Journal of Data Efficiency
Volume 1 · Number 1 · 2026

The Price of Labeling Only Severity Endpoints in Image-Quality Calibration

Vincent Knight[1] · Arthur Lawson[2] · Gavin Bell[3]

[1] Poland · [2] Greece · [3] Croatia

Original publication: October 24, 2026 | Article 621

Abstract

Could an image-quality study label only the mildest and strongest distortion levels? In KADID-10k metadata, endpoint-only linear curves used 3,250 training scores, compared with 8,125 for curves observing all five levels. On middle levels from sixteen held-out reference images, their mean absolute error was 0.484 rather than 0.341. The paired difference was 0.143 score units, with a reference-level interval of 0.131 to 0.155. Allowing quadratic curvature with the full label set reduced error further to 0.299. This small calibration exercise separates label availability from curve flexibility; it does not measure the accuracy or cost of a trained visual-quality network.

Exploratory calibration study based on public score metadata. Draft interpretation and publication decisions require accountable human review.

The Price of Labeling Only Severity Endpoints in Image-Quality Calibration

Exploratory calibration study based on public score metadata. Draft interpretation and publication decisions require accountable human review.

Abstract

Could an image-quality study label only the mildest and strongest distortion levels? In KADID-10k metadata, endpoint-only linear curves used 3,250 training scores, compared with 8,125 for curves observing all five levels. On middle levels from sixteen held-out reference images, their mean absolute error was 0.484 rather than 0.341. The paired difference was 0.143 score units, with a reference-level interval of 0.131 to 0.155. Allowing quadratic curvature with the full label set reduced error further to 0.299. This small calibration exercise separates label availability from curve flexibility; it does not measure the accuracy or cost of a trained visual-quality network.

1. A budget question before a model question

Subjective image-quality labels can be expensive to collect. When a distortion has several ordered severity settings, an appealing shortcut is to rate only its endpoints and interpolate the middle. That shortcut assumes more than monotonic deterioration. It assumes that a straight line learned at the edges gives an adequate approximation where the omitted judgments would have occurred. We examine that assumption with existing scores, using a held-reference design rather than assessing the same images used to fit each curve.

KADID-10k contains 81 reference images, 25 distortion types, and five severity levels per type (Lin et al., 2019). Its complete design permits a transparent counterfactual labeling exercise: retain only levels one and five in the training partition, then evaluate levels two through four on other references. No observers are newly recruited. No image generator is run. The available labels let us measure the predictive consequence of a reduced labeling schedule within this specific released benchmark.

The representational separation advocated in Meng, 2026 motivates careful treatment of distortion information, but our experiment operates at a different layer. It studies how the supervision grid is sampled, not whether a semantic and distortion encoder improves prediction. The source is an artificially distorted natural-image database rather than an AI-generated-image quality experiment. A positive finding about a labeling schedule here would therefore not establish its suitability for generated images or text-image correspondence.

2. Why endpoints need not describe the interior

A severity index is an ordered construction variable, not a guaranteed interval scale of perceptual damage. Equal increments in blur, compression, or other distortion settings may produce unequal score changes. Traditional structural comparison (Wang et al., 2004), natural-scene-statistics approaches (Mittal et al., 2013), and learned perceptual features (Zhang et al., 2018) all offer different representations of visible degradation. Their differences make linearity in an administrative severity label a substantive approximation to check, rather than a property to assume.

Modern predictors such as Talebi and Milanfar, 2017, Ke et al., 2021, and Yang et al., 2022 can represent complex visual information. Greater predictive flexibility does not automatically repair a training set that omits informative regions of the response curve. Language-assisted quality assessment (Wang et al.,

2022) and the joint treatment of structure and texture in Ding et al., 2020 likewise concern useful representations, whereas our budget intervention determines which response labels are supplied. These design choices can interact, but the present experiment separates them deliberately.

Authentic distortions in KonIQ-10k (Hosu et al., 2019) provide an important boundary to this exercise. They generally do not arrive with five controlled levels of one distortion applied to the same reference. Our endpoint schedule is therefore an artificial-grid strategy, not a universal plan for natural-image sampling. An investigator working with authentic mixtures would first need a meaningful ordering variable and a rationale for treating its extremes as coverage of the middle.

3. Two propositions with different comparisons

The primary proposition asks whether endpoint-only linear calibration has higher held-reference mean absolute error than full-level linear calibration. Both models use one separate intercept and slope for each distortion type. Their only intended difference is which training labels are available. This comparison estimates the consequence of the label schedule conditional on the chosen linear form. It does not combine an increase in sample information with an increase in model complexity.

The second proposition compares full-level quadratic calibration with full-level linear calibration. Here the labels are held constant while a distortion-specific squared severity term is added. This contrast examines whether curvature helps describe middle-level scores on unseen references. We keep its interpretation separate from the budget comparison. If both contrasts favor the richer option, the result still does not identify a globally optimal number of labels, a best neural architecture, or a universally suitable polynomial degree.

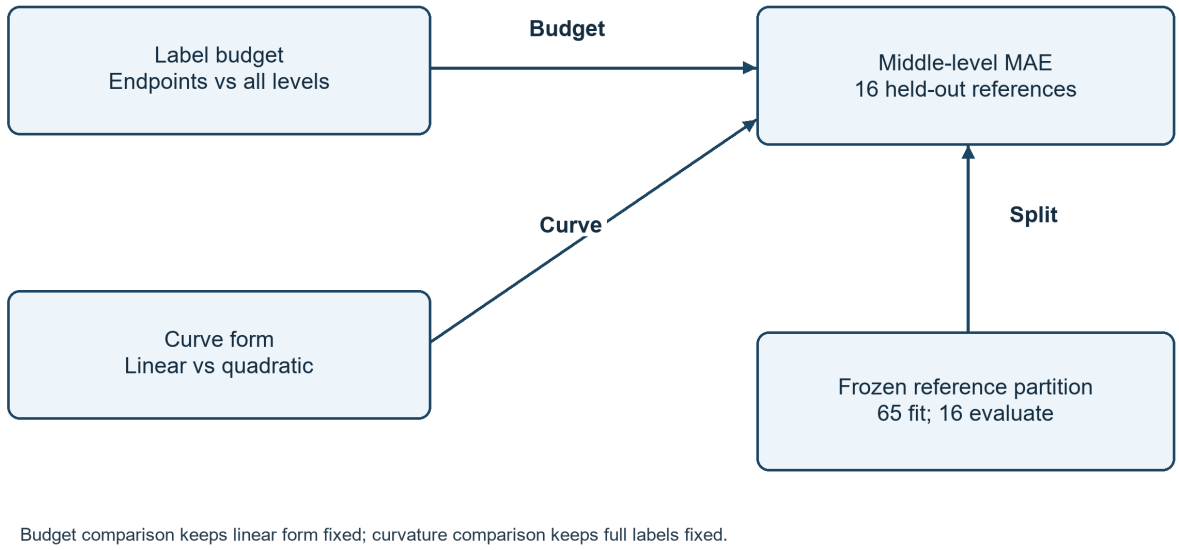


Figure 1. A fixed reference split feeds two comparisons: endpoint versus full-level labels under a linear curve, and quadratic versus linear curves under full labels. Both end at middle-level error on held-out references.

4. Metadata, scale, and partition

The acquired maintainer CSV has 10,125 unique reference-distortion-level cells. We decode the three identifiers from its filenames and verify the full 81 by 25 by five structure. References with identifiers one through sixty-five form training data. Identifiers sixty-six through eighty-one form the holdout. This is a deterministic identifier partition. It prevents direct reference reuse between fitted and evaluated curves, but it is neither a random sample of future images nor a temporal deployment simulation.

The official source describes the DMOS range as one to five, with higher values indicating better visual quality. That direction matters: the acronym alone is insufficient to infer whether a larger value means more damage. The public mirror also calls one field std, although its processing document copies the original variance entry into that field. We do not use it for weighting, divide it by a guessed observer count, or present it as a standard error.

Endpoint training contains 65 times 25 times two, or 3,250 scores. Full-level training contains 8,125 scores. The reduced schedule therefore retains forty percent of the score cells. This fraction measures label availability in the constructed exercise. It is not a measured reduction in financial cost, annotation time, energy use, or computing time. Missing-score reasoning in Little and Rubin, 2002 and measurement cautions in Carroll et al., 2006 support making that distinction explicit.

Evaluation uses only the three interior severity levels of the sixteen held-out references, yielding 1,200 cells. No endpoint score from these references is used in fitting. Each reference contributes seventy-five evaluated cells, so averaging cell absolute errors within reference and then across references produces equal reference weights. We do not select a different holdout partition after inspecting the outcome, and we do not tune a polynomial degree on its errors.

5. Estimation and uncertainty

For each distortion, ordinary least squares fits either a line in severity centered at level three or a line plus its squared centered term. The equivalent pooled specification contains distortion indicators and their interactions with the relevant severity terms. Reference intercepts are omitted because the intended prediction concerns unseen references. Adding their fitted effects and transferring them by position would create information unavailable for the holdout content.

The comparison follows the general warning about model-selection bias in Varma and Simon, 2006 and the structured-validation arguments of Roberts et al., 2017. A label-level random split would allow the same reference content into training and evaluation. Our split removes that direct overlap while retaining a common distortion taxonomy. It still conditions on the benchmark's image collection, and it does not reproduce the uncertainty of collecting a new training set or a new observer panel.

Each primary statistical model is an intercept-only regression on sixteen paired reference-level error differences. Ordinary t intervals use fifteen degrees of freedom. The standard error is the between-reference standard deviation of paired differences divided by the square root of sixteen. These intervals assume the reference differences adequately represent independent units. Bühlmann, 2002 discusses dependence-aware resampling in other settings; we do not add a bootstrap here or claim that such a procedure would repair content-selection limitations.

Absolute error is reported in the supplied score units. Forecast-error discussions in Hyndman and Koehler, 2006 and predictive-comparison work by West, 1996 emphasize that the loss function defines what advantage means. Gneiting and Raftery, 2007 provides a broader framework for aligning evaluation with

the stated target. Our target is a mean subjective score, and our descriptive loss is absolute deviation from its released value. Observer-level probabilistic prediction is outside this analysis.

Table 1. Equal-reference differences in held-out middle-level mean absolute error. Negative curvature change favors the quadratic curve; positive budget penalty favors full-level linear calibration.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	0.1428 (0.0057)	[0.1307, 0.1550]	16
M1: fit	R-squared = 0.0000	-	-	16
M2	Mean contrast	-0.0421 (0.0052)	[-0.0531, -0.0311]	16
M2: fit	R-squared = 0.0000	-	-	16

Model labels: M1 = Endpoint minus full-level linear MAE; M2 = Full-level quadratic minus linear MAE.

6. Results from the retained score grid

Endpoint-only linear calibration has held-out mean absolute error 0.4839. Full-level linear calibration has error 0.3410. Their paired difference is 0.1428, with standard error 0.0057 and interval 0.1307 to 0.1550. The label reduction therefore comes with a clear error penalty across these sixteen references. Because the linear form is held fixed, this difference cannot be attributed simply to fitting a higher-order polynomial in the full-label condition.

Using the full labels with quadratic curves lowers mean absolute error to 0.2990. Its difference from the full-level line is -0.0421, with standard error 0.0052 and interval -0.0531 to -0.0311. Thus curvature explains part of the remaining calibration mismatch. The result does not show that a quadratic curve is the true perceptual law. It shows that this particular additional term improves the selected holdout loss relative to the prespecified line.

Training R-squared values are 0.9341 for endpoint lines, 0.8292 for full-level lines, and 0.8736 for full-level quadratics. The endpoint model's larger training fit is especially instructive. Its training response omits the difficult interior region, so that number does not imply better calibration. The held-out middle-level errors provide the relevant comparison. No prediction fell outside the official one-to-five score range, and none required outcome-dependent clipping.

The mean-difference regressions have R-squared zero, as expected for intercept-only models. This statistic is included for completeness rather than as a measure of calibration quality. Calibration diagnostics for probabilistic neural predictions in Guo et al., 2017 concern another object; here the useful diagnostics are the held-reference loss, the separation of training budgets, and the provenance of the response scale. Reporting these objects together prevents a large training fit from substituting for the actual research question.

7. Limits of a low-cost surrogate

We use classical least squares, not a robustly fitted curve. White, 1980 addresses covariance under heteroskedasticity, while Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 develop different approaches to limiting observation influence. Those alternatives could be valuable if the objective were resistance to unusual rating cells. Introducing them now would change the estimator as well as the label schedule. Our narrow comparison instead preserves a common ordinary fit and reports the resulting prediction error transparently.

Influence diagnostics following Cook, 1977 could locate which held references contribute most to the penalty, but selecting a favorable subset would change the target population. Likewise, one could vary training size, severity locations, or distortion categories. Such extensions should be specified as a new design, with their own heldout strategy. We do not interpret the present sixteen-reference interval as protection against all choices an analyst could have made after seeing the data.

The statistical-significance guidance of Wasserstein and Lazar, 2016 discourages replacing these design judgments with a single probability threshold. Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 describe multiplicity control for broader testing families; the two reported contrasts are exploratory, and their unadjusted intervals are labeled accordingly. Their consistent direction is evidence about this executed comparison, not a confirmatory guarantee that endpoint annotation will fail in every quality-assessment application.

8. Reusable evidence from a small experiment

There is a geometric reason to distinguish the two comparisons. A line fitted only at the endpoints is anchored by extreme response regions. A line fitted to all five levels also minimizes residuals in the middle, where our evaluation occurs. Its advantage may therefore reflect a better compromise for the chosen test region even if the underlying response is not exactly quadratic. Adding curvature addresses a separate approximation problem. These explanations are compatible and should not be compressed into a statement that more parameters alone solved the labeling problem.

The holdout is deliberately restricted to middle levels, so the reported losses should not be advertised as average error over the entire severity range. A deployment population dominated by nearly pristine images could weight the comparison differently. Likewise, a budget policy that retains endpoints plus one interior level might offer another tradeoff. Neither alternative is evaluated here. The processed table makes the retained levels explicit so that later analyses can distinguish a new target distribution from a replication of this one.

The sixteen reference units are also a modest basis for broad generalization. Their confidence interval can be numerically narrow when the paired penalties are similar, yet that similarity says little about image categories absent from the source. More reference images and a genuinely separate collection would provide a stronger test of the annotation policy. The current experiment is most useful as evidence that endpoint interpolation deserves validation, rather than as a final estimate of how much any annotation program should spend.

The archived material retains the source CSV, decoded identifiers, fitted specifications, reference-level losses, and all holdout predictions. Following Gebru et al., 2021 and Wilkinson et al., 2016, the record distinguishes the creator's dataset from the maintainer's metadata mirror and preserves the source-field caveat. Only public score metadata were required. No claim is made to have downloaded or inspected all associated image pixels, reconstructed individual ratings, or trained a visual representation.

The practical finding is specific: for these held references and distortion levels, endpoint supervision is an inaccurate substitute for observing the interior of the score curve. Full-level labels improve prediction even with a fixed linear form, and a separate curvature term adds a smaller further gain. An annotation team could use this result to justify a pilot that preserves some interior levels before committing to an endpoint-only schedule, while recognizing that new imagery and rating procedures require their own evidence.

References

1. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.
2. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
3. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.
4. 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.
5. 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.
6. 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.
7. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.
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. 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.
10. 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.
11. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
12. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
13. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
14. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
15. 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.
16. Lin, Hanhe, et al. "KADID-10k: A Large-scale Artificially Distorted IQA Database" *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019. https://doi.org/10.1109/qomex.2019.8743252.
17. 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.
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. 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.
20. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
21. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

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. 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.

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. Bühlmann, Peter. (2002). Bootstraps for Time Series. Statistical Science. https://doi.org/10.1214/ss/1023798998.

27. 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.

28. 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.

29. 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.

30. 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.