Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

Does a Higher Distortion Label Always Mean a Lower Quality Score? Adjacent-Level Evidence from TID2013

Cole Crawford[1] · Simon Sanders[2] · Ryan King[3]

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

Original publication: November 5, 2026 | Article 574

Abstract

Increasing a distortion-level label does not mechanically decrease a released quality score. Among 2,400 adjacent-level comparisons in TID2013 metadata, 277 have higher mean opinion scores at the higher label and seven are tied. The overall increase frequency is 11.54 percent, with a reference-image-clustered interval of 10.40 to 12.68 percent. The pattern is uneven: local block-wise distortions account for seventy-eight increases, while several other types show zero or one. These comparisons concern published aggregate scores and ordinal labels. Without individual judgments or a verified uncertainty model, they cannot distinguish meaningful perceptual reversals from sampling variation or establish that the level numbering represents equal physical damage increments.

Reanalysis of released image-quality scores. The article does not claim that a score reversal proves a labeling mistake or a perceptual improvement.

Does a Higher Distortion Label Always Mean a Lower Quality Score? Adjacent-Level Evidence from TID2013

Reanalysis of released image-quality scores. The article does not claim that a score reversal proves a labeling mistake or a perceptual improvement.

Abstract

Increasing a distortion-level label does not mechanically decrease a released quality score. Among 2,400 adjacent-level comparisons in TID2013 metadata, 277 have higher mean opinion scores at the higher label and seven are tied. The overall increase frequency is 11.54 percent, with a reference-image-clustered interval of 10.40 to 12.68 percent. The pattern is uneven: local block-wise distortions account for seventy-eight increases, while several other types show zero or one. These comparisons concern published aggregate scores and ordinal labels. Without individual judgments or a verified uncertainty model, they cannot distinguish meaningful perceptual reversals from sampling variation or establish that the level numbering represents equal physical damage increments.

1. An ordering assumption hidden inside a quality benchmark

Image-quality studies often expect a more severely distorted image to receive a lower quality score. That expectation can be useful, but several different objects lie behind it. A file can carry a larger level index, a corruption parameter can change, and observers can assign a different average score. These events need not be numerically equivalent. A benchmark user who treats the level index as a guaranteed ordering of observed quality can therefore impose an assumption that should first be checked.

Our question is how often the released mean opinion score increases between adjacent level labels within the same reference image and distortion type. We also ask whether average adjacent-score changes differ across distortion types after accounting for the position of the level step. The study uses supplied scores, not new human ratings. It examines a benchmark property before introducing any image-quality predictor whose performance might obscure that property.

The representation distinction in Meng, 2026 motivates attention to what quality supervision actually measures. The present resource contains conventional distorted images rather than a generated-image semantic-alignment experiment. We do not test that architecture or infer semantic preservation from the released labels. The connection is methodological: a quality model learns from aggregate responses whose ordering and construction deserve explicit examination.

2. The resource offers a complete labeled grid

The creator documentation for Ponomarenko et al., 2015 describes twenty-five reference images, twenty-four distortion types, and five levels per type, yielding three thousand distorted-image scores. Higher mean opinion score indicates better perceived quality on the stated zero-to-nine scale. We acquire a maintainer metadata snapshot containing the image names and supplied scores. The analyzed sample is the complete reference-by-type-by-level grid, rather than the image pixels or newly computed perceptual metrics.

File names encode the reference number, distortion type, and level. We parse these fields case-insensitively and verify that no grid cell is duplicated. The metadata's predefined train/validation split fields are not used because this study fits no image predictor and holds out no images for model

selection. Their presence in the file does not create multiple independent replications of the same score observations.

The source also contains a field labeled standard deviation. Its full construction and relation to an uncertainty estimate are not established in the captured metadata alone. We therefore do not use it as an observer-level dispersion, a standard error, or an inverse-variance weight. The main comparison is between the released point scores. Uncertainty in the reference-clustered summary does not retrospectively supply a valid significance test for every individual score difference.

Related quality resources such as Hosu et al., 2019, Li et al., 2023, Gu et al., 2020, and Lin et al., 2019 involve other image populations and rating designs. We do not combine their score scales with TID2013 to enlarge the sample. Their relevance is that benchmark construction varies, so an ordering convention should be assessed within its own resource. The present three thousand values remain a single captured release with a declared interpretation of its file labels.

3. Compare adjacent labels within the same content and type

Within each reference and distortion type, we sort the five records by level. We subtract each preceding level's score from the following level's score, creating four comparisons per reference-type combination. This yields 2,400 adjacent differences. A positive difference is called a score inversion for the purpose of this audit: the higher label has a higher released score. A difference of zero is retained separately as a tie.

The term inversion describes an observed ordering relative to a conventional severity expectation. It does not diagnose a corrupted file or an incorrect observer judgment. Some distortion manipulations can interact with content in ways that a simple scalar severity description does not capture. Moreover, an aggregate score contains measurement variation. Our procedure cannot separate those explanations from the supplied point values alone, and the article does not rank them as established causes.

A level step is ordinal within a distortion type. It is not an equal physical increment across Gaussian noise, blur, contrast change, and block-wise corruption. Thus a regression comparing mean score changes across types describes their supplied schedules, not an effect per common dose of damage. This distinction remains necessary even though every type has the same number of images and level labels in the released grid.

Structural metrics such as Wang et al., 2004 and Lin Zhang et al., 2011, feature-based distance in Zhang et al., 2018, and the texture-structure perspective in Ding et al., 2020 address particular definitions of image difference. None is calculated here. A favorable value for such a metric would not itself establish that a larger TID2013 label should have a strictly lower observed mean score. The benchmark ordering audit and a metric-validation study would use different measured quantities.

4. Separate frequency from the magnitude of score movement

The first model estimates the mean of a binary indicator for a positive adjacent-score change. Each of the 2,400 comparisons receives equal weight. Because every reference supplies the same number of types and steps, this is also an equally weighted average over the twenty-five reference images. The summary does not give extra weight to a distortion type because it produces a large numerical score movement.

The second model uses the signed adjacent MOS difference as its outcome. It includes distortion-type indicators and step indicators, with the first distortion type and the first adjacent step serving as references. This additive model allows average score changes to differ by type and by step position. It does not impose monotonicity, fit an image-dependent neural representation, or assert that type and step explain all content-specific variation.

Both models use covariance clustered by reference image, with twenty-five clusters, the recorded small-sample correction, and a t reference with twenty-four degrees of freedom. This accounts for repeated comparisons sharing reference content under the chosen scheme, following the dependence considerations in Cameron and Miller, 2015. Heteroskedasticity correction in White, 1980 would not alone represent those repeated-content observations. The number of grid cells should not be mistaken for 2,400 independently sampled reference scenes.

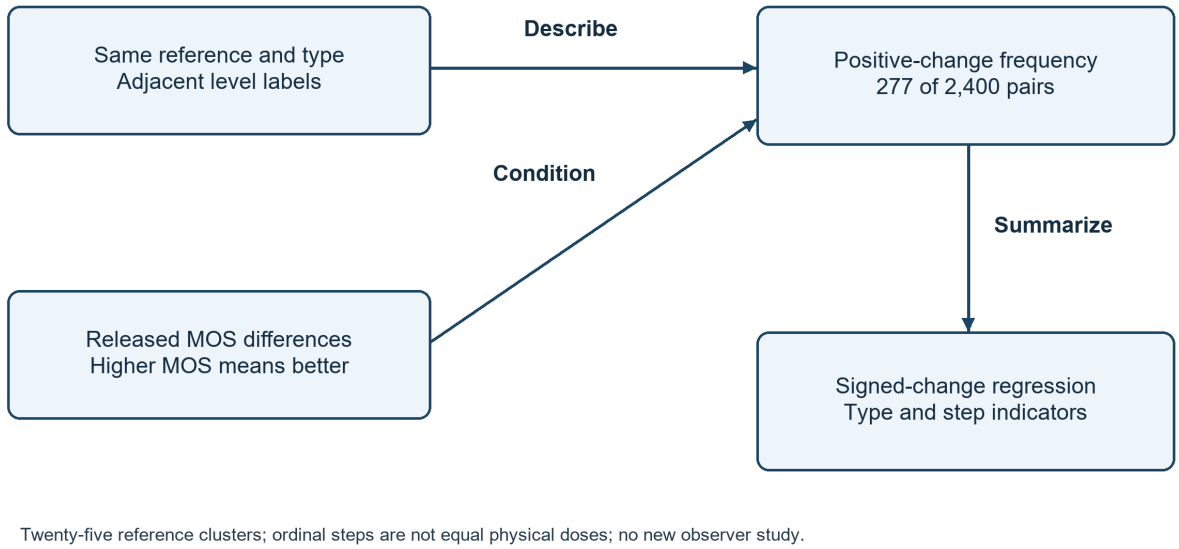

Figure 1. Within-reference and within-type score differences support a frequency summary and a separate type-and-step magnitude model.

Table 1. Released score-inversion frequency and selected distortion-type contrasts in adjacent MOS change.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.1154 (0.0055)	[0.1040, 0.1268]	2400
M1: fit	R-squared = 0.0000	-	-	2400
M2	Gaussian blur vs type 1	-0.3247 (0.0198)	[-0.3656, -0.2839]	2400
M2	Block-wise vs type 1	0.6705 (0.0308)	[0.6070, 0.7340]	2400
M2	Contrast change vs type 1	-0.0682 (0.0162)	[-0.1016, -0.0348]	2400
M2: fit	R-squared = 0.2855	-	-	2400

Model labels: M1 = Mean adjacent-level score-inversion frequency; M2 = Distortion and step differences in MOS change.

The first model has effectively zero explanatory R-squared because it estimates an overall mean. The type-and-step model has R-squared 0.2855. These statistics describe the two stated outcomes. They are not image-quality prediction correlations, agreement with individual observers, or evidence that a semantic representation explains twenty-eight percent of perceived quality. The analysis contains neither image features nor model-generated quality estimates.

5. A minority of adjacent scores increase, with concentration by type

There are 277 positive score changes, seven ties, and 2,116 negative changes. The estimated positive-change frequency is 0.11542, with standard error 0.00551 and interval 0.10405 to 0.12679. Negative changes dominate, so the general direction is consistent with worse quality at higher labels. At the same time, strict observed monotonicity fails often enough that it should not be assumed for every cell when using this captured release.

The overall average adjacent change is -0.60584 MOS units. Its sign answers a magnitude question distinct from the positive-change count. A distortion type can have several small score increases and still a strongly negative mean change if another step produces a large decrease. Consequently, the inversion fraction should not be used as a substitute for the mean decrement, or the mean decrement as proof that every adjacent ordering follows the expected direction.

Each distortion type contributes exactly one hundred adjacent pairs. Type fifteen, local block-wise distortions of different intensity under the creator's naming, contains seventy-eight positive changes and an average adjacent change of 0.19556. Contrast change, type seventeen, contains forty-eight positives while its mean change is -0.54315. Mean shift, type sixteen, has twenty-seven positives. Those differences identify where the observed ordering pattern is concentrated without demonstrating which physical or perceptual mechanism causes it.

Several other types have very few positive changes. Spatially correlated noise and JPEG compression have none in the captured grid. Impulse noise, quantization noise, Gaussian blur, JPEG2000 compression, lossy compression of noisy images, and sparse sampling/reconstruction each have one. These counts do not certify their scores as error-free or their corruption schedules as universally monotonic. They show that a pooled frequency conceals substantial heterogeneity in this particular release.

6. What a learned quality system would inherit

Rating-distribution approaches such as Talebi and Milanfar, 2017, multi-scale assessment in Ke et al., 2021, and attention-based assessment in Yang et al., 2022 offer different prediction strategies. Language-guided quality work in Wang et al., 2022 adds another representational approach. Their evaluation still requires a clearly specified target. If a training process forces every higher level to be worse while the evaluation target uses the supplied MOS, the imposed order and the observed target can disagree before the model has learned anything about image content.

The current audit does not establish that monotonic constraints are undesirable. A constraint might encode a meaningful scientific or engineering prior, reduce noise sensitivity, or improve transfer. Testing those benefits would require fitting models and evaluating them under a declared task. The observed inversions simply mean that strict monotonicity of released point scores cannot be treated as an unexamined empirical fact. A method that imposes it should explain what object it intends to regularize.

Agreement perspectives in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 help distinguish a score-order check from validation against human perception. We do not estimate new inter-observer reliability or show that two quality measures are interchangeable. Measurement-error reasoning in Carroll et al., 2006 is especially relevant to small adjacent differences whose signs may change with rating variation. Without the individual response structure, we cannot label all 277 increases as statistically meaningful perceptual improvements.

7. Shared content and selected schedules limit inference

The reference images are a fixed, small collection. Clustered covariance describes a hypothetical broader content-sampling interpretation; it does not create new scenes or new observer panels. Mixed-model ideas in Laird and Ware, 1982 provide another framework for repeated observations, but no such model is estimated here. Treating the twenty-four distortion schedules as fixed also leaves uncertainty about generalization to a new corruption family outside the reported interval.

Validation concerns in Roberts et al., 2017 and model-selection issues in Varma and Simon, 2006 matter for a later predictive extension. Randomly separating distorted versions of the same reference image can let content information cross a split. This audit does not perform that split or claim held-out prediction. Its complete-grid analysis is appropriate for the descriptive question, while a future learned metric should define a reference-level validation design and reserve data for any monotonicity-policy selection.

Missing-data principles in Little and Rubin, 2002 apply to evidence not contained in the metadata, including individual judgments and the physical parameters producing each level. Those fields are not filled with assumptions. Influence reasoning in Cook, 1977 could identify particularly consequential reference images or steps, but none is removed because it creates an unexpected score increase. The retained complete grid allows a reader to see those patterns rather than replacing them with a cleaner monotonic reconstruction.

8. Reporting the ordering check without overstating it

Guidance in Wasserstein and Lazar, 2016 and concerns in Ioannidis, 2005 favor separating observed counts, statistical summaries, and causal explanations. The empirical count is exact for the captured scores; its broader uncertainty depends on the chosen reference-clustering interpretation. No individual pairwise test is performed. If a later study tests thousands of perceptual orderings, multiplicity issues addressed by Benjamini and Hochberg, 1995 would need attention rather than treating every positive point difference as a confirmed reversal.

A practical benchmark report could state whether its method assumes monotonic physical distortion, monotonic latent quality, or monotonic observed scores. These assumptions are related but distinct. The present data directly inform only the last object and indirectly motivate checking the others with richer source information. A claimed anomaly in one should not automatically be described as evidence that the benchmark creators mislabeled a level or that a learned representation fails to understand an image.

The documentation aims of Gebru et al., 2021 support retaining the filename parser, score direction, complete grid, unused dispersion field, and adjacent-pair construction. Wilkinson et al., 2016 motivates preservation of the captured metadata and type summaries. The study's conclusion is specific: higher labels usually accompany lower released scores, yet eleven and a half percent of adjacent point comparisons go the other way, with a concentrated type pattern. That fact is sufficient to make strict observed monotonicity a checkable modeling choice rather than an assumption hidden inside quality evaluation.

References

1. 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.
2. 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.
3. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

4. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

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

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Lin Zhang, ; Lei Zhang, ; Xuanqin Mou, ; Zhang, D.. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2011.2109730.

14. 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.

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

16. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

17. 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.

18. 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.

19. 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.

20. Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.

21. 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.

22. 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.

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

24. 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.

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

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

27. 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.

28. 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.

29. 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.

30. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.