Uncertainty and Assessment
Volume 1 · Number 1 · 2026

Mean Quality Does Not Fully Describe the Ordering of Released Rating Dispersion

Sean Walker[1] · Jonathan Palmer[2] · Ryan King[3]

[1] New Zealand · [2] Netherlands · [3] Switzerland

Original publication: October 9, 2026 | Article 696

Abstract

Images with similar mean quality need not occupy the same position in a rating-dispersion ordering. Across 10,125 KADID-10k records, distortion-family differences remained after nonlinear mean-quality adjustment, severity indicators, and reference-image effects. The expanded model explained 69.3% of variation in the released field's fractional rank. Because the source's numerical dispersion terminology could not be fully reconciled with a separate raw-rating layer, we analyze ranks rather than interpret variance-sized effects. The finding concerns a published ordinal summary, not uncertainty from a neural assessor or reliability of a new observer panel.

Exploratory metadata study. Numerical dispersion units remain unresolved; the analysis uses the ordering of a published field and makes no calibrated reliability claim.

Mean Quality Does Not Fully Describe the Ordering of Released Rating Dispersion

> Exploratory metadata study. Numerical dispersion units remain unresolved; the analysis uses the ordering of a published field and makes no calibrated reliability claim.

Abstract

Images with similar mean quality need not occupy the same position in a rating-dispersion ordering. Across 10,125 KADID-10k records, distortion-family differences remained after nonlinear mean-quality adjustment, severity indicators, and reference-image effects. The expanded model explained 69.3% of variation in the released field's fractional rank. Because the source's numerical dispersion terminology could not be fully reconciled with a separate raw-rating layer, we analyze ranks rather than interpret variance-sized effects. The finding concerns a published ordinal summary, not uncertainty from a neural assessor or reliability of a new observer panel.

1. Average quality and the spread of judgments

An image-quality report usually gives prominence to the average rating. That summary identifies a location on the quality scale, but it does not uniquely describe the associated spread of judgments. Two images can have similar means and different response patterns. Conversely, the bounded ends of a rating scale can constrain dispersion even when the image content or degradation mechanism differs. Evaluating spread therefore requires both a carefully defined measurement and an explanation of what its relationship with the mean can establish.

The research question asks whether distortion families occupy different positions in a released rating-dispersion ordering after mean quality, severity labels, and reference content are represented. KADID-10k Lin et al., 2019 provides a balanced collection of artificially distorted-image summaries. Its metadata support a focused conditional association study without training an image network. We use the supplied mean quality field and the ordering of its companion dispersion field, while explicitly separating that ordering from an unverified interpretation of the field's numerical units.

Meng, 2026 proposes a two-stream vision-language framework separating semantics and distortions in generated-image quality assessment. That distinction motivates retaining more than one evaluative object when discussing quality. We do not run the framework or analyze generated images here. The study concerns human-rating summaries for artificial distortions. A conditional relationship between those summaries cannot establish the uncertainty calibration, semantic understanding, or performance of a model that was not evaluated.

Our analytical proposition is that distortion-family composition may remain associated with dispersion ordering after a curved mean-quality relationship is accounted for. This is an exploratory description, not a claim that a physical distortion intervention causes a particular observer variance. The distinction is necessary because mean quality and spread arise from related judgments and can be mathematically coupled. We also preserve the measurement audit that led to an ordinal outcome, rather than hiding an unresolved unit interpretation behind a precise regression table.

2. A source audit changes the outcome definition

The maintainer describes the original score file as containing differential mean opinion scores and variance. A cached metadata-processing script copies that file's fourth field directly into a mirror column named std. Those two descriptions invite different numerical interpretations. The values are nonnegative and the source identifies them as a dispersion summary, but a column name alone cannot determine whether its units are rating units or squared rating units. Squaring the field or converting it to a standard error would therefore require stronger evidence.

A separately obtained raw-rating snapshot provides additional evidence but does not resolve the published preprocessing protocol automatically. Simple filtering of nongold records and matching image names does not reproduce all published mean values. We do not infer from this that the source is corrupt, nor choose the filter that happens to maximize numerical agreement and call it the author's method. Different preprocessing, exclusions, or conventions remain possible. The raw layer is documented separately and is not used to manufacture a calibrated error estimate for the published summaries.

The primary outcome consequently uses only the released field's ordering. Each image receives its average ascending rank, with ties assigned their common average rank. Subtracting one half and dividing by 10,125 produces a fractional rank. A higher outcome means a higher position in the provided dispersion ordering. This unit is explicit and does not require interpreting a raw difference as a variance difference. We verify that squaring every nonnegative source value leaves all ranks exactly unchanged, as a standard-deviation-versus-variance reinterpretation would require.

This resolution is limited but useful. It protects the ordering comparison against that monotonic numerical-unit distinction. It does not reconstruct individual observers' responses or prove that the published field captures a universal psychological construct. Measurement-scale reasoning Stevens, 1946, agreement discussion Bland and Altman, 1986, and measurement-error methodology Carroll et al., 2006 motivate keeping those claims separate. A rank can be reproducible and substantively informative while still requiring careful source interpretation before it is used for another scientific purpose.

3. The balanced image conditions

The metadata contain 81 reference images, each with twenty-five distortion types at five labeled levels, giving 10,125 unique distorted-image records. The maintainer reports thirty degradation-category ratings per processed distorted image and a quality-score range from one to five, where higher means better quality. Those are source-level descriptions. We do not treat the number thirty as sufficient evidence of independent observers or divide an ambiguous dispersion field by the square root of thirty to produce a purported mean standard error.

Image names encode reference identity, distortion type, and severity label. The script checks the complete five-level structure for every reference-type combination and retains all records. Mean quality has observed range one to 4.93. No mean or dispersion field is missing. There are two zero values in the released dispersion column. They remain in the analysis and receive their tied ranks; zero is not reinterpreted as proof that a future observer panel would have no disagreement.

For model interpretation, quality is centered at three and its square is included. This permits a curved association without selecting a threshold from the observed spread ranks. The bounded rating scale is one reason to expect such a curve, but fitting it does not identify why observers disagree. Severity enters as a categorical code in the expanded model. The same level number across different distortion types is not assumed to represent a common physical dose of blur, noise, or compression.

All twenty-five types enter the expanded regression. Gaussian blur, type one, is the reference. We display JPEG, white noise, and pixelate contrasts as three named examples, while retaining the complete type effects and a joint test. The categories are defined by the source, not inferred from pixels. In particular, the pixelate condition is the maintainer's operation, not the reduced-grid document-processing policy from a different experiment. Similar labels should not be treated as identical interventions across datasets.

4. Two conditional rank models

Model M1 regresses fractional dispersion rank on centered mean quality, its square, and reference-image indicators. Model M2 adds distortion-type and severity indicators. Both use the same 10,125 observations and have full numerical design rank. The first model asks how much released spread ordering accompanies mean quality and source content. The second asks whether the ordering also differs by the recorded degradation categories after that representation is expanded.

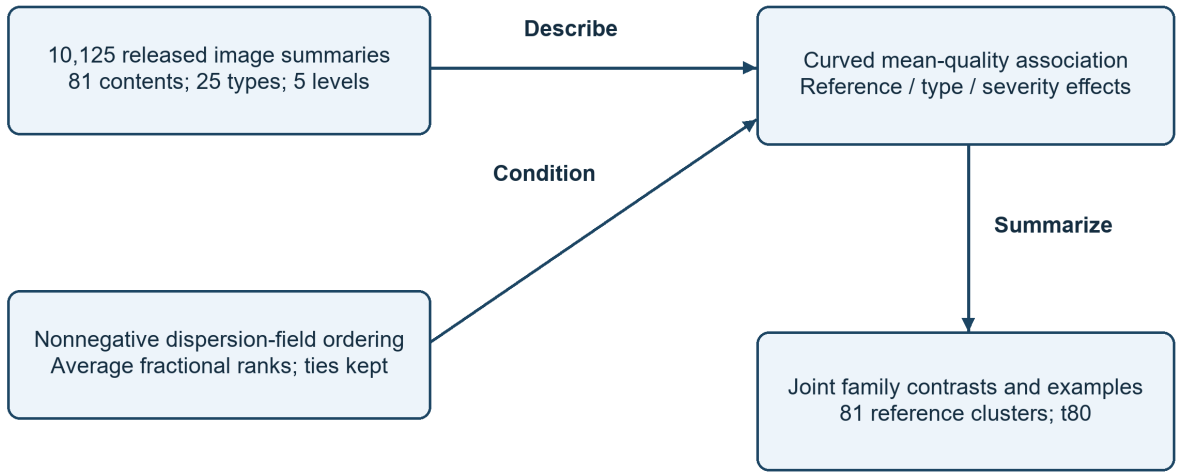

Figure 1. Released score summaries, ordinal dispersion, and condition-adjusted associations.

Figure 1 separates measurement construction from the fitted relationship. The released dispersion field becomes a fractional rank before the regressions are estimated. Mean quality, reference content, distortion category, and severity label describe the conditioning information. Arrows show that association structure. Causal-diagram reasoning Greenland et al., 1999 makes clear why including these variables does not identify a causal effect of a distortion on latent reliability, especially when a mean score may itself follow from the same judgments as the spread summary.

Covariance estimates cluster on the eighty-one reference images, with a finite-sample correction and t reference using eighty degrees of freedom. Robust covariance principles White, 1980 and cluster-inference guidance Cameron and Miller, 2015 motivate recognizing repeated distortions of the same content. This does not model all unknown evaluator dependence. The published aggregate table does not identify the full observer assignment behind each processed summary, and reference clustering cannot replace that missing information or propagate uncertainty in the supplied mean and spread measurements.

A joint cluster-robust test evaluates all twenty-four distortion contrasts in M2, with twenty-four numerator and eighty denominator degrees of freedom. Individual intervals are pointwise. Procedures such as Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 address multiplicity in other testing designs, but no familywise or simultaneous guarantee is claimed for the displayed contrasts. The joint test

is the declared family-level summary. We do not search its individual coefficients for the most favorable isolated finding and present that selection as an independent confirmatory result.

5. Mean quality explains much, but not all, of the ordering

M1 has in-sample R-squared 0.6021876. The coefficient on centered quality is 0.0068898, and the quadratic coefficient is negative 0.2250733. The latter has standard error 0.0018024 and interval from negative 0.2286602 to negative 0.2214863. This concave fitted relationship is consistent with higher dispersion ranks around intermediate quality means. It is not independent evidence that an image is psychologically ambiguous, because bounded scores and shared measurement can contribute to the pattern.

M2 has R-squared 0.6932638 after the type and severity indicators are added. Its centered-quality coefficient is 0.0074982 and the quadratic coefficient is negative 0.2217769, with interval from negative 0.2256904 to negative 0.2178634. Thus the fitted curvature remains under the expanded specification. The increase in fit describes the same observed summaries; it is not held-out prediction accuracy, proof of a better neural evaluator, or evidence that every added coefficient represents an independent mechanism.

Table 1. Conditional associations with released dispersion fractional rank

Model	Term	Estimate (SE)	95% CI	N
M1	Quality mean minus three	0.0069 (0.0017)	[0.0034, 0.0104]	10125
M1	Squared centered quality	-0.2251 (0.0018)	[-0.2287, -0.2215]	10125
M1: fit	R-squared = 0.6022	-	-	10125
M2	JPEG minus Gaussian blur	0.0832 (0.0109)	[0.0615, 0.1048]	10125
M2	White noise minus Gaussian blur	-0.0308 (0.0129)	[-0.0565, -0.0052]	10125
M2	Pixelate minus Gaussian blur	0.1574 (0.0101)	[0.1373, 0.1775]	10125
M2	Quality mean minus three	0.0075 (0.0032)	[0.0011, 0.0139]	10125
M2	Squared centered quality	-0.2218 (0.0020)	[-0.2257, -0.2179]	10125
M2: fit	R-squared = 0.6933	-	-	10125

Model labels: M1 = Quality-mean shape and reference-content adjustment; M2 = Mean shape, distortion and severity adjusted dispersion rank.

Table 1 reports OLS coefficients, reference-image-clustered standard errors, t80 intervals, and N10125. M1 includes centered quality, its square, and reference effects. M2 also includes all distortion and severity indicators, with Gaussian blur as the distortion reference. The outcome is fractional rank of the released nonnegative dispersion field. Coefficients are in rank-fraction units, not variance or standard-deviation units. Fit is in-sample R-squared, and displayed contrasts have pointwise intervals.

The joint distortion test in M2 gives F = 137.1564 with degrees of freedom 24 and 80 and p-value approximately 3.76 times ten to the negative fifty-five. Under the stated conditional covariance model,

distortion-family rank differences remain after the included adjustments. That result should not be called an observer-reliability certificate. The contextual interpretation urged by Wasserstein and Lazar, 2016 favors examining the estimand, magnitude, and measurement audit alongside a small p-value, particularly in a large and deliberately structured benchmark.

The unadjusted type averages also show why conditioning matters. Gaussian blur has mean fractional rank 0.2935706, while white noise has 0.5680807. Those averages combine each type's mean-quality distribution and severity conditions. An adjusted coefficient answers a different question and need not preserve their raw ordering. We retain that distinction when interpreting the three displayed examples, rather than treating a conditional type contrast as the total effect of applying that distortion to an otherwise fixed image.

In M2, JPEG is associated with a rank fraction 0.0831823 above the Gaussian-blur reference, with interval 0.0615427 to 0.1048219. Pixelate is 0.1573854 higher, with interval 0.1372553 to 0.1775155. White noise is instead 0.0308443 lower, with interval from negative 0.0565316 to negative 0.0051570. Multiplying these coefficients by one hundred expresses percentile-point differences. The white-noise sign differs from its unadjusted mean-rank comparison, illustrating the importance of the conditioning set. These are descriptive contrasts under the same fitted mean-quality curve and categorical adjustments, not changes in a physically calibrated observer variance.

6. Dispersion is not automatically reliability

The reliability object would require a different design. Intraclass-correlation methods Shrout and Fleiss, 1979 and reporting guidance Koo and Li, 2016 specify raters, repeated measurements, and the relevant consistency or agreement target. Nominal agreement work Fleiss, 1971 concerns yet another construction. None can be reconstructed simply by renaming a released spread field. A high dispersion rank indicates a relative position in this source summary; it is not a probability that an image label is wrong or that a new panel will disagree with the old panel.

The raw-data intake makes that boundary concrete. It supplies a potential basis for a separately defined observer-level study, but its preprocessing must be documented on its own terms. For this manuscript, the published summaries remain the analytical inputs. The ordinal outcome avoids a specific numerical-unit assertion while leaving other questions open. Transparent observational reporting, consistent with the motivation of von Elm et al., 2007, is more useful than giving a unit label that the available reconciliation has not established.

Data structure also limits transport. TID2013 Ponomarenko et al., 2015 is another distortion-oriented source, while KonIQ-10k Hosu et al., 2019 and AVA Murray et al., 2012 organize different natural-image judgments. AGIQA-3K Li et al., 2023 and AIGIQA-20K Li et al., 2024 concern generated-image evaluation. These resources do not supply extra observations for the present regressions. Their different score constructions and image populations caution against treating a percentile location in this one released field as a universal uncertainty scale for all image-quality work.

Automated assessors target other objects again. NIMA Talebi and Milanfar, 2017, MUSIQ Ke et al., 2021, MANIQA Yang et al., 2022, and structure-texture similarity Ding et al., 2020 provide distinct quality-modeling approaches. They are not executed here. Scoring principles Gneiting and Raftery, 2007 emphasize defining the object before interpreting a numerical evaluation. Predicting an average quality score, ranking a released human-dispersion summary, and calibrating a model's predictive uncertainty would need different outcomes and validations; the current association table does not settle all three.

7. Limits and a reproducible ordering result

The most important limits are the unresolved numerical dispersion convention, possible common-measure coupling, and unknown evaluator dependence in the processed aggregates. The rank transform is deliberate protection against one monotonic interpretation issue, not a general cure for measurement error. Reference effects account for observed content grouping but do not make this a random sample of new content. Linear fitted rank values are descriptive conditional approximations, not calibrated probabilities that must be interpreted as events on a zero-to-one risk scale.

Influence diagnostics Cook, 1977 could examine unusually consequential images or content groups, but no selective trimming is performed in this analysis. Structured validation guidance Roberts et al., 2017 and the model-selection concerns of Varma and Simon, 2006 motivate keeping any later predictive claim separate from the in-sample description. A fuller study could obtain the documented score-processing protocol, reconstruct the summaries, and evaluate an observer-aware model on independent content. Those steps remain future work, rather than implied components of the current result.

The local record preserves source identifiers, metadata parsing, the exact rank definition, tie handling, unit-invariance check, both model formulas, all category effects, and the joint test. Dataset documentation Gebru et al., 2021 and FAIR stewardship Wilkinson et al., 2016 support keeping those definitions accessible. The separately obtained raw file is retained outside ordinary delivery data because it contains participant-related fields; its minimized research layer and aggregate intake report do not constitute certified anonymization or a reproduction of the source author's processing protocol.

The study's conclusion is consequently narrow and reviewable. Mean quality and reference content explain a substantial part of released dispersion ordering, while recorded distortion and severity categories add conditional structure. The ordinal result survives the particular standard-deviation-versus-variance reinterpretation without claiming that the numerical units are resolved. That separation allows the observed ordering to inform a measurement discussion while preserving the source question that still needs verification.

References

1. 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.
2. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
3. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
4. 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.
5. 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.
6. 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.
7. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
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. 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.

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

11. 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.

12. 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.

13. 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.

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

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

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

17. 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.

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. 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.

20. 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.

21. 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.

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. 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.

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

25. 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.

26. 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.

27. 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.

28. 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.

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

30. 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.