annotator would not know before obtaining those votes. It is not a deployable competitor and is not claimed to be the exact integer optimum. Its role is to indicate whether the observed collection contains useful allocation headroom when the weighting information is substantially more informative than an eight-vote pilot.

Even the pilot policy assumes that the size of each captured finite vote pool is known. That is natural for retrospective retention from an existing archive, but differs from recruiting an open-ended stream of new judges. The remaining-count correction and available sample bounds refer to that known pool. An actual campaign would need to define its intended population of judgments, stopping rule, and estimator before borrowing this recipe. The experiment therefore supports a comparison of retained-record policies, with an explicit numerical budget, while leaving the cost and statistical design of future annotation unresolved.

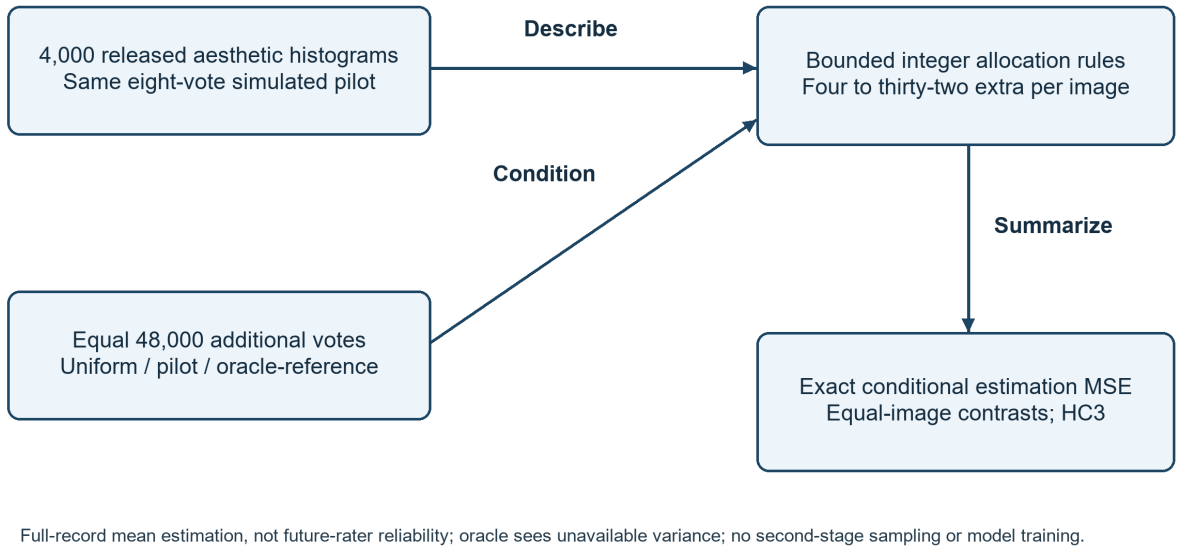


Figure 1. Shared pilot, equal total retention budgets, and conditional mean-estimation error.

Figure 1 separates the information available to the pilot policy from the remaining-histogram information used only by the oracle reference and evaluation. All three policies target the same full-record image means and include the same pilot cost. The figure contains no neural model, generated pixels, or new observer recruitment. Cost-sensitive decision reasoning Elkan, 2001 would require additional assumptions before converting retained-vote counts into a cost-benefit recommendation for an actual annotation service.

4. Why the estimator must respect adaptive retention

When allocation depends on a pilot, simply averaging all retained votes can create a subtle problem. The number of added observations is informative about the pilot outcomes. We instead estimate the full captured mean by combining the known pilot contribution with an independently sampled estimate of the remaining pool's total. The remaining-pool sample mean is multiplied by its known pool size, added to the pilot sum, and divided by the full vote count. Conditional on the pilot, this construction is unbiased for the captured full-record mean.

For a remaining pool of size R, an additional simple random sample of n votes has mean variance equal to its finite-population sample variance multiplied by the difference between one divided by n and one divided by R. The full-record estimator scales that variance by the square of R divided by the original vote count. We compute this exact conditional mean squared error from each recorded remaining histogram. No second-stage vote sample is actually drawn, and there is no Monte Carlo error from estimating the second-stage expectation.

The conditional qualifier matters. The pilot is one realized random draw, and the allocation is then fixed. Our result does not average over repeated pilots. Nor does the finite-histogram expectation characterize the behavior of future human raters, whose responses may differ from the captured multiset. Measurement-error reasoning Carroll et al., 2006 helps distinguish uncertainty in a constructed label from validity of the underlying rating construct. The known finite target is an evaluation convenience, not proof that the captured mean is an unbiased universal aesthetic truth.

Two intercept-only OLS models summarize per-image expected-error differences against uniform allocation. Standard errors use HC3 and normal-reference intervals, within the robust covariance tradition of White, 1980. Each model has 4,000 image rows, no controls, and no fixed effects. These conventional intervals summarize hypothetical image heterogeneity conditional on the selected data and pilot. They are not uncertainty in the exact analytic expectation, repeated-pilot intervals, or observer-cluster inference. The clustering principles discussed by Cameron and Miller, 2015 do not supply unavailable rater identifiers.

5. The pilot-guided rule performs worse

Uniform allocation has mean conditional MSE 0.1410928. The pilot-guided policy has 0.1554018, increasing expected error by 0.0143090, or about 10.1% of the uniform value. Its additional counts range from four to twenty-nine, with median twelve. Thirty-three images receive the minimum four, and two images have zero observed pilot variance. These allocation summaries describe the realized policy; they do not individually identify which low-budget images account for the aggregate increase.

The oracle-guided reference has mean conditional MSE 0.1378812, reducing error by 0.0032116, or about 2.3% relative to uniform. Its additional counts range from seven to twenty-one, also with median twelve. The direction of these two comparisons is informative. Better variance information can help under the declared allocation recipe, but the small pilot fails to capture that benefit in this realization. We do not infer that every adaptive strategy would fail, or that the oracle reference is an attainable annotation policy.

Table 1. Equal-image conditional MSE differences against uniform allocation

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.0143 (0.0010)	[0.0123, 0.0163]	4000
M1: fit	R-squared = 0.0000	-	-	4000
M2	Mean outcome/contrast	-0.0032 (0.0004)	[-0.0040, -0.0025]	4000
M2: fit	R-squared = 0.0000	-	-	4000

Model labels: M1 = Pilot-guided minus uniform exact conditional MSE; M2 = Oracle-guided reference minus uniform exact conditional MSE.

Table 1 reports intercept-only OLS means over 4,000 images, HC3 standard errors, and normal-reference intervals. M1 compares the pilot-guided allocation with uniform; M2 compares the nondeployable oracle-guided reference with uniform. Each policy has the same eight-vote pilot per image and 48,000 additional retained votes. The outcome is an

exact conditional expected squared error difference in numerical rating units squared. Zero R-squared is a property of these mean-only summary models.

The pilot-minus-uniform difference has standard error 0.0010021 and interval from 0.0123449 to 0.0162730. The oracle-reference difference has standard error 0.0003782 and interval from negative 0.0039528 to negative 0.0024703. These are pointwise summaries of the two declared comparisons. Multiple-testing procedures Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 would become relevant to a much larger search across allocation recipes; they do not replace designing an independent evaluation of such a search. No family of unreported pilot sizes or bounds was screened to select the displayed result.

6. Lessons for label construction

The results are compatible with a simple allocation problem: useful heterogeneity exists, but short-pilot noise can be larger than the attainable benefit under this budget. A high pilot sample variance directs additional retention toward an image, while a low estimate can withhold it. Both estimates are based on only eight values. The oracle comparison indicates modest available headroom for the particular bounded weighting rule. It does not isolate the error attributable to pilot noise, integer rounding, or the imposed allocation floor and ceiling.

Modern assessment systems make label construction relevant beyond this numerical example. NIMA Talebi and Milanfar, 2017, MUSIQ Ke et al., 2021, and MANIQA Yang et al., 2022 represent different image-quality modeling approaches, while learned perceptual similarity Zhang et al., 2018, structure-texture similarity Ding et al., 2020, and CLIP-based quality assessment Wang et al., 2022 define other evaluative constructions. No such model is trained here. An improved estimate of an aesthetic mean would not automatically improve all these systems or validate their use on technical distortions and semantic faithfulness.

Preference and quality databases also differ in their targets. Human Preference Score Wu et al., 2023, Pick-a-Pic Kirstain et al., 2023, AGIQA-3K Li et al., 2023, and KonIQ-10k Hosu et al., 2019 organize distinct kinds of evaluative evidence. They are context for designing future allocation studies, rather than extra observations in this experiment. The finite AVA calculation cannot determine how an annotation budget should be distributed across those resources, especially when rating protocols, image populations, and intended downstream decisions differ.

The interpretation of statistical evidence should remain proportionate to that scope. Influence diagnostics Cook, 1977 could identify unusually large per-image differences, but no favorable trimming is performed here. The contextual guidance of Wasserstein and Lazar, 2016 favors reporting the negative pilot result and the estimator construction over calling the policy successful because adaptation sounds efficient. A future study should repeat pilots, consider regularized dispersion estimates, and evaluate its allocation recipe on images not used to choose its tuning constants. Those extensions remain unexecuted.

7. Reproducible budget accounting

The saved image table contains source identifiers, pilot counts in every category, full vote counts, both observed variance summaries, integer allocations, and exact expected errors. Source attribution and reuse conditions accompany the metadata. Dataset documentation Gebru et al., 2021 makes the simulated-retention interpretation and unavailable observer information explicit, while FAIR stewardship Wilkinson et al., 2016 supports retaining the identifiers and calculation record. The account distinguishes a fixed numerical vote budget from actual annotation time or payment.

For this cohort, pilot-guided allocation increased expected mean-estimation error despite equal total retained judgments. The information-rich reference achieved a smaller improvement in the opposite direction. The answer is therefore conditional rather than an endorsement of unequal allocation: a disagreement-based budget rule needs evidence that its pilot estimates are accurate enough to justify redistribution. The current finite-record experiment supplies one such test and preserves its unfavorable result.

References

1. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.
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. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
4. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
5. 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.
6. 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.
7. 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.
8. 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.
9. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
10. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
11. 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.
12. 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.
13. 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.
14. Neyman, Jerzy. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society,97(4),558-606. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.
15. 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.
16. 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.

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

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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

26. 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.

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

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

29. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

30. 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.