

Evaluation and Comparison of Statistical Models for Estimating Large-Scale MOS from Small Subjective Subsets

Javidi, K. and Martini, M.G.

This is a copy of the author's accepted version of a paper subsequently published in the proceedings of the Quality of Multimedia Experience, Cardiff, UK, 29 Jun - 03 Jul 2026.

The final published version will be available online at:

<https://doi.org/10.1109/QoMEX69967.2026.11618375>

© 2026 IEEE . This manuscript version is made available under the CC-BY 4.0 license

<https://creativecommons.org/licenses/by/4.0/>

For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

Evaluation and Comparison of Statistical Models for Estimating Large-Scale MOS from Small Subjective Subsets

Kamran Javidi[†] Maria G. Martini^{*}

^{*}*Faculty of Engineering, Computing and the Environment, Kingston University, London, UK*
{k.javidi, m.martini}@kingston.ac.uk

[†]*School of Computer Science and Engineering, University of Westminster, London, UK*
k.javidi@westminster.ac.uk

Abstract—The rapid growth of multimedia products and services creates strong competition in terms of content quality and user acceptance. While objective quality assessment methods provide useful technical insights, subjective evaluation remains essential for capturing true human perception. Most subjective quality studies are conducted in controlled laboratory environments with a number of viewers adequate to enable statistically significant results. This presents two main limitations: it is difficult and time expensive to involve a large number of participants in subjective tests; not all subjects have the same accuracy in assessing the quality of the content. To address these limitations, an alternative approach is to collect ratings from a smaller group of subjects and apply statistical models to estimate the mean opinion score (MOS) as provided by a large group of viewers. In this paper, several statistical models for MOS estimation are evaluated using small subsets of a subjective dataset. Specifically, EPMOS, based on ordinal logistic regression, is examined alongside the standardized ITU-T P.910, ITU-T P.913, and ITU-R BT.500 models. The estimated results from each model are compared with the reference MOS computed from the full dataset. The results demonstrate that EPMOS provides estimation errors and correlations with the reference scores that closely match the best performance across models, even if it does not always achieve the absolute optimum. The complete set of results is made publicly available upon publication of this paper.

Index Terms—Subjective quality assessment, MOS, estimation population MOS, ITU-T P.910, ITU-T P.913, ITU-R BT.500.

I. INTRODUCTION

The increasing use of multimedia services requires reliable methods to assess quality. Unlike objective measures, subjective evaluation reflects human perception and is essential for understanding how users experience content. Subjective testing, however, faces several challenges, such as controlling the environment, following standardized protocols, recruiting enough participants, and selecting appropriate post-processing methods. As a result, most tests rely on a limited number of subjects. While around 24 participants are typically recommended [1], researchers continue to explore ways to estimate subjective outcomes for larger populations. To address this, Pérez et al. [1] proposed using a small group of four to six subjects repeating tests multiple times, then validating the results against objective assessments. Statistical models have become

the most common solution for predicting perceived quality in larger groups. For instance, Nawala et al. [2] introduced the Generalized Score Distribution (GSD), a two-parameter model capturing the distribution of subjective responses.

The generalized linear model (GLM) family, particularly ordinal logistic regression, was also applied to estimate mean opinion score (MOS). Pezzulli et al. [3] proposed estimated population mean opinion score (EPMOS), which is well-suited for categorical data and large subject populations. Other models include those recommended in ITU-T P.910 [4], which uses maximum likelihood estimation to model subject bias and inconsistency, ITU-T P.913 [5], which removes subject bias to adjust MOS, and ITU-R BT.500 [6], which uses subject rejection to improve robustness. Probabilistic approaches were further developed to handle noisy ratings and detect peculiar subject behaviors with strong theoretical foundations [7]–[9].

The standard MOS was calculated by averaging the scores assigned by all participants to a given stimulus. Even without further statistical adjustments to generalize the results to a larger population, this average was widely accepted as a reliable measure of subjective quality [10]–[15]. However, statistical methods were not limited to multimedia applications. For instance, approaches based on GLM models were effectively applied in other subjective assessment contexts. A Bayesian ordinal logistic regression was employed by Sirengo et al. [3] to identify risk factors affecting prostate cancer screening outcomes across three categories: no cancer, early detection, and advanced stage. Similarly, a Bayesian ordinal logistic model was applied by Mwalili et al. [16] to correct inter-observer measurement errors in ordinal responses, and calibration data were incorporated to adjust for examiner-specific scoring biases.

In this paper, we use a large dataset [17] containing per-subject scores for six videos, each with nine degraded versions (stimuli used the subjective tests). The scores for each video are split into four subsets of 15, 20, 30, and 40 subjects. Various statistical models are applied to estimate the MOS for each subset and compared with the actual MOS computed from all available raw scores for each stimulus. The results showed that the ordinal logistic regression model proposed

by Pezzulli et al. [18] achieved performance that was often comparable to or better than the other selected models in estimating the population MOS, with only minor differences observed in some cases. Furthermore, even with a smaller number of subjects (e.g., 15), the EPMOS model continued to deliver performance that was comparable to, and in many cases better than, the alternative models.

The main novel contribution of this paper consists in the comparative performance evaluation of different statistical models for estimating the MOS obtained with a substantially larger number of participants. The dataset was carefully trimmed and formatted to ensure consistency and reliability. For the EPMOS model in particular, the effective coefficients were proposed by identifying the strongest contributing effects, including the independent effects of the videos, the display resolutions, and the logarithmic and square-root effects of quantization parameter (QP). This study investigates the reliability and accuracy of each model in estimating the MOS when applied to smaller subsets of the dataset (15, 20, 25, and 30 participants), and compares these results with the standard MOS derived from a larger and more representative portion of the dataset. The findings provide valuable insights for the design and optimization of future subjective quality assessment studies.

The remainder of this paper is organized as follows. Section II describes the dataset details and the statistical models employed. Section III presents the analysis of the performance of these models and their comparison. Finally, Section IV provides concluding remarks and outlines potential directions for future work.

II. METHODOLOGY

This section is divided into four parts: first, a review of the dataset used for the experiments (Section II-A); second, an explanation of the EPMOS model (Section II-B); third, a description of the P.910, P.913, and BT.500 models (Section II-C); and finally, the comparison criteria (Section II-D).

A. Dataset details

The dataset used for evaluating MOS estimation was the Netflix Screen Size Crowdsourcing Open Dataset [17], which includes 31 videos. For this study, six videos (T2–T7) were selected because they had the largest number of per-subject scores, ensuring more reliable analysis. Each video is available in three display resolutions— 1280×720 , 1920×1080 , and 517×291 —with corresponding encoded resolutions of 1280×720 , 1920×1080 , and 320×180 [19]. Details of the QP values for each video and resolution are provided in Table I. In total, 54 stimuli were used, with the number of raw subject scores per stimulus ranging from 96 to 332. The reference dataset was then divided into four subsets containing the same stimuli, but with the number of scores limited to 15, 20, 30, and 40. All estimation methods considered in this study were applied to each subset independently to evaluate how well the models can estimate MOS from a limited number of scores while approximating the results obtained from larger populations.

TABLE I: Netflix dataset [17] QP levels by resolution

Display Resolution	QP	T2	T3	T4	T5	T6	T7
A: 1280×720	QP1	23	23	23	23	23	23
	QP2	27	27	27	27	27	27
	QP3	31	31	31	31	31	31
B: 1920×1080	QP1	21	21	21	21	21	21
	QP2	25	25	25	25	25	25
	QP3	29	29	29	29	29	29
C: 517×291	QP1	27	27	27	27	27	27
	QP2	31	31	31	31	31	31
	QP3	34	34	34	34	34	34

All videos use identical QP values per resolution.

The original dataset contains scores in the range of 0 to 100. For this study, the scores were categorized into five categories: 0–19 (category 1), 20–39 (category 2), 40–59 (category 3), 60–79 (category 4), and 80–100 (category 5). This transformation was motivated by the characteristics of the EPMOS model, which is based on categorical regression.

During model fitting, cutpoints (thresholds, denoted as α) are estimated between adjacent categories. Using the full 0–100 scale would require a large number of cutpoints, resulting in an unnecessarily complex model. Therefore, the scores were mapped to a five-level numerical scale for subsequent analysis.

B. EPMOS model

The EPMOS model is a GLM model, specifically developed based on ordinal logistic regression, which is useful for categorical regression purposes [18]. Due to its GLM-based structure, the linear predictor is defined in Equation (1). The model is trained on each subset to estimate the coefficients of this predictor, which capture the effects of different input features. Specifically for this project, the coefficients represent the independent effects of each video, the combined effects of video and display resolution, and the logarithmic and square-root effects of the QP on the scores. The linear predictor of the model is defined as

$$\mu = \eta_l + \eta_m + \eta_{lm} + \gamma\sqrt{QP} + (\delta + \delta_l + \delta_m)\ln(QP). \quad (1)$$

Here, η_l represents the effect of video l , η_m denotes the effect of resolution m , and η_{lm} captures the interaction between the two. In this study, $l = 1, 2, \dots, 6$ and $m = 1, 2, 3$ (corresponding to A, B, and C in Table I). Moreover, γ represents the coefficient of the independent square-root effect of QP, while δ captures the independent logarithmic effect. The coefficient δ_l reflects the combined effect of the logarithmic QP and the video, whereas δ_m represents the combined effect of the logarithmic QP and the display resolution. The matrix representation of the linear predictor can be written as $\mu = \beta^\top \mathbf{x}$, where β denotes the vector of model coefficients (including intercept terms), and $\mathbf{x}$ is the input feature vector. The value of μ is then used as the location parameter of the ordinal regression model. Let

$$a_{lm} = \eta_l + \eta_m + \eta_{lm}, \quad b = \gamma, \quad c_{lm} = \delta + \delta_l + \delta_m.$$

Then, the linear predictor can be expressed in matrix form as

$$\begin{aligned}\mu &= \beta^\top \mathbf{x}, \\ \beta^\top &= [a_{lm} \quad b \quad c_{lm}], \quad \mathbf{x} = \begin{bmatrix} 1 \\ \sqrt{QP} \\ \ln(QP) \end{bmatrix}, \\ \mu &= a_{lm} + b\sqrt{QP} + c_{lm} \ln(QP),\end{aligned}\quad (2)$$

where QP denotes the quantization parameter of the encoded video. For the ordinal logistic regression model, the cumulative probability of observing category j or lower is given by

$$F_j = P(Y \leq j) = G(\alpha_j - \mu), \quad j = 1, 2, \dots, K-1, \quad (3)$$

where K is the total number of categories, α_j are the cutpoints (thresholds), and μ is the linear predictor defined in (2). The logistic function $G(z)$ is defined as

$$G(z) = \frac{1}{1 + \exp(-z)}. \quad (4)$$

The EPMOS can be calculated from the cumulative probabilities F_j as

$$\begin{aligned}\text{EPMOS} &= 1 \cdot F_1 + 2 \cdot (F_2 - F_1) + 3 \cdot (F_3 - F_2) \\ &\quad + 4 \cdot (F_4 - F_3) + 5 \cdot (1 - F_4) \\ &= 5 - (F_1 + F_2 + F_3 + F_4) \\ &= 5 - S,\end{aligned}\quad (5)$$

where $S = F_1 + F_2 + F_3 + F_4$ represents the sum of the cumulative probabilities up to the second-to-last category. This expression directly gives the expected score on the five-level MOS scale.

The variance of a function of the model parameters can be approximated using the delta method [18]:

$$\text{Var}(g(\mathbf{b})) \approx g'(\hat{\mathbf{b}})^\top \Sigma g'(\hat{\mathbf{b}}), \quad (6)$$

where Σ is the covariance matrix of the estimated parameters $\hat{\mathbf{b}}$, and $g'(\hat{\mathbf{b}})$ is the gradient of $g(\mathbf{b})$ evaluated at $\hat{\mathbf{b}}$.

For the cumulative sum of probabilities

$$S = \sum_{j=1}^4 F_j = g(\mathbf{b}) = \sum_{j=1}^4 \frac{1}{1 + e^{-\alpha_j + \beta^\top \mathbf{x}}}, \quad (7)$$

the gradient vector can be expressed as

$$g'(\mathbf{b})^\top = \left[\frac{\partial g(\mathbf{b})}{\partial \alpha}^\top \quad \frac{\partial g(\mathbf{b})}{\partial \beta}^\top \right]. \quad (8)$$

The partial derivatives with respect to the cutpoints are

$$\frac{\partial g(\mathbf{b})}{\partial \alpha_h} = \frac{e^{-\alpha_h + \beta^\top \mathbf{x}}}{(1 + e^{-\alpha_h + \beta^\top \mathbf{x}})^2}, \quad h = 1, 2, 3, 4, \quad (9)$$

and with respect to the coefficients β :

$$\frac{\partial g(\mathbf{b})}{\partial \beta_l} = -x_l \sum_{j=1}^4 \frac{e^{-\alpha_j + \beta^\top \mathbf{x}}}{(1 + e^{-\alpha_j + \beta^\top \mathbf{x}})^2}, \quad l = 1, 2, \dots, 26. \quad (10)$$

These expressions allow computing the variance of S , and consequently the variance of the expected MOS, via the delta method. For model fitting and estimation of the coefficients, as well as the covariance matrix, the statistical analysis system (SAS) software [20] was used.

C. BT.500, P.913, P.910 models

The statistical model defined in BT.500 follows a traditional approach based on data cleaning followed by score averaging to obtain the MOS [6]. In this model, outliers are first identified using the statistical procedures described in the standard, and unreliable subjects are removed before computing the final MOS. This makes it one of the most conventional methods for subjective quality analysis, where the focus is on filtering inconsistent ratings and then averaging the remaining scores. The P.913 model adopts a different strategy. Instead of removing subjects, it aims to correct subject bias. All participants are retained, but their scores are adjusted to compensate for personal rating tendencies so that the scales become more comparable across subjects. In practice, this works as a normalization step applied prior to MOS computation, rather than a rejection-based method [5]. P.910 introduces a more advanced statistical framework by explicitly modeling subject behavior. It accounts for both bias and inconsistency and uses maximum likelihood estimation (MLE) to estimate the underlying true quality alongside subject-specific parameters such as reliability. By weighting subjects according to their consistency, the model produces a consistency-aware MOS estimate [4].

D. Comparison criteria

The EPMOS model was trained on all four available subsets, resulting in four distinct series of estimated coefficients for the linear predictor μ . Consequently, using the equations presented in Section II-B, four corresponding series of EPMOS values were calculated for the subsets. Notably, EPMOS was trained on a raw subset without any subject outlier rejection, and the reference MOS was computed from the raw Netflix dataset [19]. The variance of each EPMOS estimate was computed using the delta method described in Section II-B. Similarly, the P.910, P.913, and BT.500 models were applied to each subset using the SUREAL toolkit [21], developed by Netflix, to compute the estimated MOS, with the results displayed in Figures 1 and 2. For all subsets generated from the reference dataset, the standard form of MOS was obtained by averaging the scores without introducing additional outlier rejection. The evaluation metrics used to quantify prediction error and correlation between the reference MOS—defined as the average of all available scores for each stimulus—and the model predictions were root mean squared error (RMSE) [22], mean absolute error (MAE), Spearman rank correlation coefficient (SRCC) [23], Kendall's rank correlation coefficient (KRCC) [24], and Pearson linear correlation coefficient (PLCC) [25].

III. NUMERICAL RESULTS

All calculated key performance indicator (KPI) results, reported in Tables II and III, are derived from the corresponding results for each video across its associated stimuli. Each video includes nine stimuli in total—three resolutions with three QP levels each. Table II presents a comparison of all the employed statistical models (rows) using subsets with 30 subject scores. All evaluations were performed against

the same reference MOS, calculated from the full reference dataset, which includes between 96 and 332 subjective scores per stimulus. The results for different videos are reported in columns. Table II shows that EPMOS generally achieves lower RMSE values than the other models, indicating better overall accuracy, while correlation metrics also suggest competitive performance. Since P.910 and P.913 do not remove outliers, using the reference MOS from all raw scores provides a fair baseline for comparing all models. Table II shows that, when predicting MOS for the larger dataset from smaller subsets, no single model consistently outperforms the others.

Table III indicates that, across subsets of 15, 20, 30, and 40 scores, the performance of EPMOS improves with increasing sample size. However, even with the smallest subset, it provides competitive RMSE and MAE values (even compared with other models with larger samples) and preserves satisfactory correlation measures, demonstrating its ability to reliably approximate MOS for larger populations from limited samples.

Figure 1 compares reference MOS calculated with all per-subject scores with EPMOS and MOS estimates from the 30-score subset, showing across videos (T2–T7) and three display resolutions that EPMOS follows a smoother, more monotonic trend. Figure 2 shows the MOS plots for P.913, P.910, and BT.500, also using the 30-score subset. Each row corresponds to a video, and columns show the three models. Across both figures, the 95% confidence interval (CI) for P.913 and P.910 is consistently narrower than that of the EPMOS model, whereas the widest CI is associated with BT.500 and the standard MOS.

TABLE II: Performance comparison of standardized P.910, P.913, BT.500, and EPMOS models for a subset with $N = 30$ scores per subject. Best performances per metric are bold and underlined.

Method	Metric	T2	T3	T4	T5	T6	T7
P910	RMSE	0.127	0.145	0.106	0.190	0.104	0.094
	MAE	0.107	0.103	0.097	0.142	0.079	0.075
	SRCC	0.983	1.000	0.967	0.950	0.983	1.000
	KRCC	0.944	1.000	0.889	0.833	0.944	1.000
	PLCC	0.991	0.984	0.991	0.969	0.995	0.998
P913	RMSE	0.093	0.156	0.143	0.154	0.083	0.103
	MAE	0.071	0.107	0.131	0.122	0.070	0.082
	SRCC	0.983	1.000	0.933	0.950	0.983	0.967
	KRCC	0.944	1.000	0.833	0.833	0.944	0.889
	PLCC	0.996	0.982	0.983	0.978	0.998	0.995
BT500	RMSE	0.127	0.143	0.105	0.191	0.105	0.095
	MAE	0.109	0.102	0.096	0.143	0.078	0.074
	SRCC	0.983	0.996	0.954	0.950	0.992	1.000
	KRCC	0.944	0.986	0.873	0.833	0.972	1.000
	PLCC	0.991	0.984	0.991	0.969	0.995	0.998
EPMOS	RMSE	0.074	0.108	0.062	0.136	0.086	0.117
	MAE	0.062	0.094	0.055	0.116	0.076	0.096
	SRCC	0.983	0.983	0.983	0.967	0.983	1.000
	KRCC	0.944	0.944	0.944	0.889	0.944	1.000
	PLCC	0.997	0.992	0.997	0.988	0.997	0.993

IV. CONCLUSION

In this paper, we evaluated several statistical models using smaller subsets of subjective scores to estimate MOS cal-

TABLE III: Performance of the EPMOS model evaluated on subsets containing $N = 15, 20, 30$, and 40 subject scores per stimulus.

N	Metric	T2	T3	T4	T5	T6	T7
15	RMSE	0.105	0.089	0.102	0.196	0.094	0.205
	MAE	0.086	0.081	0.085	0.162	0.083	0.182
	SRCC	0.967	1.000	0.917	0.917	0.983	0.917
	KRCC	0.889	1.000	0.778	0.778	0.944	0.778
	PLCC	0.996	0.993	0.995	0.971	0.998	0.974
20	RMSE	0.147	0.087	0.075	0.191	0.091	0.119
	MAE	0.125	0.079	0.069	0.154	0.080	0.109
	SRCC	0.967	0.983	0.933	0.917	1.000	0.983
	KRCC	0.889	0.944	0.833	0.778	1.000	0.944
	PLCC	0.988	0.994	0.995	0.967	0.995	0.991
30	RMSE	0.074	0.108	0.062	0.136	0.086	0.117
	MAE	0.062	0.094	0.055	0.116	0.076	0.068
	SRCC	0.983	0.983	0.983	0.967	0.983	1.000
	KRCC	0.944	0.944	0.944	0.889	0.944	1.000
	PLCC	0.997	0.992	0.997	0.988	0.997	0.993
40	RMSE	0.091	0.098	0.079	0.130	0.062	0.080
	MAE	0.068	0.083	0.057	0.114	0.055	0.059
	SRCC	0.967	0.983	0.983	0.967	0.983	1.000
	KRCC	0.889	0.944	0.944	0.889	0.944	1.000
	PLCC	0.995	0.983	0.997	0.988	0.998	0.996

culated from the scores from a greater number of subjects. The aim was to examine how effectively each model can predict reliable MOS values from limited data. To this end, we tested five different approaches: the conventional MOS computed by simple averaging, the EPMOS model based on ordinal logistic regression, and the P.910, P.913, and BT.500 models implemented in the SUREAL toolkit. The reference dataset was divided into four subsets, and each model was applied independently to each subset. The estimated results were then compared with the reference MOS calculated from the full dataset containing all available per-subject scores for each stimulus. The findings indicate that EPMOS often achieves the best estimation errors and correlations, though not for all contents and performance indicator. Moreover, we observed that applying EPMOS to the smallest subset, with only 15 scores per stimulus, still provided competitive results compared to the other models evaluated on larger subsets. The detailed results for all subsets, along with the coefficients estimated using EPMOS, will be made publicly available upon acceptance of the paper. In the future, this study can serve as a benchmark for developing new statistical models, allowing the comparison of their performance with the results presented here.

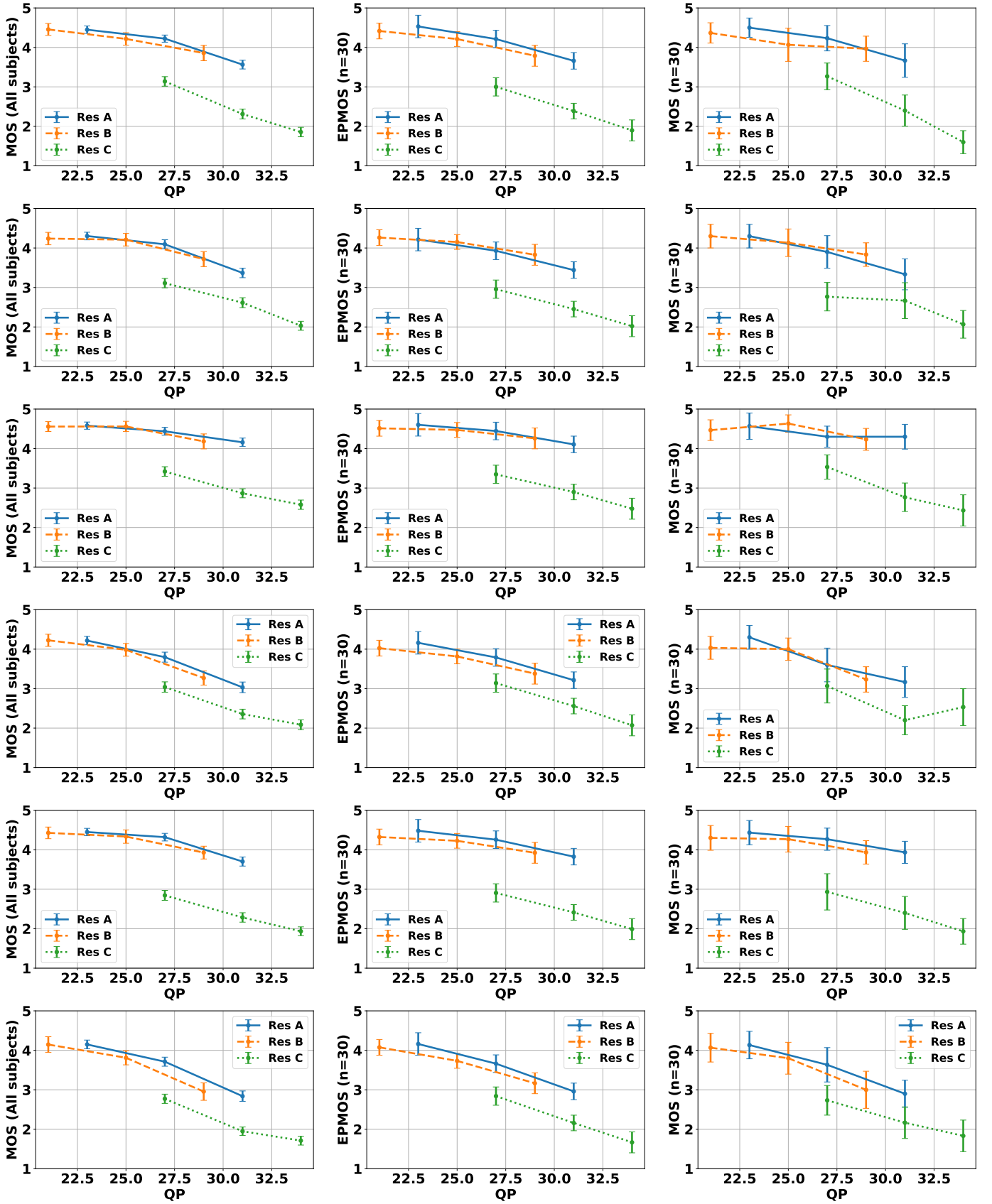


Fig. 1: Each row (top to bottom) corresponds to videos T2-T7 from the Netflix dataset [19]. From left to right, the plots show MOS using all available scores, EPMOS computed from 30 scores, and MOS computed from the same 30-score subset. Each plot contains three curves corresponding to resolutions A, B, and C, evaluated over their respective QP ranges.

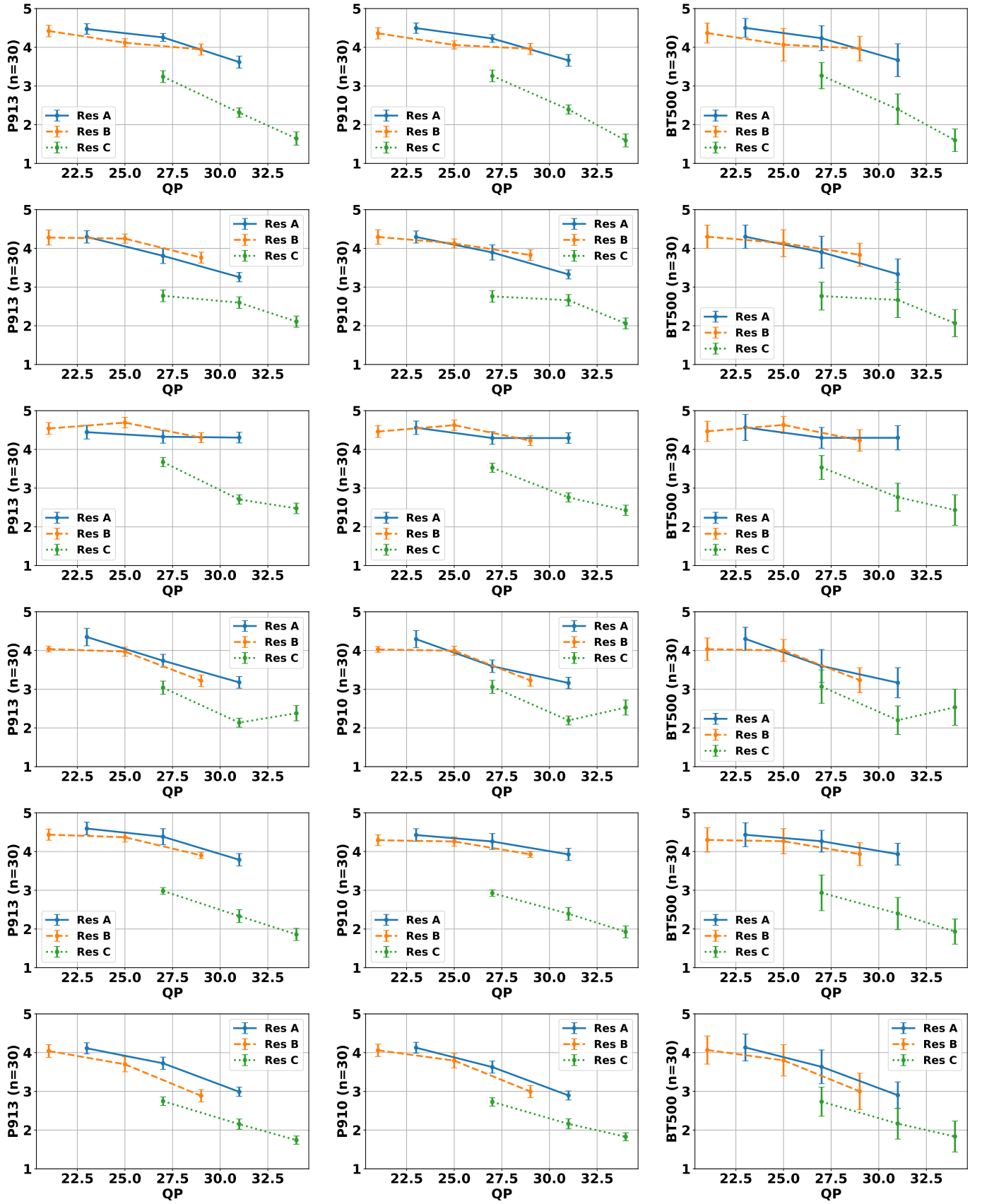


Fig. 2: Each row (top to bottom) corresponds to videos T2-T7 from the Netflix dataset [19]. From left to right, the plots show P913, P910, and BT.500 computed from 30 scores. Each plot contains three curves corresponding to resolutions A, B, and C, evaluated over their respective QP ranges.

REFERENCES

- [1] P. Pérez, L. Janowski, N. García, and M. Pinson, "Subjective assessment experiments that recruit few observers with repetitions (FOWR)," *IEEE Transactions on Multimedia*, vol. 24, pp. 3442–3454, 2021.
- [2] J. Nawala, L. Janowski, B. Ćmiel, K. Rusek, and P. Pérez, "Generalized score distribution: A two-parameter discrete distribution accurately describing responses from quality of experience subjective experiments," *IEEE Transactions on Multimedia*, vol. 25, pp. 6090–6104, 2022.
- [3] J. Sirengo, D. Alilah, D. Mbete, and R. Keli, "Estimation of risk factors affecting screening outcomes of prostate cancer using the Bayesian ordinal logistic model," *Journal of Probability and Statistics*, vol. 2023, no. 1, p. 4987764, 2023.
- [4] "Subjective video quality assessment methods — ITU-T Recommendation P.910, Annex E," International Telecommunication Union (ITU), Nov. 2021. [Online]. Available: <https://www.itu.int/rec/T-REC-P.910/>
- [5] "Methods for the subjective assessment of video quality — ITU-T Recommendation P.913, Section 12.4," International Telecommunication Union (ITU), Jun. 2021. [Online]. Available: <https://www.itu.int/rec/T-REC-P.913/>
- [6] "Methodologies for the subjective assessment of the quality of television pictures — ITU-R Recommendation BT.500-14, Section A1-2.3.1," International Telecommunication Union (ITU), Oct. 2019. [Online]. Available: <https://www.itu.int/rec/R-REC-BT.500/>
- [7] L. F. Tsiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Modeling subject scoring behaviors in subjective experiments based on a discrete quality scale," *IEEE Transactions on Multimedia*, vol. 26, pp. 8742–8757, 2024.
- [8] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.
- [9] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.
- [10] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, "NR-GVQM: A no reference gaming video quality metric," in *2018 IEEE international symposium on multimedia (ISM)*. IEEE, 2018, pp. 131–134.
- [11] H. Amirpour, J. Zhu, W. Zhou, P. Le Callet, and C. Timmerer, "VQM4HAS: A real-time quality metric for HEVC videos in HTTP adaptive streaming," *IEEE Transactions on Multimedia*, vol. 27, pp. 9619 – 9631, 2025.
- [12] M. Neri and F. Battisti, "Low-complexity patch-based no-reference point cloud quality metric exploiting weighted structure and texture features," *IEEE Transactions on Broadcasting*, 2025.
- [13] B. Chen, D. Huang, H. Zhu, L. Zhu, W. Zhou, S. Wang, Y. Fang, and W. Lin, "From global to granular: Revealing iqa model performance via correlation surface," *arXiv preprint arXiv:2601.21738*, 2026.
- [14] K. Mitra, H. S. Rossi, J. Gavrell, and C. Åhlund, "Quality of experience assessment for streamed social extended reality applications over heterogeneous access networks," in *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2025, pp. 1–7.
- [15] L. Peroni, S. Gorinsky, F. Tashtarian, and C. Timmerer, "Empowerment of atypical viewers via low-effort personalized modeling of video streaming quality," *Proceedings of the ACM on Networking*, vol. 1, no. CoNEXT3, pp. 1–27, 2023.
- [16] S. M. Mwalili, E. Lesaffre, and D. Declerck, "A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 54, no. 1, pp. 77–93, 2005.
- [17] C. G. Bampis, L. Krasula, Z. Li, and O. Akhtar, "Measuring and predicting perceptions of video quality across screen sizes with crowdsourcing," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2023, pp. 13–18.
- [18] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests - beyond subjects' MOS," *IEEE Transactions on Multimedia*, vol. 23, pp. 2505–2519, 2020.
- [19] C. Bampis, L. Krasula, Z. Li, and O. Akhtar, "Netflix public subjective quality dataset," https://docs.google.com/document/d/1Ly-r196i6ekIMATuA_qKZCSNpGoKj4shqhyXU8gbSnQ, Aug. 2023, accessed: 23 February 2026.
- [20] SAS Institute Inc., "SAS OnDemand for Academics: SAS Studio," <https://welcome.oda.sas.com/>, 2026, base SAS version 9.4_M8, Accessed: 2026-02-24.
- [21] Netflix Technology, "SUREAL: A library for subjective and objective video quality evaluation," <https://github.com/Netflix/sureal>, 2019, accessed: 2026-02-24.
- [22] D. Saupe and T. Bleile, "Robustness and accuracy of mean opinion scores with hard and soft outlier detection," in *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2025, pp. 1–7.
- [23] J. H. Zar, "Significance testing of the spearman rank correlation coefficient," *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 578–580, 1972.
- [24] H. Abdi *et al.*, "The kendall rank correlation coefficient," *Encyclopedia of measurement and statistics*, vol. 2, pp. 508–510, 2007.
- [25] P. Sedgwick, "Pearson's correlation coefficient," *Bmj*, vol. 345, 2012.