Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

The Dataset Menu behind an IQA Average: A Reporting-Pattern Sensitivity Analysis

Charles Porter[1] · Brandon Porter[2] · Ryan King[3]

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

Original publication: November 7, 2026 | Article 706

Abstract

An average benchmark correlation is defined partly by which datasets enter it. We recovered six incomplete reporting patterns from a published no-reference image-quality table and applied them to twelve fully reported metrics. Exchanging the five-dataset menu for one of these observed subsets changed the average score by an absolute 0.0442 on average, with equal weights for metrics and distinct menus. Two menus also changed the leading metric. This exercise exposes a reporting sensitivity; it does not infer the missing scores of partially reported methods or establish that any method selected a favorable menu intentionally.

Analysis of captured published correlations. Missing score patterns are applied to complete known vectors; unreported model performance is not estimated.

The Dataset Menu behind an IQA Average: A Reporting-Pattern Sensitivity Analysis

> Analysis of captured published correlations. Missing score patterns are applied to complete known vectors; unreported model performance is not estimated.

Abstract

An average benchmark correlation is defined partly by which datasets enter it. We recovered six incomplete reporting patterns from a published no-reference image-quality table and applied them to twelve fully reported metrics. Exchanging the five-dataset menu for one of these observed subsets changed the average score by an absolute 0.0442 on average, with equal weights for metrics and distinct menus. Two menus also changed the leading metric. This exercise exposes a reporting sensitivity; it does not infer the missing scores of partially reported methods or establish that any method selected a favorable menu intentionally.

1. A score average has a population of datasets

Benchmark summaries compress many results into a number that readers can compare quickly. The apparent simplicity can conceal a changing denominator. If one metric has results on five datasets and another has results on only two, their arithmetic means refer to different collections of tasks. Even when every available correlation is accurately transcribed, the aggregate comparison can reflect dataset selection as well as predictive performance.

Our research question concerns the size of that menu sensitivity in an actual published table. How much would complete, known image-quality score vectors change if they were summarized over the dataset subsets used by incompletely reported rows? We deliberately ask the question on complete vectors. This design measures the consequences of a changed denominator without inventing the scores that are missing for other methods.

Meng, 2026 distinguishes semantics from distortions in vision-language alignment for generated-image quality assessment. That substantive separation motivates explicit measurement choices when several quality datasets are combined. Our study examines the reporting layer around quality evaluation rather than running the proposed two-stream system. It supplies no estimate of that system's performance, and it does not assume that the public benchmark table evaluates the same construct as generated-image semantic alignment.

The analysis has two propositions. First, a changed dataset menu can alter a metric's average even though none of its individual scores changes. Second, the same menu change can affect methods differently and thus change their ordering. The first proposition follows directly from averaging unequal values. The second depends on the empirical score profiles and is evaluated using the captured complete cohort.

2. Why interchangeable correlation columns need scrutiny

IQA datasets differ in their images and rating conditions. TID2013 Ponomarenko et al., 2015 and KADID-10k Lin et al., 2019 organize deliberately distorted content, while KonIQ-10k Hosu et al., 2019 and SPAQ Fang et al., 2020 address other photographic quality settings. PIPAL Gu et al., 2020 focuses on perceptual restoration assessment. Aesthetic judgments in AVA Murray et al., 2012 concern another evaluation objective. We cite these resources to motivate the dataset-menu question, not to pool their underlying observers or assume that their score distributions are equivalent.

Generated-image assessments introduce further distinctions. AGIQA-3K Li et al., 2023 provides a setting where quality and alignment need separate interpretation. A high correlation on one dataset can therefore be useful without constituting a universal measure of visual quality. Measurement-scale reasoning Stevens, 1946 is pertinent: a shared numerical range does not establish a shared substantive scale. Our equal-dataset average is an explicit descriptive convention, rather than a validated latent quality index.

The IQA-PyTorch repository in Chen and Mo, 2022 reports multiple metric families and dataset columns. Feature similarity in Lin Zhang et al., 2011, perceptual distances in Zhang et al., 2018, and texture-sensitive comparison in Ding et al., 2020 illustrate different assessment objectives. BRISQUE in Mittal et al., 2012 and NIQE in Mittal et al., 2013 add no-reference approaches. Neyman, 1934 distinguishes probability sampling from purposive selection, relevant to our treatment of captured reporting menus as fixed patterns rather than random missing panels. None of these references identifies why a particular source cell was not reported.

Modern learned metrics such as MUSIQ Ke et al., 2021 and MANIQA Yang et al., 2022 also appear in the table. Their model names can include training-source or variant information, so we retain exact published identifiers rather than treating every spelling as an independent scientific contribution. In the primary calculation, however, identifiers define the finite rows being compared. A separate family grouping affects only the working uncertainty summary, not which score vector belongs to which row.

3. Capturing coverage before comparing performance

We use the no-reference correlation table at repository commit d47b7ac72ed25388371d0b637076c125cdb95a3c. The source file contains twenty-five metric rows. Each nonempty dataset cell records three rounded absolute correlations in a fixed order. This study reads only the Spearman component and uses the first five dataset columns: LIVE Challenge, KonIQ-10k, TID2013, FLIVE, and SPAQ. The separate KonIQ-10k-1024 column is excluded because it represents a different resolution condition and is very sparsely populated.

Column coverage is uneven. LIVE Challenge has twenty-five reported rows, KonIQ-10k twenty-one, TID2013 twenty-one, FLIVE twenty-three, and SPAQ sixteen. Twelve rows have a score in all five primary columns. Thirteen are incomplete, and their observed presence-or-absence patterns form six distinct menus. A menu is a set of dataset columns, not an explanation for why those results were or were not reported.

The complete cohort consists of brisque, nrqm, niqe, ilniqe, nima, pi, cnniqa, paq2piq, dbcnn, topiq_nr, maniqa, and musiq-koniq. All twelve are retained irrespective of their average score. Completeness is required to observe the full and restricted averages directly. That requirement defines a selected cohort: the results do not describe the full twenty-five-row field, and the omitted newer or partially evaluated variants may have quite different profiles.

Missing-data theory Little and Rubin, 2002 cautions against assuming that the complete cases represent the incomplete cases. We do not assume missing at random, estimate a missingness model, or fill an empty cell with zero. Benchmark documentation Gebru et al., 2021 is relevant precisely because an empty cell can have several meanings, including unavailable evaluation, incompatible settings, or omitted reporting. The captured CSV alone does not resolve those alternatives.

4. Applying empirical menus to complete score vectors

For each complete metric, the full summary is the arithmetic mean of its five published absolute Spearman correlations. Spearman's association measure Spearman, 1904 supplies an ordering-based relationship between predictions and ratings; our average combines already reported coefficients. We do not calculate a new image-level correlation, average raw ranks across datasets, or transform the coefficients through Fisher's correlation transform.

Each of the six incomplete menus is then applied to each complete vector. The restricted summary averages only the included dataset scores, with equal weight within that menu. Subtracting the full five-dataset mean yields a signed shift. Taking its absolute value measures how far the reported scalar moves, regardless of whether the new menu raises or lowers it. This produces seventy-two metric-menu cells whose values are fully observed functions of the captured table.

The primary aggregation first averages the six shifts within each metric, then treats the twelve resulting metric summaries equally. Distinct menus receive equal weights even when more source rows use one menu than another. Consequently, the result is not the expected effect of drawing a partially reported row at random. It is a balanced sensitivity summary over the observed set of reporting patterns. Readers with another menu distribution would obtain another estimand.

Figure 1 shows that the incomplete rows contribute their coverage patterns, while the complete rows contribute all performance values. This separation is essential. A restricted score for a complete row is observed, but a hypothetical missing score for an incomplete row remains unknown. The diagram describes a controlled reporting transformation rather than a learned predictor of absent evaluations.

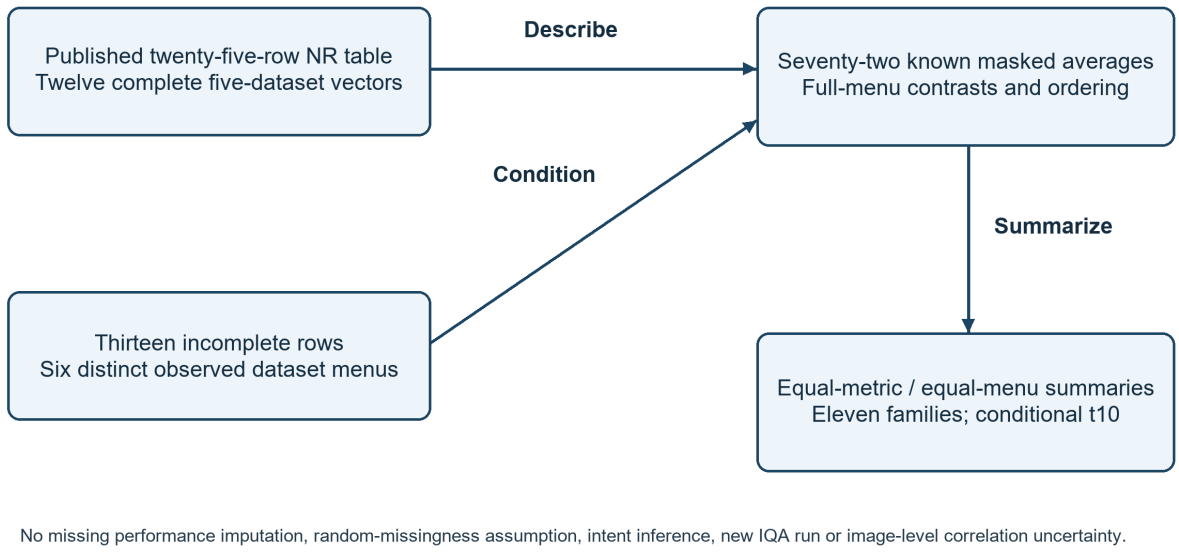

Figure 1. *Six coverage patterns and twelve complete vectors generate seventy-two known contrasts. Metric-level mean absolute and signed changes enter separate descriptive regressions.*

Two intercept-only OLS models report the mean absolute and mean signed changes. Heteroskedasticity-robust reasoning White, 1980 motivates specifying uncertainty explicitly. We cluster by identifier family, grouping niqe and ilniqe together and retaining the other complete rows as separate families. There are eleven clusters, with t intervals using ten degrees of freedom. Cluster guidance Cameron and Miller, 2015 does not turn these families into a random sample of all possible metrics; common training ancestry and shared evaluation datasets remain additional dependence sources.

5. Score displacement and ordering changes

The mean absolute menu-induced change is 0.044220, with standard error 0.004416 and 95% interval [0.034381, 0.054060]. The signed mean is -0.011749, with standard error 0.002532 and interval [-0.017391, -0.006107]. These intervals concern the selected family-level descriptive summaries under the working assumptions. They are not uncertainty bands for the original image correlations, and they do not estimate a population frequency of incomplete reporting.

Table 1. Descriptive changes in equally weighted correlation summaries.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.0442 (0.0044)	[0.0344, 0.0541]	12
M1: fit	R-squared = 0.0000	-	-	12
M2	Mean outcome/contrast	-0.0117 (0.0025)	[-0.0174, -0.0061]	12
M2: fit	R-squared = 0.0000	-	-	12

Model labels: M1 = Average absolute menu-induced mean-score change; M2 = Average signed menu-induced mean-score change.

Each model contains twelve metric-average observations and an intercept only. Both have effectively zero R-squared because their fitted value is the sample mean. Uncertainty is clustered over eleven declared identifier families. The outcome of Model 1 is an average absolute change, while Model 2 uses signed changes; their coefficients therefore answer different questions. A small signed average could coexist with a substantial absolute sensitivity if upward and downward movements cancel.

The two-dataset LIVE Challenge-TID2013 menu produces the largest mean absolute shift, 0.069229. Its signed mean is -0.068326, although not every complete row moves downward: its maximum positive shift is about 0.00542. At the opposite directional extreme, the four-dataset menu omitting TID2013 raises the mean by 0.041822. Exchanging menus therefore need not create a constant adjustment that could be subtracted from every model's result.

Ordering changes confirm the second proposition. There are sixty-six unordered metric pairs in each twelve-row comparison. The six menus generate between one and nine strict pair reversals relative to the full-menu ordering. The LIVE Challenge-TID2013 menu has nine reversals. The menu omitting TID2013 has only one. These are exact counts for the rounded source values, not significance tests of whether two metrics differ in the original observer population.

The complete five-dataset mean selects topiq_nr as its leader. Two observed menus instead select musiq-koniq: the two-column LIVE Challenge-TID2013 menu and the four-column menu omitting KonIQ-10k. The other four preserve topiq_nr. This result does not identify the best metric in an application. It shows that a leading name attached to an unqualified arithmetic average can depend on the average's dataset denominator.

6. Reporting implications without imputing intent

Measurement-error analysis Carroll et al., 2006 helps distinguish uncertainty in an original score from uncertainty introduced by reporting transformations. Here the input coefficients are treated as fixed captured values. Their sampling errors, differing image counts, and shared observers are not available in the aggregate file. A confidence interval for the menu summary cannot stand in for paired image-level tests or establish that one reported correlation is significantly superior.

Selection and transport arguments Stuart, 2010 and Ovadia et al., 2019 provide a second boundary. Applying a coverage pattern to a complete model illustrates what that pattern can do to a known vector. It does not show what the same pattern does to the incomplete model from which it was obtained. Those rows might differ in training data, supported inputs, or intended use. The design avoids filling their cells precisely because such extrapolation would need unobserved evidence.

Benchmark comparison guidance Demšar, 2006 and structured validation principles Roberts et al., 2017 suggest choosing the unit of comparison deliberately. In the present study, the primary mean weights metric identifiers and distinct menus. It does not weight datasets by image count or methods by deployment prevalence. A reader should therefore see the coverage matrix, included dataset list, and aggregation rule next to any aggregate leaderboard value.

Technical-debt arguments Sculley et al., 2015 motivate documenting evaluation choices that might otherwise become difficult to reconstruct. The table does not reveal deliberate selection, so no intent is inferred. Different menus can arise for legitimate reasons. The finding supports disclosure of those reasons and denominator-specific comparisons, rather than accusations based on missing cells alone.

The ASA statement Wasserstein and Lazar, 2016 further cautions against reducing this result to whether an interval excludes zero. The finite shifts and reversals are directly calculable for the captured table. Their substantive importance depends on the decision the average supports, the desired datasets, and the cost of an ordering change. None of those deployment utilities is estimated in this study.

7. What should accompany the scalar summary

A minimally interpretable aggregate would name its included datasets, score orientation, treatment of absent cells, and weighting rule. It would also preserve the individual results so that readers can reconstruct the scalar. A statement such as an average correlation without its dataset denominator leaves open whether an improvement reflects stronger results on common tasks or the replacement of difficult tasks by different ones. Our controlled masks demonstrate the latter mechanism without changing any complete row's scores.

The source's absolute-correlation convention creates another boundary. A coefficient magnitude summarizes association strength after the sign has been removed. This analysis follows that published convention, but it cannot recover the original direction from an absolute value. It therefore does not certify that every metric's raw output has the correct orientation for quality or that all methods can be compared without preprocessing. A signed, image-level evaluation would be needed for those claims.

The six observed menus also should not become a universal stress-test distribution. They are the patterns present in one captured table. Their frequencies, available methods, and intended datasets may change in a later edition. Giving them equal weight makes the present calculation transparent and reproducible, but does not imply that the menus are equally likely in practice. Repeating the exercise with a new table should begin by recording its coverage anew, rather than carrying these six patterns forward as a fixed property of IQA reporting.

8. Conclusion

Observed reporting menus can move IQA averages appreciably even when every underlying value is held fixed. In the twelve-row complete cohort, the balanced mean absolute change is 0.0442, and two of six menus change the leading metric. An aggregate comparison should therefore report its dataset membership and coverage alongside its value. This descriptive audit leaves missing performance unknown and preserves the distinction between a reporting sensitivity and evidence of model superiority.

References

1. 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.
2. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.
3. Demšar, Janez. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. https://www.jmlr.org/papers/v7/demsar06a.html.
4. 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.
5. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
6. 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.
7. 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.
8. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
9. 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.
10. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
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. 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.
13. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
14. 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.
15. 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.
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. 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. 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.
19. Neyman, Jerzy. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society,97(4),558-606. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.
20. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
21. Chen, Chaofeng; Mo, Jiadi. (2022). IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. Software repository. https://github.com/chaofengc/IQA-PyTorch/tree/d47b7ac72ed25388371d0b637076c125cdb95a3c.
22. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
23. Spearman, C.. (1904). The Proof and Measurement of Association between Two Things. The American Journal of Psychology. https://doi.org/10.2307/1412159.
24. 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.

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

26. 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.

27. 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.

28. Hosu, Vlad, et al. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.

29. 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.

30. 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.