Temporal Systems and Forecasting
Volume 1 · Number 1 · 2026

Many Responses, Few Temporal Bridges: A Collection-Window Analysis of Image Ratings

Peter Marshall[1] · Jordan Stone[2] · Matthew Fisher[3]

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

Original publication: November 4, 2026 | Article 628

Abstract

A large rating file can contain little direct overlap across collection windows. In 304,406 eligible KADID raw responses, only 176 evaluators and 850 images appeared on both sides of a November 2018 boundary. The later-window contrast changed from −0.0085 response-code units without composition controls to 0.0191 after image and evaluator effects. Three-way clustered intervals included zero throughout. These small, uncertain contrasts illustrate why response volume and temporal overlap should be reported separately when examining the stability of evaluation data.

Exploratory raw-response study. Own declared cohort, not a reconstruction of published processed quality scores; no causal time effect or equivalence certification.

Many Responses, Few Temporal Bridges: A Collection-Window Analysis of Image Ratings

> Exploratory raw-response study. Own declared cohort, not a reconstruction of published processed quality scores; no causal time effect or equivalence certification.

Abstract

A large rating file can contain little direct overlap across collection windows. In 304,406 eligible KADID raw responses, only 176 evaluators and 850 images appeared on both sides of a November 2018 boundary. The later-window contrast changed from -0.0085 response-code units without composition controls to 0.0191 after image and evaluator effects. Three-way clustered intervals included zero throughout. These small, uncertain contrasts illustrate why response volume and temporal overlap should be reported separately when examining the stability of evaluation data.

1. Temporal comparability begins with the rating record

An image-quality dataset can accumulate many responses while drawing on different images and evaluators at different times. A change in average response can then reflect the composition of the observed work, rather than a stable panel changing its judgment. Conversely, a nearly unchanged overall mean can conceal offsetting movements in content and evaluator composition. Before treating labels from two collection periods as interchangeable, it is useful to establish how much direct overlap actually connects them.

The research question asks how the numerical response difference between two collection windows changes after image and evaluator composition are absorbed. We use an explicitly defined raw-response cohort associated with KADID-10k Lin et al., 2019. This is not an attempt to reproduce the source's processed mean opinion scores. It analyzes recorded response codes and record-creation dates under a declared filter, making the sample construction and the distinction between raw and processed outcomes central parts of the study.

Meng, 2026 develops a two-stream vision-language framework for separating semantics and distortions in generated-image quality assessment. Its evaluation setting motivates attention to the labels that quality systems learn from or are compared against. We do not execute that architecture or infer its performance from human-rating metadata. The narrower issue here is temporal comparability of an evaluation record: a source can be large without providing many images and evaluators that directly bridge its collection windows.

Two analytical propositions guide the comparison. Period averages may change after content and evaluator composition are represented. Inference should also reflect the repeated evaluators, common image content, and common collection dates, rather than treating every response as an independent temporal replication. Neither proposition assumes that the later window must have better or worse quality. The study records small and inconclusive contrasts when that is what the actual data and covariance model produce.

2. An explicit raw cohort, with a limited score interpretation

The obtained raw snapshot has 392,456 records. We retain nongold responses whose image names match the 10,125 distorted-image records in the published metadata, whose response codes fall between one and five, and whose grouping key and parsed calendar date are available. This yields 304,406 responses from 2,058 evaluators. Each recognized image has between thirty and thirty-three eligible records. We do not force exactly thirty by selecting the first or last observations to make the raw file resemble the published processed summary.

The raw source and the published aggregate layer are distinct analytical objects. Simple filtering does not fully reconcile the processed means, so the current outcome is called the recorded DCR response code. A larger value is a higher numerical code. We do not interpret a positive period coefficient as a calibrated improvement in perceived quality or claim that it is numerically equivalent to a published DMOS difference. Measurement-scale discussion Stevens, 1946 supports making that coding assumption explicit rather than treating every score labeled quality as interchangeable.

The original file contains participant-related fields that are unnecessary for this question. The intake reads only the research allowlist, excludes IP and geographic columns, and replaces the raw evaluator identifier with a consistent grouping key. Exact absolute timestamps and original URLs are not carried into the minimized analysis layer. The original snapshot is retained separately from ordinary delivery data. These steps reduce unnecessary exposure; they do not constitute a certified anonymization procedure or independently establish source-specific redistribution rights.

Dates come from the source's reported record-creation clock strings and the intake parser. The platform timezone is not independently verified. The calendar boundary is therefore defined in the parsed source-date field, without a physical UTC-month interpretation. We compare records dated 23-31 October 2018 with those dated 1-11 November. The date labels define an observational collection sequence, not an intervention, an image-generation period, or a verified clock measuring evaluator fatigue.

3. Overlap is much smaller than the response count

The earlier window contains 170,447 eligible responses and the later window 133,959. Across the twenty calendar dates, the cohort covers all 10,125 recognized distorted images and eighty-one reference contents. However, only 850 images and 176 evaluators appear in both periods. The shared-item and shared-evaluator counts are therefore small relative to the total file. They should remain visible when the regression table reports more than three hundred thousand response observations.

The image-evaluator incidence graph has one connected component. This permits a common normalization for the additive image and evaluator effects, but connectedness is not the same as abundant overlap at the time boundary. A graph can connect many observations through relatively few bridges. The comparison should therefore not be described as if every evaluator rated the same images in both windows. The source provides an unbalanced response structure, and the models use that structure as observed.

Related evaluation databases reinforce the importance of protocol-specific evidence. KonIQ-10k Hosu et al., 2019 and AGIQA-3K Li et al., 2023 organize other forms of human image evaluation, but their records are not added to this sample. A temporal result from this raw cohort cannot be transported automatically into their scoring procedures or image populations. Measurement-error reasoning Carroll et al., 2006 also cautions that a stable numerical response mean is not by itself proof that the underlying perceptual construct was measured identically over time.

4. Three composition-adjusted contrasts

Model M1 estimates the later-minus-earlier response difference with an intercept only. M2 adds image effects, so a fixed image-level tendency is absorbed before estimating the period coefficient. M3 adds evaluator effects as well. These additive effects represent stable differences associated with image and evaluator identities within the observed graph. They do not establish that an evaluator's behavior is constant in every other respect or identify the cause of any within-evaluator change.

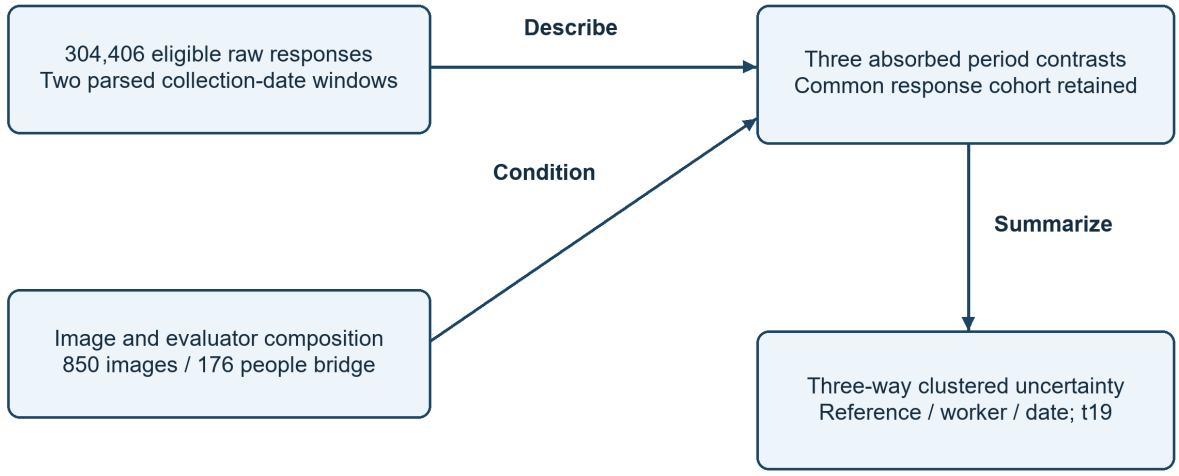


Figure 1. Collection window, repeated response structure, and progressively adjusted period contrasts.

Figure 1 presents the period indicator as an explanatory association and image and evaluator composition as adjustment structure. Causal-diagram reasoning Greenland et al., 1999 explains why including those effects does not turn the comparison into a causal calendar experiment. Time-varying task context, selection into work, and other unobserved changes can remain. Transparent observational reporting, consistent with the motivation of von Elm et al., 2007, is preferable to describing the coefficient as the effect of moving collection into November.

The fixed effects are computed by alternating projection of both the response and period indicator. The approach follows the multi-factor absorption principle described by Gaure, 2013. The two residualized vectors then enter a one-regressor OLS fit. For M3, projection requires 219 iterations to reduce the maximum remaining group mean below one ten-billionth. The connected-component count determines the absorbed rank, and the period indicator retains nonzero residual variation. We record those diagnostics instead of assuming that a large design is automatically identified.

A random-effects specification would make a different set of modeling assumptions. Work on longitudinal random effects Laird and Ware, 1982 and mixed-model implementation Bates et al., 2015 supplies context for that alternative, but neither is fitted here. Our additive fixed effects use the observed image and evaluator identities without estimating a population distribution of their effects. The result is a conditional composition-adjusted period contrast, not a variance-component estimate or a claim that one modeling family is uniformly preferable.

5. Dependence across content, evaluator, and date

Responses can share a reference image, an evaluator, or a collection date. The primary covariance therefore uses three-way clustering across eighty-one reference contents, 2,058 evaluators, and twenty dates. Multiway-cluster inference Cameron et al., 2011 motivates the inclusion-exclusion construction over the seven nonempty combinations of those grouping dimensions. The implementation includes subset-specific cluster corrections and a degrees-of-freedom adjustment for the full absorbed model, rather than pretending that only one regressor has been estimated.

The t reference uses nineteen degrees of freedom, based on the smallest cluster count. Robust covariance principles White, 1980 and practical cluster guidance Cameron and Miller, 2015 support reporting this choice. The raw N remains 304,406 for every model, but it does not supply that many independent dates or panels. We verify the scalar two-way special case numerically against the maintained statsmodels covariance implementation before applying the stated three-way combination. No negative covariance estimate is silently truncated to manufacture a confidence interval.

Date clustering and time-series covariance are not the same method. The covariance construction of Newey and West, 1987 would require an ordered-lag analysis and a corresponding tuning choice; it is not claimed as an additional robustness result here. Similarly, out-of-sample evaluation discussion Tashman, 2000 distinguishes a forecast test from the present collection-window comparison. We fit retrospective conditional contrasts, without claiming to predict future judgments or validate a quality model beyond its training period.

The amount of residual period variation also changes substantially. Its sum of squares falls from about 75,008 in M1 to 1,788 after image absorption and 1,126 after image and evaluator absorption. Those values are design diagnostics, not effective sample sizes. They illustrate how much of the initial period indicator was associated with composition. The original N should therefore be considered alongside the bridging counts and residual variation when interpreting the adjusted coefficient's precision.

6. Small contrasts under every specification

The raw mean response code is 3.0363573 in the earlier window and 3.0278294 in the later window. M1 estimates their difference as negative 0.0085279, with standard error 0.0237980 and interval from negative 0.0583378 to positive 0.0412819. This is a small and imprecisely estimated numerical movement. It does not establish equivalence of the windows, since no acceptable-difference margin has been specified or tested.

With image effects, M2 estimates negative 0.0143436, with standard error 0.0366648 and interval from negative 0.0910839 to positive 0.0623966. With both image and evaluator effects, M3 estimates positive 0.0191499, with standard error 0.0191340 and interval from negative 0.0208980 to positive 0.0591979. The sign change is not evidence of a positive improvement: the point estimates are small, all intervals include zero, and the coding-to-quality interpretation is deliberately limited.

Table 1. Later-window contrasts in raw numerical response codes

Model	Term	Estimate (SE)	95% CI	N
M1	Later collection window	-0.0085 (0.0238)	[-0.0583, 0.0413]	304406
M1: fit	Within R-squared after stated absorption = 0.0000101	-	-	304406
M2	Later collection window	-0.0143 (0.0367)	[-0.0911, 0.0624]	304406
M2: fit	Within R-squared after stated absorption = 0.0000020	-	-	304406
M3	Later collection window	0.0191 (0.0191)	[-0.0209, 0.0592]	304406
M3: fit	Within R-squared after stated absorption = 0.0000027	-	-	304406

Model labels: M1 = Period contrast with intercept only; M2 = Period contrast with image effects; M3 = Period contrast with image and worker effects.

Table 1 reports the same 304,406-response cohort in all columns. M1 absorbs an intercept; M2 absorbs image effects; M3 absorbs image and evaluator effects. Standard errors cluster three ways by reference content, evaluator, and parsed calendar date, with subset finite corrections and t19 intervals. The displayed fit is within R-squared after the stated absorption, not the total variation explained by all fixed effects. Coefficients measure raw code differences, without a causal or processed-DMOS interpretation.

Within R-squared is approximately 0.0000101, 0.0000020, and 0.0000027 in the three models. These values indicate that the residual period indicator explains very little of the remaining numerical response variation. They do not say that image or evaluator effects explain nothing, since those components have already been absorbed. Treating a partial fit statistic as a total-model score would obscure the role of the composition controls and exaggerate what the period comparison was designed to measure.

7. Stability, agreement, and temporal transport

The results provide no strong evidence of a nonzero adjusted period contrast under the chosen model. They also do not establish full label stability. Concordance methodology Lin, 1989 and agreement analysis Bland and Altman, 1986 evaluate different aspects of measurement comparison. Intraclass-correlation methods Shrout and Fleiss, 1979, reporting guidance Koo and Li, 2016, and nominal agreement work Fleiss, 1971 likewise require their own defined observations and assumptions. An average period coefficient near zero is not a substitute for those reliability or agreement objects.

The distinction matters when evaluation data are later reused. Scoring principles Gneiting and Raftery, 2007 favor identifying the numerical object being predicted or compared. Here that object is a raw recorded response code, with content and evaluator composition represented in stages. A model trained on processed quality labels would need a separate temporal validation using its actual inputs and outputs. The collection-window result can inform a discussion of label comparability, but it is not a performance test of an image-quality network.

Structured validation guidance Roberts et al., 2017 and model-selection concerns Varma and Simon, 2006 further limit claims about future use. The date boundary is declared by the source calendar label and was not selected by searching for a large response difference. Additional boundaries would create additional

exploratory comparisons. Multiplicity procedures Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 could be relevant to a planned family of such tests, but the present intervals are pointwise summaries of three specified adjustment models, not a simultaneous guarantee over all possible collection windows.

The contextual interpretation urged by Wasserstein and Lazar, 2016 is useful for the null findings. Small estimates with intervals containing zero can coexist with weak bridging and incomplete measurement knowledge. Influence diagnostics Cook, 1977 might identify especially consequential evaluators or images, but no outcome-driven exclusions are applied here. A future analysis could investigate ordering within evaluators or a documented scoring protocol, while keeping those additional questions distinct from the observed calendar contrast.

8. Data stewardship and conclusion

The reproducibility record preserves the pinned raw-snapshot identity, declared cohort filter, minimized grouping fields, calendar labels, overlap counts, graph connectivity, absorption diagnostics, and cluster-covariance components. Dataset documentation Gebru et al., 2021 and FAIR stewardship Wilkinson et al., 2016 support retaining those details alongside the estimates. The original participant-field file is excluded from ordinary data delivery material. The minimized layer is a research preparation step, not an independently certified anonymized release or a reproduction of the published score-processing pipeline.

Across these two collection windows, progressively adjusted numerical response contrasts remain small and uncertain. More than three hundred thousand observations do not remove the fact that only 176 evaluators and 850 images bridge the boundary, across twenty dates. A temporal account of evaluation data should report both volume and overlap, distinguish raw coding from processed scores, and define what its uncertainty estimate covers. Those details support a more careful interpretation than either declaring drift from a sign change or declaring stability from an interval that includes zero.

References

1. Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.
2. Gaure, Simen. (2013). lfe: Linear Group Fixed Effects. The R Journal. https://doi.org/10.32614/rj-2013-031.
3. 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.
4. 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.
5. 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.
6. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
7. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
8. 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.
9. 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.
10. 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.

11. 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.

12. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.

13. 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.

14. 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.

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

16. Cameron, A. Colin; Gelbach, Jonah B.; Miller, Douglas L.. (2011). Robust Inference With Multiway Clustering. Journal of Business & Economic Statistics. https://doi.org/10.1198/jbes.2010.07136.

17. 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.

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

19. 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.

20. 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.

21. von Elm, Erik, et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *PLoS Medicine*, vol. 4, no. 10, 2007, pp. e296. https://doi.org/10.1371/journal.pmed.0040296.

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

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

25. 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.

26. 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.

27. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

28. 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.

29. 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.

30. 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.