Temporal Systems and Forecasting
Volume 1 · Number 1 · 2026

A Distortion-Condition Score Map Retained Predictive Value Across Later Rating Dates

Cameron Wilson[1] · Luke Reed[2] · Evan Anderson[3]

[1] Ireland · [2] Denmark · [3] Norway

Original publication: November 4, 2026 | Article 618

Abstract

Do known distortion conditions retain predictive value when the rating collection moves into later dates? We froze one hundred twenty-five distortion-type and severity cell means after 170,447 earlier responses, then predicted 133,959 later responses. Equal-day MSE was 0.7567 for the cell map and 1.7549 for the earlier global-mean reference; MAE was 0.7003 and 1.1414. Both paired loss differences had working intervals below zero across eleven later dates. The transfer concerns supplied distortion metadata and source numerical response codes, with changing image and evaluator composition still embedded in the later cohort.

Chronological prediction of source-coded individual ratings. The date boundary uses parsed source calendar labels with unverified timezone; no published processed MOS or neural image-quality model is reconstructed.

A Distortion-Condition Score Map Retained Predictive Value Across Later Rating Dates

> Chronological prediction of source-coded individual ratings. The date boundary uses parsed source calendar labels with unverified timezone; no published processed MOS or neural image-quality model is reconstructed.

Abstract

Do known distortion conditions retain predictive value when the rating collection moves into later dates? We froze one hundred twenty-five distortion-type and severity cell means after 170,447 earlier responses, then predicted 133,959 later responses. Equal-day MSE was 0.7567 for the cell map and 1.7549 for the earlier global-mean reference; MAE was 0.7003 and 1.1414. Both paired loss differences had working intervals below zero across eleven later dates. The transfer concerns supplied distortion metadata and source numerical response codes, with changing image and evaluator composition still embedded in the later cohort.

A temporal question about the rating process

An image-quality dataset can contain a relationship between its known distortion conditions and the numerical scores people assign. Whether that relationship remains useful on later response dates is an empirical question. An aggregate mean difference between periods cannot answer it by itself: period averages may be similar while a condition-specific predictor succeeds, or change while a frozen conditional relationship still helps.

We ask whether a score map fitted on earlier distortion-type and severity cells improves prediction of later individual rating codes relative to the earlier global mean. The source contains calendar labels for responses, allowing a chronological split with no retrospective reassignment based on prediction errors. The analysis concerns a frozen mapping across recorded dates, not a causal effect of calendar time.

Meng, 2026 studies separating semantics from distortions in vision-language alignment for AI-generated image quality assessment. We do not implement its two-stream system or apply it to this dataset. Our narrower contribution concerns the predictive role of known distortion conditions over time. A useful condition map does not demonstrate that semantic content and visual distortion have been disentangled.

The source benchmark Lin et al., 2019 supplies synthetically distorted images with type and severity labels. Ponomarenko et al., 2015 and Gu et al., 2020 illustrate other controlled quality-assessment designs. Such design metadata can be informative, but access to it distinguishes our predictor from a no-reference image model that must infer quality from the image itself.

Why metadata transfer is a limited form of generalization

A distortion-condition mean can capture average score differences associated with a known manipulation and severity category. It cannot describe every image's content, every evaluator's preference, or each individual's response variability. If those omitted factors change in later collection, the frozen map can retain some value without being perfectly calibrated for the destination cohort.

The shift perspective in Ovadia et al., 2019 motivates describing which distribution changes are observed rather than assuming that a date boundary isolates one mechanism. Here response dates, image composition, and evaluator composition can change together. We evaluate the operational later cohort as recorded and do not attribute its errors to a pure temporal drift parameter.

Validation guidance in Roberts et al., 2017 emphasizes aligning evaluation with the intended transport question. The predictor sees earlier responses only, but some image identities appear in both periods. The analysis therefore tests later response prediction with mixed image familiarity. It is not a strict unseen-image, unseen-evaluator, or unseen-distortion holdout.

In-the-wild collections such as Hosu et al., 2019 and Fang et al., 2020 present different relationships between quality, content, and naturally occurring capture conditions. A supplied synthetic distortion label is not necessarily available there. Our favorable temporal result cannot be extended to those datasets without acquiring equivalent destination information and executing a separate assessment.

Response provenance and the calendar split

The analysis uses the locally minimized raw-rating table associated with the KADID source, retaining the same explicit cohort used by an earlier collection-window study in this project. Eligible responses are nongold observations with recognized distorted-image names, numerical codes from one to five, and valid source date and pseudonymous evaluator fields. This yields 304,406 responses over 10,125 images.

The earlier period covers source calendar dates from October twenty-third through October thirty-first, 2018. It contains 170,447 responses to 5,974 images. The later period covers November first through November eleventh and contains 133,959 responses to 5,001 images. Eight hundred fifty image identities occur in both periods. These counts are retained as distinct response and image denominators.

Source timezone has not been independently verified. We therefore refer to the parsed calendar-day labels rather than claim physically standardized UTC days or a precise local midnight intervention. Their ordering supports the declared chronological partition, with that limitation attached to any time-transport interpretation.

Only the distortion type and severity category parsed from the recognized source filename enter prediction. Neither evaluator identity, response duration, trust score, demographic information, nor image pixels are model features. A pseudonymous evaluator field is used only for the common eligibility rule; it is not written into the new daily analysis output or sent to an external service.

The original intake contains participant-related fields and remains in the restricted local source area. Minimization is not claimed as a formal anonymization certification. The present regression table and daily prediction summaries contain no participant keys. This boundary makes the new analysis narrower than a publication-ready release of the underlying raw records.

Fitting a transparent condition map

All twenty-five distortion types and five severity categories have observations in the earlier period, giving one hundred twenty-five cells. We estimate the mean numerical response within every cell. This is equivalent to a saturated distortion-by-severity OLS mapping fitted to the individual earlier responses; no assumption of a common linear severity slope is imposed.

For computational economy, the executed fit uses the earlier cell means weighted by their response counts. Since every observed cell has its own parameter, the prediction is its empirical earlier mean. We do not report a training confidence interval from the saturated aggregate regression or confuse its one hundred twenty-five rows with the number of independent rating dates.

The constant reference predicts the earlier overall mean code, 3.036357, for every later response. Both predictors are frozen before the later losses are calculated. Predicted values are constrained to the source

coding range from one to five; no held prediction required clipping. The comparison therefore does not arise from repairing unbounded outputs after inspecting later errors.

Model-selection concerns in Varma and Simon, 2006 and leakage analysis in Kapoor and Narayanan, 2023 motivate freezing both the period boundary and predictor specification without using the later outcomes. The study is nevertheless secondary and exploratory, with no preregistration claim. Chronological separation protects this prediction calculation while leaving the broader analyst choice of question transparent.

The outcome is a numerical response, not a processed quality label

We predict each individual's coded response. The paper does not reconstruct the published processed DMOS, reproduce observer screening, or equate the raw coding with a calibrated perceptual-quality scale. No claim is made that a larger source code universally means better visual quality. The loss simply measures numerical prediction error under the source coding convention.

Rating measurement in Fleiss, 1971 motivates distinguishing individual responses, their aggregates, and agreement. Murray et al., 2012 supplies a separate example of aesthetic rating distributions, while Li et al., 2023 and Li et al., 2024 address generated-image assessment with their own tasks and score definitions. These are different measurement designs; our response-code predictor cannot be relabeled as an aesthetic or semantic-alignment estimator.

The scale perspective in Stevens, 1946 keeps that distinction explicit. Carroll et al., 2006 also cautions that measured scores contain variation beyond the latent characteristic one might wish to infer. The cell map does not estimate evaluator-independent truth, and its residual is not a direct amount of unmeasured semantic failure.

We evaluate both squared and absolute loss. The choice follows the objective-oriented perspective in Gneiting and Raftery, 2007. Robust-statistical considerations in Rousseeuw, 1984 help explain why extreme residuals can affect these losses differently. Neither loss is declared universally correct; reporting both makes the prediction claim less dependent on one numerical criterion.

Eleven later dates, not one hundred thousand temporal replications

For each later date, we average cell-map-minus-constant squared loss and separately average the absolute-loss difference. Each date therefore provides one paired outcome for each model comparison. The final regressions are intercept-only OLS over eleven dates, using HAC covariance with one lag, finite-sample correction, and a t10 reference distribution.

The temporal-covariance rationale comes from Newey and West, 1987. White, 1980 provides the broader heteroskedasticity-robust context, but the executed table uses HAC rather than independent-response HC3. There are no explanatory controls or fixed effects in these daily mean regressions. R-squared is zero because they estimate average loss contrasts.

Forecast-evaluation guidance in Hyndman and Koehler, 2006 and temporal assessment in Tashman, 2000 motivate keeping fitting and evaluation information ordered. Our horizon is a later collection window rather than a fixed number of hours ahead, and its predictor conditions on supplied distortion labels. Large numbers of responses do not create an equally large number of independent time periods.

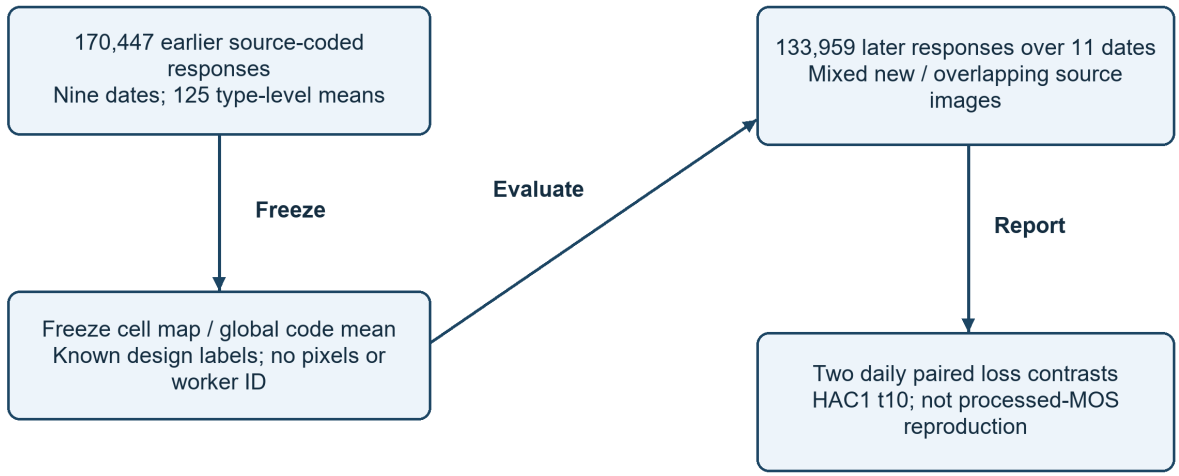


Figure 1. *Earlier numerical responses fit the condition map; later responses assess it. Supplied distortion labels are inputs, while pixels, evaluator identities, and later outcomes do not fit the predictor.*

The condition map retained substantial predictive information

Equal-day later MSE was 0.756670 for the condition map and 1.754907 for the constant reference. Their paired difference was -0.998238, with standard error 0.025168 and working interval -1.054315 to -0.942160. The distortion-condition mapping therefore predicted later numerical responses substantially better under squared loss in the supplied window.

Later MAE was 0.700276 for the map and 1.141377 for the reference. The daily paired difference was -0.441100, with standard error 0.010434 and interval -0.464350 to -0.417851. The favorable comparison did not depend exclusively on squaring larger residuals. Both working intervals remained below zero under the same eleven-date assessment.

Table 1. *Frozen distortion-cell map minus earlier global-mean prediction loss on later responses. N equals eleven source calendar-date means. Intercept-only OLS with HAC lag one, finite-sample correction, and t10 intervals; no controls or fixed effects. MSE is in squared code units, MAE in code units.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.998238 (0.025168)	[-1.054315, -0.942160]	11
M1: fit	R-squared = 0.0000	-	-	11
M2	Mean outcome/contrast	-0.441100 (0.010434)	[-0.464350, -0.417851]	11
M2: fit	R-squared = 0.0000	-	-	11

Model labels: M1 = Later-date type-level minus constant MSE; M2 = Later-date type-level minus constant MAE.

The result does not show that ratings were temporally invariant. The predictor could have systematic bias, composition-specific error, or varying performance across distortion categories while still outperforming a global mean. We do not infer stability of every cell or uniform improvement for every individual response from the aggregate loss differences.

The cell map's information is also unusually explicit. It is supplied the known manipulation type and severity, whereas a practical blind image assessor may receive only pixels. The favorable comparison establishes a useful metadata reference under the observed collection process; it does not constitute a neural IQA benchmark result or a substitute for testing an image-based model.

What image-quality methods would add

Blind quality estimators such as Mittal et al., 2012 and Mittal et al., 2013 infer properties from visual statistics rather than receive the source distortion label directly. Yang et al., 2022 and Ke et al., 2021 use learned visual representations for quality assessment. Their performance cannot be predicted from the cell-map result because their inputs, representations, and fitting objectives differ.

Vision-language quality work in Wang et al., 2022 adds another information source. Structural and texture comparison in Ding et al., 2020 answers yet another measurement question. None of those methods was executed here, and no generated image, embedding, or perceptual distance is used in the chronological prediction. The comparison remains a transparent condition-level baseline.

Agreement arguments in Bland and Altman, 1986 and Lin, 1989 distinguish predictive association from interchangeability of measurements. A metadata prediction and an individual rating can track each other while disagreeing substantially. The remaining error of about 0.70 code units under absolute loss should remain visible alongside the large improvement over the constant reference.

What the saturated map preserves and what it ignores

The condition map does not impose equally spaced physical distortion doses. Severity labels index five source settings within each distortion type, and their perceptual distances can differ. Allowing a separate mean for every type-level cell avoids forcing a universal linear severity slope. It also limits transport to the observed menu: the map supplies no principled extrapolation to a new distortion type or an unobserved setting between the source categories.

The fitting mean weights responses, not images equally. An image with more eligible earlier responses contributes more to its cell's mean. That matches the individual-response prediction target used for fitting, but it differs from predicting an image-equal MOS. The later comparison adds another explicit weighting choice by averaging daily losses equally, so a date with fewer responses has the same influence on the reported temporal mean as a busier date.

This hierarchy explains why the source cohort counts must travel with the result. One cannot infer the equal-day loss difference merely by subtracting two pooled individual-response error averages, and the training response count is not the number of independent temporal observations. The processed daily file retains both response count and loss difference, permitting a reader to assess an alternative weighting objective without changing what the current table estimates.

The predictor has no time variable. It does not learn a trend or update itself as November responses arrive. Its successful transfer means that an earlier condition-specific mapping remains useful in the later collection window. It does not show that a temporal model correctly forecast a change in preferences, that evaluator behavior remained constant, or that the same mapping would survive a much longer interval.

Finally, the known condition labels are available because the source images belong to a controlled distortion experiment. A new AI-generated image may not have a single known distortion type or ordinal severity label. The present metadata result therefore provides a transparent source-task reference rather than a ready-made deployment predictor for generated content. Evaluating that destination would require observable inputs and outcomes appropriate to it.

Conclusion and transport boundaries

The earlier distortion-type and severity map retained predictive value over eleven later source response dates. Its success is conditional on known manipulation metadata, mixed familiar and new images, and the numerical coding of the raw cohort. Evaluator and image composition changes remain part of the destination evidence rather than identified causes. Reusing the same local records does not create an independent replication of the earlier period-contrast analysis; it answers a different frozen-prediction question. The appropriate conclusion is temporal usefulness of this explicit metadata reference, with no claim to establish semantic disentanglement, processed-MOS reproduction, or external image-model accuracy.

References

1. 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.
2. Mittal, A.; Moorthy, A. K.; Bovik, A. C.. (2012). No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2012.2214050.
3. 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.
4. 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.
5. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
6. 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.
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. 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.
9. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
10. 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.
11. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
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. 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.
14. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
15. Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.
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. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.
18. 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.

19. Tashman, Leonard J.. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(00)00065-0.

20. 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.

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

22. 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.

23. Li, Chunyi; Kou, Tengchuan; Gao, Yixuan; Cao, Yuqin; Sun, Wei; Zhang, Zicheng; Zhou, Yingjie; Zhang, Zhichao; Zhang, Weixia; Wu, Haoning; Liu, Xiaohong; Min, Xiongkuo; Zhai, Guangtao. (2024). AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. arXiv:2404.03407. https://arxiv.org/abs/2404.03407.

24. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804.

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

26. 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.

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

28. Ponomarenko, Nikolay; Jin, Lina; Ieremeiev, Oleg; Lukin, Vladimir; Egiazarian, Karen; Astola, Jaakko; Vozel, Benoit; Chehdi, Kacem; Carli, Marco; Battisti, Federica; Jay Kuo, C.-C.. (2015). Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication. https://doi.org/10.1016/j.image.2014.10.009.

29. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

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