2025 7th International Conference on Applied Computational Intelligence in Information Systems (ACIIS) | 979-8-3315-8877-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ACIIS66255.2025.11402948

Relatively Robust QoS and QoE Score Aggregation

Thomas A. Weber
Chair of Operations, Economics and Strategy
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
thomas.weber@epfl.ch

***Abstract*—We present a robust methodology for aggregating Quality of Service (QoS) and Quality of Experience (QoE) scores across multiple criteria in settings where the relative importance of individual metrics (i.e., weights) is unknown or ambiguous. Building on recent advances in relatively robust multicriteria decision theory, we define a Robust Aggregate Score (RAS) based on the worst-case performance ratio with respect to all admissible weight configurations. The method is weight-agnostic, invariant to rescaling of individual metrics, and accommodates criteria with both positive and negative preference direction. We illustrate the approach in the context of video-streaming service evaluation and demonstrate how it enables interpretable and defensible rankings across heterogeneous quality indicators. The proposed framework is broadly applicable to quality assessment tasks that require reliability under weight uncertainty.**

***Index Terms*—Performance evaluation, relative robustness, score aggregation, quality of experience (QoE), quality of service (QoS)**

I. Introduction

Quality assessment across multiple service dimensions (e.g., latency, packet loss, bitrate, user experience) is central to selecting optimal configurations in telecommunications, streaming, and gaming. When comparing multiple providers or configurations, users often lack precise preferences or knowledge of weightings for each criterion. Hence, robust aggregation mechanisms that protect against worst-case interpretations of preferences become critical. This paper introduces a robust score aggregation method for evaluating service alternatives, based on a general approach to relatively robust decision-making developed by Weber [1], [2].

Service	QoS Metrics	QoE Metrics
Audio	Latency, loss, jitter	MOS, PESQ, POLQA
Video	Bitrate, buffering, resolution, frame rate	VMAF, SSIM, PSNR, MOS
VR/AR	Frame rate, motion delay, sync, tracking	SSQ, presence scales, ratings
Gaming	Ping, jitter, loss, input lag, FPS	GEQ, MOS, engagement scores

TABLE I
QoS and QoE Metrics by Service Type

Table I summarizes typical Quality of Service (QoS) metrics and Quality of Experience (QoE) evaluation methods across four key service domains: audio, video streaming, virtual/augmented reality (VR/AR), and gaming. QoS metrics refer to objective, network-level measurements such as latency (the delay between sending and receiving data), jitter (the variability in packet arrival times), packet loss (the percentage of data lost during transmission), bitrate (the volume of data transmitted per unit time), and frame rate (the number of frames rendered per second). These parameters directly affect user-perceived quality. The corresponding QoE metrics aim to quantify subjective user satisfaction. For audio, common methods include MOS (Mean Opinion Score), PESQ (Perceptual Evaluation of Speech Quality), and POLQA (Perceptual Objective Listening Quality Assessment). In video, tools such as VMAF (Video Multi-Method Assessment Fusion), SSIM (Structural Similarity Index Measure), and PSNR (Peak Signal-to-Noise Ratio) are often used to estimate visual fidelity. VR/AR services are evaluated through SSQ (Simulator Sickness Questionnaire), as well as presence and immersion scales that assess user comfort and realism. Gaming QoE is measured using instruments like GEQ (Game Experience Questionnaire), Gaming MOS, and engagement scoring frameworks that assess responsiveness and playability under varying network conditions.

II. Literature

The evaluation of services based on multiple quality dimensions has long been studied in the fields of networking, multimedia systems, economics, and decision theory. This section reviews the main strands of literature that motivate our objective: to develop a robust aggregation method for QoS and QoE scores in the presence of weight ambiguity.

A. Multiattribute Utility

Lancaster [3] describes any product as a bundle of attributes. These attributes, which may be exploited for price discrimination (see, e.g., [4], [5]), can also be interpreted as objectives or criteria in their own right, thereby inducing a multicriteria decision problem for any agent seeking to identify the most preferred alternative [6], [7].

B. QoS and QoE Metrics

Traditional Quality of Service (QoS) metrics—such as latency, jitter, packet loss, and throughput—provide objective, infrastructure-level indicators of system performance. These metrics are often specific to service types: for example, latency and jitter are critical for VoIP [8], whereas bitrate and rebuffering are central to video streaming performance [9].

979-8-3315-8877-9/25/$31.00 ©2025 IEEE

Quality of Experience (QoE), while arguably related to QoS [10], captures subjective user satisfaction and is commonly assessed through standardized tools such as the Mean Opinion Score (MOS) [11], PESQ [12], and POLQA [12]. For video, perceptual metrics such as VMAF [13], SSIM, and PSNR are widely used to estimate visual fidelity. In interactive environments like gaming and VR/AR, methods such as the Game Experience Questionnaire (GEQ) [14], the Simulator Sickness Questionnaire (SSQ), and presence or immersion scales are commonly applied [15].[1] The synthesis of these metrics across domains has been discussed in comprehensive overviews such as the Qualinet white paper on QoE [16].

C. Score Aggregation Approaches

A central challenge in aggregating diverse QoS and QoE metrics lies in selecting and applying appropriate criterion weights. Classical methods include weighted sums, geometric means, and more general power means, all of which can be unified through the Kolmogorov (or quasi-arithmetic) mean framework [2]. Other techniques include multiattribute utility theory [6], which typically yields concave objectives, and ordered weighted averaging (OWA) operators [17]; both approaches tend to rely heavily on subjective input.

In network QoE research, various attempts have been made to model the relationship between QoS inputs and a single QoE output using empirical fitting or data-driven methods [18]. However, such models often suffer from limited generalizability due to their dependence on specific training datasets and fixed user profiles. More importantly, they typically aim to reduce the dimensionality of the criterion space, without addressing the more fundamental issue of ambiguity in the relative importance of the criteria themselves.

D. Robust and Weight-Agnostic Methods

To address the fundamental issue of weight ambiguity, recent work in robust optimization proposes decision-theoretic frameworks that remain defensible under all plausible weight configurations. Relative robustness measures—initially explored in evaluating online vs. offline performance of algorithms [19] and later formalized in the context of fair resource allocation [20], as well as other economic and operational decisions [1], [2]—considers the worst-case performance ratio across all admissible weights. Instead of targeting optimality under a fixed weight vector, relatively robust methods identify alternatives that achieve a guaranteed fraction of the best possible weighted score. The performance index introduced by Weber [1] provides a principled means of comparing alternatives when the true weighting is unknown. It has been extended to discrete decision sets with efficient computational procedures [2], making it applicable to QoS/QoE score aggregation problems where decisions are finite and data-driven.

E. Contributions

While much existing work assumes known weights or emphasizes parametric learning, our contribution lies in applying relative robustness to QoS/QoE score aggregation. We propose a robust, weight-agnostic index for evaluating alternatives (e.g., service providers or configurations) across multiple performance metrics. The resulting robust aggregate score (RAS) provides a quantitative robustness guarantee with respect to all possible weights, avoids subjective bias, is invariant under linear scaling of any given criterion, and supports modular aggregation (e.g., separate RAS values for QoS and QoE criteria).

III. Robust Average Scoring

Let $\mathcal{X} = \{1, \ldots, J\}$ be a finite set of services, each evaluated using n criteria $f_1, \ldots, f_n : \mathcal{X} \to \mathbb{R}_{++}$. Thus, any alternative $j \in \mathcal{X}$ receives a score $f_i(j) > 0$ when evaluated by the criterion with index $i \in \mathcal{N} = \{1, \ldots, n\}$. For a given weight vector $\boldsymbol{\lambda} \in \Delta = \big\{(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n_+ : \lambda_1 + \cdots + \lambda_n = 1\big\}$, the decision-maker would want to choose the service j that maximizes the weighted score,

$$F(j|\boldsymbol{\lambda}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}(j) = \sum_{i=1}^{n} \lambda_i f_i(j), \quad j \in \mathcal{X},$$

where $\boldsymbol{f} = (f_1, \ldots, f_n)$ denotes the (vector-valued) multicriteria function used to evaluate the various available alternatives.

Remark 1 (Preference Direction). When maximizing the weighted score $F(\cdot|\boldsymbol{\lambda})$ to obtain a most preferred alternative, an implicit assumption is that all criteria describe "desirable" attributes—in the sense that higher values are better. Because this may not always apply,[2] the proposed robust average scoring method is extended in Sec. III-E to accommodate criteria with arbitrary preference direction. Until then, we maintain the assumption that higher criterion values are preferred.

A. Scoring Requirements

Being unsure about which weight vector $\boldsymbol{\lambda} \in \Delta$ would accurately represent the relevant tradeoffs, the decision-maker seeks a *robust aggregate score* (RAS) $\rho : \mathcal{X} \to [0, 1]$, based on which a most preferred service option can be determined. To this end, we posit three natural *scoring requirements*:

R1) **Robustness Guarantee.** The RAS ensures the best possible score achievement of the chosen option relative to an "ex-post optimal score" (obtained for any known weight), evaluated over all possible weights in Δ.

R2) **Scaling Invariance.** The RAS is invariant under multiplying any of the criteria by a positive scaling factor.

R3) **Modular Aggregation.** Joining two multicriteria functions, $\boldsymbol{f}$ (of dimension n) and $\boldsymbol{g}$ (of dimension m), into $\boldsymbol{h} = (\boldsymbol{f}, \boldsymbol{g})$ yields the RAS $\rho_{\boldsymbol{h}}$ as a *symmetric* function of the scores $\rho_{\boldsymbol{f}}$ and $\rho_{\boldsymbol{g}}$, independent of n and m.

Requirement R1 ensures that the relative regret of the most preferred service is minimized with respect to any other

[1] Objective perceptual measures such as PESQ (P.862) were long regarded as state-of-the-art. However, the entire P.862 series was officially withdrawn in Jan. 2024 and superseded by POLQA (P.863) [12], which now serves as standardized reference method for narrow- and wideband speech quality.

[2] For example, a lower startup delay is preferred for a video-streaming service; cf. Sec. IV.

available option—under any possible weighting of the criteria. Requirement R2 states that a positive linear scaling of units does not influence the RAS for any feasible alternative in $\mathcal{X}$. For example, multiplying criterion f_1 by 10 while keeping $f_2, \ldots, f_n$ unchanged would leave all RAS values invariant. Finally, requirement R3 stipulates that the joint evaluation of two sets of criteria (e.g., QoS and QoE) should involve only the two already computed RAS values. The resulting aggregate must be independent of (i) more granular information (such as individual criterion scores or dimensions), and (ii) the order of aggregation—that is, whether $\boldsymbol{h} = (\boldsymbol{f}, \boldsymbol{g})$ or $\boldsymbol{h} = (\boldsymbol{g}, \boldsymbol{f})$.

B. Relative Robustness Measure

To evaluate the achievement of a feasible alternative for any given weight, consider the *performance ratio*,

$$\varphi(j|\boldsymbol{\lambda}) = \frac{F(j|\boldsymbol{\lambda})}{F^*(\boldsymbol{\lambda})}, \quad (j, \boldsymbol{\lambda}) \in \mathcal{X} \times \Delta,$$

where $F^*(\boldsymbol{\lambda}) = \max_{j \in \mathcal{X}} F(j|\boldsymbol{\lambda})$ is the best possible value of the weighted objective under the weight $\boldsymbol{\lambda}$. Thus, the *robust aggregate score* (RAS) (or "performance index" [1], [2]),

$$\rho(j) = \min_{\boldsymbol{\lambda} \in \Delta} \varphi(j|\boldsymbol{\lambda}), \quad j \in \mathcal{X},$$

is the worst-case performance ratio with respect to all possible weights. It provides a relative performance guarantee (as stipulated by requirement R1), in the sense that

$$F(j|\boldsymbol{\lambda}) \geq \rho(j)\, F^*(\boldsymbol{\lambda}), \quad (j, \boldsymbol{\lambda}) \in \mathcal{X} \times \Delta.$$

More specifically, a "robust alternative" $j^* \in \mathcal{X}$, such that

$$j^* \in \arg\max_{j \in \mathcal{X}} \rho(j),$$

achieves the *optimal* RAS $\rho^* = \max_{j \in \mathcal{X}} \rho(j)$, so that

$$F(j^*|\boldsymbol{\lambda}) \geq \rho^* F^*(\boldsymbol{\lambda}) = \rho^* \cdot \left(\max_{j \in \mathcal{X}} F(j|\boldsymbol{\lambda}) \right), \quad \boldsymbol{\lambda} \in \Delta.$$

Thus, the robust alternative j^* is always within $(1 - \rho^*)$ (in percent) of the weighted score of any other feasible option, with respect to any possible weight $\boldsymbol{\lambda} \in \Delta$.

Remark 2 (Relative Regret). The RAS is closely related to the concept of relative regret, since $\rho(j) = 1 - \mathrm{RR}(j)$, where $\mathrm{RR}(j) = \max_{\boldsymbol{\lambda} \in \Delta} \{(F^*(\boldsymbol{\lambda}) - F(j|\boldsymbol{\lambda}))/F^*(\boldsymbol{\lambda})\}$ denotes the (maximum) relative regret incurred by alternative $j \in \mathcal{X}$.

C. Representation of the RAS

Given that the RAS is the worst-case performance ratio with respect to all possible weights, it is useful to note that the level sets of $\varphi(j|\cdot)$ are convex for any fixed alternative j.

Lemma 1. *For any option $j \in \mathcal{X}$, the performance ratio $\varphi(j|\cdot)$ is quasiconcave on Δ.*

Proof. Fix $j \in \mathcal{X}$, and define the upper contour set

$$\mathcal{C}_\gamma(j) = \{\boldsymbol{\lambda} \in \Delta : \varphi(j|\boldsymbol{\lambda}) \geq \gamma\}, \quad \gamma \in [0, 1].$$

Since $\mathcal{C}_0 = \Delta$, there exists $\gamma \in [0, 1]$ such that $\mathcal{C}_\gamma(j) \neq \emptyset$. Choose two weights $\boldsymbol{\lambda}', \boldsymbol{\lambda}'' \in \mathcal{C}_\gamma(j)$ and a scalar $\kappa \in (0, 1)$. The convex combination $\boldsymbol{\lambda}_\kappa = \kappa \boldsymbol{\lambda}' + (1 - \kappa)\boldsymbol{\lambda}''$ lies in $\mathcal{C}_\gamma(j)$ if and only if $\varphi(j|\boldsymbol{\lambda}_\kappa) = F(j|\boldsymbol{\lambda}_\kappa)/F^*(\boldsymbol{\lambda}_\kappa) \geq \gamma$.

Given that $\varphi(j|\boldsymbol{\lambda}')$ and $\varphi(j|\boldsymbol{\lambda}'')$ both exceed γ (weakly), we have

$$\begin{aligned} F(j|\boldsymbol{\lambda}_\kappa) &= F(j|\kappa \boldsymbol{\lambda}' + (1 - \kappa)\boldsymbol{\lambda}'') \\ &= \kappa F(j|\boldsymbol{\lambda}') + (1 - \kappa) F(j|\boldsymbol{\lambda}'') \\ &\geq \gamma \left(\kappa F^*(\boldsymbol{\lambda}') + (1 - \kappa) F^*(\boldsymbol{\lambda}'') \right). \end{aligned}$$

Since $\boldsymbol{f}(\mathcal{X}) \subset \mathbb{R}^n_{++}$ by hypothesis, $F^*(\boldsymbol{\lambda}')$ and $F^*(\boldsymbol{\lambda}'')$ are strictly positive. Now define

$$B = \kappa F^*(\boldsymbol{\lambda}') + (1 - \kappa) F^*(\boldsymbol{\lambda}'').$$

Observe that

$$\begin{aligned} B &= \max_{j', j'' \in \mathcal{X}} \{\kappa F(j'|\boldsymbol{\lambda}') + (1 - \kappa) F(j''|\boldsymbol{\lambda}'')\} \\ &\geq \max_{j \in \mathcal{X}} \{\kappa F(j|\boldsymbol{\lambda}') + (1 - \kappa) F(j|\boldsymbol{\lambda}'')\} \\ &= \max_{j \in \mathcal{X}} \{F(j|\kappa \boldsymbol{\lambda}' + (1 - \kappa)\boldsymbol{\lambda}'')\} = F^*(\boldsymbol{\lambda}_\kappa). \end{aligned}$$

Thus, $F(j|\boldsymbol{\lambda}_\kappa) \geq \gamma F^*(\boldsymbol{\lambda}_\kappa)$, which implies that $\boldsymbol{\lambda}_\kappa \in \mathcal{C}_\gamma(j)$. Hence, the upper contour set is convex, and $\varphi(j|\cdot)$ is quasiconcave. □

The preceding result implies that, in computing the RAS, it is sufficient to restrict attention to the (unique) set of extreme points of Δ.[3]

Proposition 1. *The RAS can be represented as*

$$\rho(j) = \min\{\varphi_1(j), \ldots, \varphi_n(j)\}, \quad j \in \mathcal{X},$$

where $\varphi_i(j) = f_i(j)/f_i^*$ for all $i \in \{1, \ldots, n\}$ measures the i-th criterion-specific achievement relative to the corresponding maximum criterion score $f_i^* = \max_{\ell \in \mathcal{X}} f_i(\ell)$.

Proof. By Lemma 1, the minimum of the performance ratio over all possible weights is attained on the boundary of Δ. Hence—again by quasiconcavity—it is attained on the set of extreme points of Δ (cf. footnote 3), so

$$\rho(j) = \min\{\varphi(j|\boldsymbol{e}_1), \ldots, \varphi(j|\boldsymbol{e}_n)\}, \quad j \in \mathcal{X},$$

which is equivalent to the claimed representation. □

D. Properties of the RAS

The representation of ρ as the minimum of n criterion-specific performance ratios (cf. Prop. 1) immediately implies that the RAS satisfies requirement R2.

Proposition 2. *The RAS is invariant with respect to a rescaling of criteria. That is, for any $\boldsymbol{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_n) \succ 0$, if $\hat{\boldsymbol{f}}(\cdot \mid \boldsymbol{A}) = \boldsymbol{A}\boldsymbol{f}(\cdot)$, then*

$$\hat{\rho}(j \mid \boldsymbol{A}) = \min_{\boldsymbol{\lambda} \in \Delta} \left\{ \frac{\boldsymbol{\lambda} \cdot \hat{\boldsymbol{f}}(j \mid \boldsymbol{A})}{\hat{F}^*(\boldsymbol{\lambda} \mid \boldsymbol{A})} \right\} = \rho(j), \quad j \in \mathcal{X},$$

where $\hat{F}^*(\boldsymbol{\lambda} \mid \boldsymbol{A}) = \max_{\ell \in \mathcal{X}} \boldsymbol{\lambda} \cdot \hat{\boldsymbol{f}}(\ell)$.

[3] The set of extreme points of Δ is $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$, where $\boldsymbol{e}_i$ is the i-th Euclidean unit vector, for all $i \in \mathcal{N}$.

Proof. Consider any $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_{++}$, and let $\boldsymbol{A} = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \succ 0$. Then, by Prop. 1,

$$\hat{\rho}(j \mid \boldsymbol{A}) = \min_{i \in \mathcal{N}} \left\{ \frac{\boldsymbol{e}_i \cdot \hat{\boldsymbol{f}}(j \mid \boldsymbol{A})}{\hat{F}^*(\boldsymbol{e}_i \mid \boldsymbol{A})} \right\} = \min_{i \in \mathcal{N}} \left\{ \frac{\alpha_i f_i(j)}{\alpha_i f_i^*} \right\} = \rho(j),$$

for all $j \in \mathcal{X}$, which completes the proof. □

As a corollary of the preceding result (in conjunction with Prop. 1), it follows that if one criterion is a positive affine transformation of another, then that criterion may be dropped without altering the RAS. Meanwhile, a simple modular method for aggregating evaluations based on different multicriteria functions—as required by R3—is established next.

Proposition 3. *The RAS is min-additive. That is, when combining $\rho_{\boldsymbol{f}}$ for the multicriteria vector $\boldsymbol{f} = (f_1, \ldots, f_n)$ and $\rho_{\boldsymbol{g}}$ for the multicriteria vector $\boldsymbol{g} = (g_1, \ldots, g_m)$, with given integers $n, m \geq 1$, the joint RAS satisfies*

$$\rho_{\boldsymbol{h}}(j) = \min \left\{ \rho_{\boldsymbol{f}}(j), \rho_{\boldsymbol{g}}(j) \right\}, \quad j \in \mathcal{X},$$

where $\boldsymbol{h} = (f_1, \ldots, f_n, g_1, \ldots, g_m)$.

Proof. Applying Prop. 1 to the criterion vector $\boldsymbol{h}$ of length $n + m$ yields

$$\begin{aligned} \rho_{\boldsymbol{h}}(j) &= \min \left\{ \frac{f_1(j)}{f_1^*}, \ldots, \frac{f_n(j)}{f_n^*}, \frac{g_1(j)}{g_1^*}, \ldots, \frac{g_m(j)}{g_m^*} \right\} \\ &= \min \left\{ \rho_{\boldsymbol{f}}(j), \rho_{\boldsymbol{g}}(j) \right\}, \quad j \in \mathcal{X}, \end{aligned}$$

as claimed. □

It is evident that the min-additive aggregation of RAS values across different criterion sets, as described by Prop. 3, is symmetric and does not require knowledge of the number of criteria in each subset.[4]

E. Criteria with Arbitrary Preference Direction

As mentioned in Remark 1, some QoS/QoE criteria are such that higher scores are better, while others are preferable when lower. Given the representation of the RAS in Prop. 1, it is natural to include criteria with negative preference direction (i.e., where smaller values are more desirable) by inverting the ratio—taking the theoretically achievable minimum score divided by the actual score.

Let the index set of all criteria be partitioned as $\mathcal{N} = \mathcal{N}_+ \cup \mathcal{N}_-$, with $\mathcal{N}_+ \cap \mathcal{N}_- = \emptyset$, where $\mathcal{N}_+$ contains criteria with positive preference direction, and $\mathcal{N}_-$ those with negative direction. Then, the generalized RAS can be written as

$$\rho(j) = \min \left\{ \rho_+(j), \rho_-(j) \right\}, \quad j \in \mathcal{X},$$

where

$$\rho_+(j) = \min_{i \in \mathcal{N}_+} \left\{ \frac{f_i(j)}{f_i^*} \right\} \quad \text{and} \quad \rho_-(j) = \min_{i \in \mathcal{N}_-} \left\{ \frac{f_i^\circ}{f_i(j)} \right\}.$$

Here, $f_i^* = \max_{\ell \in \mathcal{X}} f_i(\ell)$ for $i \in \mathcal{N}_+$ is the maximum criterion-specific score, while $f_i^\circ = \min_{\ell \in \mathcal{X}} f_i(\ell)$ for $i \in \mathcal{N}_-$ is the (strictly positive) minimum score. This ensures that each normalized score remains within the unit interval and interpretable in terms of relative performance.

IV. Application: Video Streaming

To illustrate the robust aggregate scoring method, we now apply it to the selection of a video-streaming service based on both QoS and QoE criteria. The synthetic data used in this example, while not drawn from actual measurements, reflects "reasonable" values given the current state of the art. Its sole purpose is to provide a realistic use case for aggregate scoring in environments characterized by multiple, and at times conflicting, quality indicators.

A. Quality of Service (QoS)

In addition to natural service attributes such as bitrate, frame rate, and resolution, it is common to include metrics like startup delay, buffering ratio, packet loss rate, round-trip time, and jitter when assessing the quality of a video-streaming service [9], [21]–[23]. Table II describes the eight QoS metrics used in this evaluation.

Metric	**Description**	**Unit**
Bitrate	Video encoding rate	kbps
Startup Delay	Time until playback begins	ms
Buffering Ratio	Share of time spent buffering	%
Frame Rate	Frames displayed per second	fps
Resolution	Display resolution	pixels (e.g., 1080p)
Packet Loss Rate	Lost packets during delivery	%
Round-Trip Time (RTT)	Time for round-trip transmission	ms
Jitter	Packet inter-arrival variability	ms

TABLE II
QoS Metrics for Video Streaming and Their Units

We assume that five different video-streaming providers (labeled A–E) have been evaluated with respect to these QoS criteria, as shown in Table III. These measurements are assumed to be reasonably accurate.

Prov.	**BR**	**SD**	**Buf**	**FPS**	**Res**	**PLR**	**RTT**	**Jit**
A	5500	2500	1.1	30	1080	1.5	42	5.5
B	6000	3000	1.4	45	1080	1.8	55	6.0
C	4200	1050	0.9	29	1080	0.4	40	4.5
D	2800	350	1.0	27	720	0.5	75	3.0
E	3900	200	0.8	24	720	0.3	50	2.0

TABLE III
QoS Metrics for Video Streaming Providers. BR: Bitrate (kbps), SD: Startup Delay (ms), Buf: Buffering Ratio (%), FPS: Frame Rate (#/s), Res: Resolution (p), PLR: Packet Loss Rate (%), RTT: Round-Trip Time (ms), Jit: Jitter (ms).

While providers A, B, and C tend to outperform D and E in terms of throughput (bitrate) and frame rate, they also exhibit higher jitter and more frequent buffering. As such, no single provider dominates across all dimensions, making robust aggregation particularly valuable in this context.

B. Quality of Experience (QoE)

Video-streaming QoE metrics—such as MOS, SSIM, and PSNR—have been extensively validated through subjective studies [24]–[26]. The VMAF metric, developed by Netflix,

[4] This contrasts with standard averaging. For instance, combining the arithmetic means of n and m positive numbers into a single arithmetic mean of all $n + m$ values requires knowledge of both n and m.

has demonstrated strong correlation with MOS ratings [27]. In addition, session-level indicators such as rebuffer counts, playback failures, and QoE drop rates have been discussed in detail in [28]. Table IV summarizes eight key QoE metrics relevant for video streaming.

Metric	Description	Unit
MOS	Mean user quality rating	1–5
VMAF	Perceptual quality score (Netflix)	0–100
SSIM	Structural similarity to reference image	$[0, 1]$
PSNR	Peak signal-to-noise ratio	dB
AWR	Average Watch Ratio: (watch time)/(video length)	%
EBVS	Exit Before Video Starts: (# exits)/(# plays)	%
ER	Engagement Rate: (# engagements)/(# viewers)	%
QoE Drop Rate	Sessions with MOS below threshold	%

TABLE IV
QoE Metrics for Video Streaming and Their Units

We assume that a complete data set is available for five streaming providers, as shown in Table V. These QoE scores are naturally related to the previously shown QoS results. For instance, providers with lower QoS—e.g., high buffering or packet loss—often exhibit degraded MOS values, reduced engagement, or elevated drop rates.

Prov.	MOS	VMAF	SSIM	PSNR	AWR	EBVS	ER	QDR
A	4.3	87	0.93	40.0	85%	3%	3.5%	1.0%
B	3.1	94	0.96	48.5	73%	5%	2.0%	0.5%
C	3.9	70	0.80	45.0	82%	2%	1.5%	0.8%
D	3.5	66	0.77	31.5	65%	7%	4.0%	0.9%
E	4.1	83	0.90	36.5	88%	3%	3.0%	1.2%

TABLE V
QoE Metrics for Video Streaming Providers. MOS: Mean Opinion Score (1–5), VMAF: Video Multi-method Assessment Fusion (0–100), SSIM: Structural Similarity Index (0–1), PSNR: Peak Signal-to-Noise Ratio (dB), AWR: Average Watch Ratio (%), EBVS: Exit Before Video Start (%), ER: Engagement Rate (%), QDR: QoE Drop Rate (%).

Metric (Abbreviation)	Preference Direction
QoS Metrics	
Bitrate (BR)	$+$
Startup Delay (SD)	$-$
Buffering Ratio (Buf)	$-$
Frame Rate (FPS)	$+$
Resolution (Res)	$+$
Packet Loss Rate (PLR)	$-$
Round-Trip Time (RTT)	$-$
Jitter (Jit)	$-$
QoE Metrics	
Mean Opinion Score (MOS)	$+$
VMAF	$+$
SSIM	$+$
PSNR	$+$
Average Watch Ratio (AWR)	$+$
Exit Before Video Start (EBVS)	$-$
Engagement Rate (ER)	$+$
QoE Drop Rate (QDR)	$-$

TABLE VI
Preference Direction of QoS and QoE Metrics.
[$+$ indicates higher is better, $-$ indicates lower is better.]

Prov.	BR	SD	Buf	FPS	Res	PLR	RTT	Jit
A	0.917	0.080	0.727	0.667	1.000	0.200	0.952	0.364
B	1.000	0.067	0.571	1.000	1.000	0.167	0.727	0.333
C	0.700	0.190	0.889	0.644	1.000	0.750	1.000	0.444
D	0.467	0.571	0.800	0.600	0.667	0.600	0.533	0.667
E	0.650	1.000	1.000	0.533	0.667	1.000	0.800	1.000

TABLE VII
Normalized QoS Performance Ratios for Video Streaming Providers.
Values scaled to [0,1] based on preference direction.

Prov.	MOS	VMAF	SSIM	PSNR	AWR	EBVS	ER	QDR
A	1.000	0.926	0.969	0.825	0.966	0.667	0.875	0.500
B	0.721	1.000	1.000	1.000	0.830	0.400	0.500	1.000
C	0.907	0.745	0.833	0.928	0.932	1.000	0.375	0.625
D	0.814	0.702	0.802	0.649	0.739	0.286	1.000	0.556
E	0.953	0.883	0.938	0.753	1.000	0.667	0.750	0.417

TABLE VIII
Normalized QoE Performance Ratios for Video Streaming Providers.
Values scaled to [0,1] based on metric preferences.

C. Score Aggregation

Before aggregating the different criterion-specific scores into an overall RAS, it is important to partition the index sets according to the preference directions, as described in Sec. III-E. Let $\boldsymbol{f} = (f_1, \ldots, f_8)$ be the multicriteria function for the QoS metrics (with index set $\mathcal{N} = \{1, \ldots, 8\}$), and let $\boldsymbol{g} = (g_1, \ldots, g_8)$ be the multicriteria function for the QoE metrics (with index set $\mathcal{M} = \{1, \ldots, 8\}$). The corresponding preference direction partitions are summarized in Table VI, yielding $\mathcal{N}_+ = \{1, 4, 5\}$ and $\mathcal{N}_- = \{2, 3, 6, 7, 8\}$ for QoS, and $\mathcal{M}_+ = \{1, 2, 3, 4, 5, 7\}$ and $\mathcal{M}_- = \{6, 8\}$ for QoE.

The normalized criterion-specific performance ratios for all providers are shown in Table VII for QoS and in Table VIII for QoE. Each value reflects a provider's relative standing in that dimension, scaled to $[0, 1]$ based on preference direction.

Based on the QoS-RAS alone, provider E achieves the best robust aggregate score with $\rho^*_{\boldsymbol{f}} = 53.3\%$, followed by provider D with 42.9%. In terms of QoE-RAS alone, provider A is ranked highest with $\rho^*_{\boldsymbol{g}} = 50.0\%$, followed by provider E with 41.7%.

The combined RAS—obtained by aggregating the QoS and QoE scores via the min-additive rule from Prop. 3—is reported in Table IX. Provider E emerges as the most robust overall option, with $\rho^*_{\boldsymbol{h}} = \min\{0.533, 0.417\} = 41.7\%$, where $\boldsymbol{h} = (\boldsymbol{f}, \boldsymbol{g})$ denotes the concatenation of the QoS and QoE criteria. An RAS of 41.7% indicates that, regardless of how the 16 criteria are weighted, provider E's overall score will be at least 41.7% of the best possible weighted score attainable by any provider. No other alternative achieves such a high guaranteed minimum, meaning that all other services deviate more substantially from the optimal under some admissible weighting configurations.

Provider	QoS-RAS	QoE-RAS	Overall-RAS	Rank
A	0.080	0.500	0.080	4
B	0.067	0.400	0.067	5
C	0.190	0.375	0.190	3
D	0.467	0.286	0.286	2
E	0.533	0.417	0.417	1

TABLE IX
Robust Average Scores (RAS) for Video Streaming Providers.
QoS-RAS and QoE-RAS are the worst (minimum) normalized scores across respective dimensions; Overall-RAS is the minimum of these two. The best provider achieves the highest Overall-RAS.

V. Conclusion

This paper introduced a robust score aggregation method for evaluating service alternatives across multiple Quality of Service (QoS) and Quality of Experience (QoE) metrics. Grounded in relative robustness, the method enables decision-making under complete uncertainty about criterion weights. By minimizing the worst-case performance ratio over all admissible weight vectors, the Robust Aggregate Score (RAS) provides a weight-agnostic benchmark that avoids subjective preference elicitation and yields interpretable, defensible quality assessments, together with an explicit performance guarantee over the entire weight space.

The RAS satisfies key axiomatic properties, including invariance under positive rescaling of individual metrics and modular min-additivity, which enables consistent aggregation across disjoint metric groups (e.g., QoS and QoE). The method accommodates criteria with arbitrary preference direction—whether higher or lower values are desirable—and admits a direct computational representation via criterion-specific performance ratios against best or worst attainable values. Importantly, the RAS is insensitive to redundant information: adding any number of metrics that are perfectly correlated with an already included metric does not change the score. Hence combining QoS and QoE is not redundant: if QoE carried no additional information beyond QoS, the overall RAS would coincide with the QoS-only RAS; if QoE adds distinct information (e.g., perceptual or behavioral outcomes), the combined RAS properly reflects it without double counting.

A detailed application to video-streaming service evaluation demonstrated how the RAS identifies robustly optimal providers without requiring knowledge of user preferences or any training procedure. Compared to classical weighted averages or empirical models, the method offers transparent guarantees on worst-case relative performance, independent of any assumed weighting configuration.

The robust aggregation framework extends naturally to other multi-criteria decision problems where weight ambiguity or conflicting objectives are central concerns, including network selection, cloud/edge resource allocation, immersive media environments, and interactive applications such as online gaming. Future work may explore partial weight information (e.g., in terms of ambiguity sets), richer temporal models with online updates, and distributed multi-agent evaluation scenarios.

Overall, the RAS provides a principled and practical tool for robust quality assessment in high-dimensional service environments subject to structural uncertainty, a challenge that increasingly arises in modern digital ecosystems characterized by complex tradeoffs, heterogeneous users, and dynamic performance constraints.

References

[1] T. A. Weber, "Relatively robust decisions," *Theory Decis.*, vol. 94, no. 1, pp. 35–62, Jan. 2023.

[2] T. A. Weber, "Relatively robust multicriteria decisions," *Manage. Sci.*, forthcoming, 2025. [DOI 10.1287/mnsc.2025.00510]

[3] K. J. Lancaster, "A new approach to consumer theory," *J. Polit. Econ.*, vol. 74, no. 2, pp. 132–157, Apr. 1966.

[4] K. S. Moorthy, "Market segmentation, self-selection, and product line design," *Mark. Sci.*, vol. 3, no. 4, pp. 288–307, Nov. 1984.

[5] T. A. Weber, "Optimal multiattribute screening," *Ural Math. J.*, vol. 2, no. 2, pp. 87–107, Dec. 2016.

[6] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives*, Cambridge University Press, Cambridge, UK, 1993.

[7] M. Ehrgott, *Multicriteria Optimization*, 2nd ed., Springer, Berlin, 1993.

[8] H. Toral-Cruz, A.-S. Khan Pathan, and J. C. Ramírez Pacheco, "Accurate modeling of VoIP traffic QoS parameters in current and future networks with multifractal and Markov models," *Math. Comput. Model.*, vol. 57, nos. 11–12, pp. 2832–2845, Jun. 2013.

[9] M. Ghasemi, P. Kanuparthy, A. Mansy, T. Benson, and J. Rexford, "Performance characterization of a commercial video streaming service," *arXiv:1605.04966*, May 2016.

[10] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Mar./Apr. 2010.

[11] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," ITU, Geneva, 1996.

[12] ITU-T Recommendation P.863, "Perceptual Objective Listening Quality Analysis (POLQA)," ITU, Geneva, 2011.

[13] A. Saha, S. K. Pentapati, Z. Shang, R. Pahwa, B. Chen, H. E. Gedik, S. Mishra, and A. C. Bovik, "Perceptual video quality assessment: the journey continues!" *Front. Signal Process.*, vol. 3, Frontiers Media, Lausanne, Switzerland, Jun. 2023.

[14] M. Claypool and K. Claypool, "Latency and player actions in online games," *Commun. ACM*, vol. 49, no. 11, pp. 40–45, Nov. 2006.

[15] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Phil. Trans. R. Soc. B*, vol. 364, no. 1535, pp. 3549–3557, Dec. 2009.

[16] P. Le Callet, S. Möller, and A. Perkis (Eds.), "Qualinet white paper on definitions of quality of experience," Version 1.2, European Network on Quality of Experience in Multimedia Systems, Novi Sad, Mar. 2013.

[17] R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making," *IEEE Trans. Syst., Man, Cybern.*, vol. 18, no. 1, pp. 183–190, Jan./Feb. 1988.

[18] M. Yang, S. Wang, R. N. Calheiros, and F. Yang, "Survey on QoE assessment approach for network service," *IEEE Access*, vol. 6, pp. 48374–48390, Aug. 2018.

[19] D. D. Sleator and R. E. Tarjan, "Amortized efficiency of list update and paging rules," *Commun. ACM*, vol. 28, no. 2, pp. 202–208, Feb. 1985.

[20] A. Goel, A. Meyerson, and T. A. Weber, "Fair welfare maximization," *Econ. Theory*, vol. 41, no. 3, pp. 465–494, Dec. 2009.

[21] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik, and Z. Tu, "Video quality assessment: A comprehensive survey," *arXiv:2412.04508*, Dec. 2024.

[22] J. Bienik, M. Uhrina, L. Ševčík, and A. Holesová, "Impact of packet loss rate on quality of compressed high-resolution videos," *Sensors*, vol. 23, no. 5, art. 2744, Mar. 2023.

[23] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Study of the subjective and objective quality of high motion live streaming videos," *IEEE Trans. Image Process.*, vol. 31, pp. 1027–1041, 2022 (Dec. 2021).

[24] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, Apr. 2010.

[25] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[26] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[27] R. Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *Proc. 2017 IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Cagliari, Italy, pp. 1–2, Jun. 2017.

[28] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: An in-depth look at QoE via better metrics and their relation to MOS," *Qual. User Exp.*, vol. 1, art. 2, Sep. 2016.