AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA NAUKI INŻYNIERYJNO-TECHNICZNE

DYSCYPLINA INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA

ROZPRAWA DOKTORSKA

Czynniki wpływające na jakość postrzeganą przez użytkowników usług wideo - badania z użyciem trafnych ekologicznie standardów oceny

Autor: Mgr Kamil Koniuch

Promotor rozprawy: Dr hab. inż. Lucjan Janowski
Drugi promotor: Prof. Michał Wierzchoń

Praca wykonana: AGH, Wydział Informatyki, Elektroniki i Telekomunikacji

Kraków, 2025

AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

FIELD OF SCIENCE ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE INFORMATION AND COMMUNICATION TECHNOLOGY

DOCTORAL THESIS

Factors Influencing Quality of Experience by Users of Video Services – Research Using Ecologically Valid Assessment Criteria

Author: Mgr Kamil Koniuch

First supervisor: Dr hab. inż. Lucjan Janowski
Second supervisor: Prof. Michał Wierzchoń

Completed in: AGH University of Krakow, The Faculty of Computer Science, Electronics and Telecommunications

Kraków, 2025

This work was possible thanks to funding from the Norwegian Financial Mechanism 2014–2021 under project 2019/34/H/ST6/00599. The project was titled „Towards Better Understanding of Factors Influencing the QoE by More Ecologically-Valid Evaluation Standards", acronym TUFIQoE.

I would like to thank all of my colleagues from the TUFIQoE project. Particularly; to Katrien De Moor for being the kindest soul that I met during this adventure and for all comments and insights that made this dissertation in its current form. To Lucjan Janowski for challenging me to be better and showing me the complex world of data analysis, including providing parts of the code for data analysis in this work. To Gabriela Wielgus, who gathered lots of crucial data in the project. To Rafał Figlus, who wrote the Chrome extension used in chapter 6 of this work. To Michał Wierzchoń, who provided a methodological thoroughness and psychological perspective on the problem. To Mikołaj Leszczuk, who gathered the data for the experiment described in chapter 5. To Natalia Jakubiec and Anna Wilusz, who provided tremendous help as students involved in the project. To Mateusz Zduński, who helped collect the data for Chapter 6.

Moreover, I would like to thank all the scientists, collaborators, and co-authors I met along the way. Especially; to Pablo Pérez, for inspiring discussions and the opportunity to apply my model in future works. To Jesús Gutiérrez Sánchez, Marta Orduna Cortillas, and Carlos Cortés Sánchez for their appreciation of my ideas and for introducing me to the world of XR communication. To Narciso García, for providing a unique perspective on science and life. To Sabina Baraković and Jasmina Baraković Husić for the time and effort spent on understanding and improving my theoretical model. To Margaret Pinson, for a warm welcome in the Video Quality Expert Group community, and her invaluable knowledge. To Borys Paulewicz for showing me the beauty of causal inference.

Finally, I would like to thank my long-time friends and family. To my mother, who pushed me against my fear to start this project. To Jacek Dubikowski, Karolina Łagodzka-Dubikowska, Mikołaj Thal, Alicja Kępa, Julia Pudo, Piotr Tomal, Bartek Kubiczek, Julia Kubiczek, Wojciech Szumański, who have been with me even in my darkest hour, unconditionally. To my fiancée, Karolina Machalska, who is always there for me and at this point probably understands this thesis as good as I do.

Abstract

The use of subjective studies in telecommunications has a tradition nearly as long as the field itself. From early investigations of telegraph circuits and telephone transmission, through the adaptation of the Mean Opinion Score (MOS), to the E-model and learned quality metrics, user judgments have continuously guided the advancement of communication systems. In this sense, Quality of Experience (QoE) is not a new perspective, but rather a modern articulation of historical principles. With the rise of adaptive video streaming and perceptually trained quality metrics, QoE is now an important part of telecommunications advancement.

Yet a persistent gap between theoretical richness and practical application of QoE remains a crucial shortcoming. Theoretical models of QoE describe complex causal structures with multiple Influential Factors (IFs) crossing human, system, and context domains. By contrast, standardized measurements such as ACR or JND reduce this complexity to ensure reliability and comparability, leaving most IFs unmeasured. This contrast raises fundamental questions about the external validity of current QoE tests and the extent to which they capture the lived experience of video service users.

This dissertation addresses this challenge through a three-part research design. **First**, it introduces and validates a memory recall-based questionnaire for the exploration of Influential Factors in ecologically valid contexts. Applied across different substudies (VoD, live streaming, video chat), the method was used to capture distinct cognitive, affective, and behavioral dimensions of QoE. **Second**, based on these findings, the thesis proposes a parsimonious theoretical model of video QoE in the form of a path diagram linking QoS, content, perceived multimedia signal quality, delight/annoyance, and behavioral outcomes. **Third**, two experimental studies tested the most important Influential Factor derived from the exploratory stage: one examined whether emotionally evocative content biases ACR ratings, while the other investigated whether social presence alters evaluations when watching together versus alone.

The results demonstrated that the recall questionnaire provided a consistent and reliable way to identify and rank Influential Factors across different service types. After the memory reconstruction procedure, participants gave more stable responses and placed greater emphasis on factors such as content quality and social presence, while technical aspects like device or network conditions played a smaller role. The proposed theoretical model integrated these findings into a path structure, providing distinct Quality of Experience layers. The two experimental studies showed that, despite being highlighted as important in the exploratory stage, emotionally evocative content and social presence did not alter ACR ratings. Instead, ACR scores were almost entirely determined by quality manipulation.

The main contributions of this work are therefore fourfold. (1) It establishes memory recall as a valid and practical method for identifying IFs in naturalistic contexts. (2) It develops a theoretical model that

clarifies how technical and experiential factors interact within video QoE. (3) It provides empirical evidence that ACR, while reliable, has clear boundaries: it captures perceived signal quality but not the broader experiential dimensions of content or social context. (4) It formulates concrete recommendations for academia and industry on when and how to complement standardized tests and quality metrics, and outlines what is needed for domain advancement.

Beyond its empirical findings, the work has influenced international discussions within the Video Quality Experts Group (VQEG), where new layered definitions of QoE are being elaborated. In sum, this dissertation strengthens the methodological foundations of QoE research and provides guidance for designing, interpreting, and applying video quality tests in the evolving landscape of telecommunication services.

Streszczenie

Wykorzystanie badań subiektywnych w telekomunikacji ma tradycję niemal tak długą jak sama ta dziedzina. Od wczesnych badań nad obwodami telegraficznymi i transmisją telefoniczną, poprzez adaptację Mean Opinion Score (MOS), aż po model E i wyuczone metryki jakości – oceny użytkowników nieustannie kierowały rozwojem systemów komunikacyjnych. W tym sensie Quality of Experience (QoE) nie jest nową perspektywą, lecz współczesnym rozwinięciem zasad historycznych. Wraz z rozwojem adaptacyjnego strumieniowania wideo i percepcyjnie trenowanych metryk jakości, QoE stało się dziś istotnym elementem postępu w telekomunikacji.

Jednak utrzymująca się luka między obszernym opisem teoretycznym a praktycznym zastosowaniem QoE pozostaje kluczowym brakiem. Teoretyczne modele QoE opisują złożone struktury przyczynowe z wieloma czynnikami wpływającymi (Influential Factors, IFs), obejmującymi domeny ludzką, systemową i kontekstową. Natomiast znormalizowane pomiary, takie jak ACR czy JND, redukują tę złożoność, aby zapewnić rzetelność i porównywalność, pozostawiając większość IFs niezmierzonych. Ten kontrast rodzi fundamentalne pytania o zewnętrzną trafność obecnych testów QoE oraz o to, w jakim stopniu oddają one rzeczywiste doświadczenie użytkowników usług wideo.

Niniejsza rozprawa podejmuje ten problem poprzez projekt badawczy złożony z trzech części. **Po pierwsze**, wprowadza i waliduje kwestionariusz oparty na pamięciowym odtwarzaniu doświadczeń do eksploracji czynników wpływających w ekologicznie trafnych kontekstach. Zastosowany w różnych podbadaniach (VoD, transmisja na żywo, wideorozmowa) pozwolił uchwycić odrębne wymiary poznawcze, afektywne i behawioralne QoE. **Po drugie**, na podstawie tych wyników rozprawa proponuje parsymoniczny model teoretyczny wideo QoE w formie diagramu ścieżkowego, łączącego QoS, treść, postrzeganą jakość sygnału multimedialnego, doznania przyjemności/irytacji oraz rezultaty behawioralne. **Po trzecie**, dwa badania eksperymentalne przetestowały najważniejsze czynniki wyłonione w etapie eksploracyjnym: jedno sprawdzało, czy emocjonalnie poruszające treści zniekształcają oceny ACR, a drugie badało, czy obecność społeczna wpływa na oceny podczas wspólnego oglądania w porównaniu z oglądaniem w pojedynkę.

Wyniki pokazały, że kwestionariusz pamięciowy stanowi spójną i rzetelną metodę identyfikacji oraz porządkowania czynników wpływających w różnych typach usług. Po procedurze odtwarzania pamięciowego uczestnicy udzielali stabilniejszych odpowiedzi i większą wagę przypisywali takim czynnikom jak treść i obecność społeczna, podczas gdy aspekty techniczne – jak urządzenie czy jakość sieci – odgrywały mniejszą rolę. Zaproponowany model teoretyczny zintegrował te wyniki w strukturę ścieżkową, wyraźnie wyróżniając warstwy QoE. Dwa badania eksperymentalne pokazały natomiast, że mimo iż treść i obecność

społeczna zostały uznane za istotne w badaniu eksploracyjnym, nie wpłynęły one na oceny ACR. Oceny te były niemal całkowicie determinowane przez manipulację jakością sygnału.

Główne wkłady niniejszej pracy mają zatem cztery aspekty. (1) Prezentuje ona pamięciowe odtwarzanie doświadczeń jako ważną i praktyczną metodę identyfikacji czynników wpływających w naturalistycznych kontekstach. (2) Rozwija model teoretyczny, który wyjaśnia, jak czynniki techniczne i doświadczeniowe oddziałują w ramach wideo QoE. (3) Dostarcza empirycznych dowodów, że ACR, mimo rzetelności, ma wyraźne granice: obejmuje postrzeganą jakość sygnału, lecz nie szersze wymiary doświadczeniowe związane z treścią czy kontekstem społecznym. (4) Formułuje konkretne rekomendacje dla środowiska akademickiego i przemysłu dotyczące tego, kiedy i jak uzupełniać testy znormalizowane i metryki jakości oraz wskazuje, co jest potrzebne dla dalszego rozwoju dziedziny.

Poza wynikami empirycznymi, praca ta wpłynęła także na międzynarodowe dyskusje w ramach Video Quality Experts Group (VQEG), gdzie opracowywane są nowe warstwowe definicje QoE. Podsumowując, niniejsza rozprawa wzmacnia metodologiczne fundamenty badań QoE i dostarcza wskazówek dotyczących projektowania, interpretacji i stosowania testów jakości wideo w zmieniającym się krajobrazie usług telekomunikacyjnych.

Contents

List of Figures

Chapter 1

Introduction and Theoretical Background

In the initial phase of working on this thesis, I had a superstition that QoE is a relatively new domain, which was established to enrich the technical Quality of Service perspective with human perception. Thanks to Pablo Pérez's work [144], I found this superstition to be false. Due to that, I would like to start this introduction with the description of the work of giants on whose shoulders today's telecommunication research stands. While some of the discussed papers are almost 100 years old, they provide a blueprint for how subjective studies have been used in the advancement of telecommunications.

Later, I will show how this original approach of incorporating subjective studies into telecommunication development is still a scientific frontier. With this background, I will describe the theoretical frame of QoE, its standardized methods, models, and applications, highlighting the gap between theory and practice. Then, I will describe the state-of-the-art work that aims to fill this gap. With this overview, I will present the motivation for the original research presented in the rest of this thesis, together with hypotheses and objectives.

1.1 Subjective Tests in Telecommunications: A Historical Perspective

The use of subjective quality tests has a history almost as long as telecommunication itself. In the context of telegraphy, Herman [68] studied the relation between signal distortion and perceived transmission quality in circuits using the American Morse Code. He examined three types of distortion: bias, characteristic, and fortuitous, and analyzed how they affected both the accuracy of message reception and the opinions of operators regarding circuit usability. To carry out the study, he built an artificial telegraph circuit where controlled amounts of distortion could be introduced. Four operators transmitted messages with a semi-automatic key at a rate of 13.5 dots per second, while receivers copied Hungarian text unfamiliar to them to reduce reliance on context. Distortion levels were measured with dedicated apparatus, and after each message, the operators reported their judgment of circuit quality. The results showed a gap between technical accuracy and subjective evaluation. Operators often considered a circuit unsatisfactory before errors became frequent, pointing to the extra effort and fatigue caused by distorted signals. This result shows that subjective assessment might differ from objective quality even in the context of telegraphy. To highlight the historical context, it is worth noting that this publication predates the first working transistor by 18 years [157].

Herman's investigation into distortion effects reflected a wider concern with the reliability of telegraph systems. Already in 1884, Rex [155] discussed this problem from a legal perspective, showing that quality issues were tied to responsibility and trust in telegraph services. In this sense, subjective studies of telegraph quality can be seen as part of a broader response to the challenge of reliability. This explains a strong motivation for the incorporation of subjective studies into the development of the telegraph grid.

With the development of telephony, Bell Labs continued the subjective test tradition, conducting several studies to model perceived quality. Martin [127], for example, proposed a rating method based on the rate of repetitions in natural telephone conversations. Instead of relying only on physical circuit characteristics such as attenuation or frequency response, he emphasized the joint action of talker, listener, and circuit. The method showed that users adjust loudness, articulation, and effort depending on circuit conditions, and that these human factors strongly influence conversational success. By using repetition rate as the main indicator, Bell Labs introduced a user-centered measure of transmission performance that linked technical parameters with perceived communication quality.

Further development of subjective tests brought systematic scale-based quality assessments. For example, Coolidge and Reier [35] examined how received speech volume affects user satisfaction. Through controlled laboratory experiments, they mapped listener judgments into categories such as Good, Fair, or Poor, and showed that preferences depend not only on technical volume levels but also on the effort required from the listener. Their work highlighted that acceptable transmission quality is best understood statistically, as a distribution of opinions across users, and established "grade of service" objectives that explicitly incorporated human factors into network design.

With the introduction of standardized rating procedures, quality judgments were placed on a fixed five-point scale ranging from Excellent to Bad. By attaching numerical labels to these categories, researchers were able to compute the Mean Opinion Score (MOS), defined as the arithmetic mean of all individual ratings. This transformation of subjective opinions into a single comparable index provided a robust way to compare results across laboratories and countries. The adoption of MOS by the International Telegraph and Telephone Consultative Committee (CCITT) in the 1956 [75] marked its recognition as the official standard for subjective telephone quality assessment. The document that is now in force, describing MOS, was provided by the International Telecommunication Union Standardization Sector in [86] and defines MoS with equation (1.1).

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{1.1}$$

where R_n are the individual ratings for a given stimulus by N subjects.

With the foundation of MOS established, subjective tests could move from description to prediction further. Cavanaugh, Hatch, and Sullivan [28] extended MOS-based studies by introducing a generalized transmission-rating scale R. This scale was anchored at reference conditions and allowed results from different tests to be combined, reducing variability due to subject groups or test design. Subjective opinion remained expressed through MOS, but now was mapped probabilistically from the transmission rating, linking technical parameters directly to expected proportions of calls judged "good or better" or "poor or worse." Their analysis drew on seven large-scale tests conducted between 1965 and 1972, including more than 8000

ratings from Bell Labs employees under controlled variations of loudness loss, circuit noise, and echo-path delay. The resulting models included separate expressions for loss–noise, echo, and their combined effect. For example, the transmission rating for loss and noise was modeled as

$$R_{LN} = 147.76 - 2.257\sqrt{(L-7.2)^2+1} - 2.009\sqrt{N'} + 0.02037\,(L \cdot N'), \tag{1.2}$$

where L is the connection loudness loss (dB), N is the circuit noise (dBrnC), and $N' = \sqrt{N^2 + 27.37^2}$ represents the combined noise power.

This framework became a cornerstone of network planning, allowing engineers to quantify how technical impairments would be perceived by users and to set grade-of-service objectives.

This brief historical overview illustrates the foundations of using subjective studies for the development of telecommunication systems. The general approach was to identify factors that influence the acceptance level, manipulate objective signal quality, collect subjective judgments, and derive models that predict user satisfaction. While methodologies have evolved, the principle has stayed the same. The next section will outline how subjective studies are comparably adapted to modern systems and services.

1.2 Subjective Tests in Modern Systems and Services

Modern video services typically rely on HTTP adaptive streaming protocols, which are standardized, for example, in ISO/IEC 23009 - 1 [2]. In this method, the client dynamically selects media segments of appropriate quality based on current network conditions and device capabilities, enabling uninterrupted playback despite varying bandwidth. To achieve this, Content and Application Providers (CAPs) like Netflix use encode optimization [6] to derive a content-specific encoding (bitrate) ladder. Encoding laddering is a technique used in video streaming to generate multiple versions of the same content at different resolutions and bitrates. This allows adaptive bitrate streaming systems to dynamically switch between versions based on the user's network conditions and device capabilities, ensuring an optimal balance between video quality and playback smoothness. The set of these bitrate-resolution pairs is referred to as the encoding ladder and is typically designed to maximize user satisfaction while minimizing encoding and delivery costs. As there are many strategies that can be applied to build such a ladder, subjective studies provide insights that are used in decision-making, e.g. [94].

Moreover, during the development of video codecs, various design decisions must be made regarding compression tools, trade-offs between bitrate and quality, and optimization strategies. While objective quality metrics such as PSNR or SSIM are commonly used to guide this process, they may not fully capture human perception. Therefore, subjective studies, in which users rate the perceived video quality under controlled conditions, play a crucial role in validating and comparing codec performance. Bodies like MPEG or JPEG use subjective scores as benchmarks in codec development [193, 87]. Such evaluations are particularly important when deciding which codec version delivers superior user satisfaction under realistic viewing conditions, as objective indicators alone may underestimate perceptual gains. For example, in an 8K comparison between VVC and HEVC codecs, researchers demonstrated that while objective metrics

suggested a bitrate saving of only 26–35% (depending on metric), subjective tests revealed a substantially higher gain of 41% for the same perceived visual quality [20].

In terms of modeling quality, the direct link to historical research I described above can be found in modern quality models. Most prominently, ITU-T Recommendation G.107 [84] formalized the *E-model*, which extends the transmission-rating scale into a comprehensive planning framework. The E-model preserves the basic principle of mapping technical impairments to a scalar rating R, but expands the formulation to account for a broader set of factors, including simultaneous impairments (I_s), delay impairments (I_d), equipment impairments (I_e), and an advantage factor (A) compensating for contextual benefits such as mobility. Its general equation is expressed as:

$$R = R_0 - I_s - I_d - I_{e,\text{eff}} + A, \tag{1.3}$$

where R_0 represents the basic signal-to-noise ratio of the connection.

Like the Bell Labs models, the E-model outputs a transmission rating that can be transformed into an estimated MOS and user satisfaction categories (e.g., percentage of calls rated "good or better"). Thus, while originating in controlled laboratory tests of loss, noise, and echo, the concept evolved into a standardized and modular tool for end-to-end quality planning in modern networks.

In the context of video, modern video quality metrics directly use subjective data for development. Those metrics aim for automated predictions of quality, which helps to make codec development and encoding laddering faster, cheaper, and better. Subjective tests were crucial for the development of relevant quality metrics. Those metrics use not only objective data, such as the peak signal-to-noise ratio (PSNR), but also model predictions based on subjective studies. They combine knowledge from the human visual system, signal data, and subjective scores. For example, one of the most recognizable video quality metrics, VMAF [119], not only utilizes information about physical changes in the video signal, but also incorporates scores from subjective tests. It is designed to predict subjective video quality based on a combination of multiple objective metrics. VMAF integrates various quality factors, such as detail loss, motion, and spatial-temporal complexities. Finally, VMAF uses data from subjective tests based on the subjective test to model the predictions. Similarly, the FovVideoVDP metric [125] uses spatiotemporal contrast sensitivity, peripheral vision sensitivity, and the physical specifications of the display to predict quality. Subjective data from human observers was crucial for tuning this model, ensuring that the metric's predictions align closely with human perception.

With these metrics, global challenges such as infrastructure scaling and network planning can be addressed using automated predictions of user satisfaction [161]. By modeling quality based on perceptual metrics validated through subjective studies, network providers [59] and service platforms [168] can anticipate user experience without the need for constant human evaluation. This enables large-scale deployment of adaptive strategies such as dynamic optimization [93], QoE-aware congestion control [197], or resource allocation—guided by estimated satisfaction [133]. As a result, service quality can be optimized in real time, improving user satisfaction while maintaining cost-efficiency and operational stability in high-traffic environments.

Despite the shift from telegraph circuits to adaptive streaming and learned metrics, the core approach remains the same: collect human judgments, relate them to controllable technical variables, and use these relations to guide system design. From this perspective, optimizing for human users is not an additional step but the main objective, since the value of a system is realized only through the quality users perceive.

1.3 Theoretical Perspectives on Quality of Experience

As I argue above, subjective studies are an inseparable part of telecommunication advancement. Yet, until recently, there was no common consensus on the definition and the scope of this part of telecommunication. For example, in the US National Telecommunications and Information Administration Report from 1985 [69], it is stated that „the end-user's perception of the entire process of call setup, conversation (which may be thought of as the information transfer phase), and call disengagement" is the definition of the Quality of Service. On the other hand, in the 2007 version of ITU-T Recommendation P.10/G.100 [76], "the overall acceptability of an application or service, as perceived subjectively by the end-user" was defined as Quality of Experience, with the notes that it includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.) and that overall acceptability may be influenced by user expectations and context. At that stage, the distinction between QoS and QoE was overlapping and inconsistently applied. Although methodologies for subjective testing existed, a coherent theoretical framework to connect them was still missing.

The biggest advance on this problem came with the COST action Qualinet in 2013 [114]. With this project, subjective tests and their legacy gained a well-recognized name: Quality of Experience (QoE) and definition. QoE is currently defined as „the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state." Currently, this is recognized as a working definition by the International Telecommunication Union – Telecommunication Standardization Sector [80]. One of the key outcomes of this project was the publication of the „Quality of Experience Advanced Concepts, Applications, and Methods" book [150], which became a main guideline for QoE research and a starting point for new phd students.

In this book's opening, the authors justify the need for QoE research with the fast development of information and communication technology. This development elevates demand for better and more widely adopted quality control. As they argue, the term QoE was established as a "counter-balance" to Quality of Service (QoS), which is only focused on technical parameters of the network. They advocate for optimizing systems and services for QoE, not for QoS, as this is a better predictor of high acceptance.

To establish QoE as a research domain, authors not only needed the definition but also a clear distinction from other closely related concepts: User Experience (UX) and QoS. In the 3rd chapter of this book [189] authors provided the following distinction between QoE and UX. They define UX based on a survey study of UX experts [113] as „dynamic, context-dependent and subjective, stemming from a broad range of potential benefits users may derive from a product." This definition highlights the dynamic and subjective character of the experience. What is more important is that this definition puts the human in the center, and treats

„product" very broadly, while QoE is always linked to the QoS. This is because QoE has its origin in telecommunications, while UX is in Human-Computer Interaction.

Furthermore, the distinction between QoE and QoS was provided in chapter 6 [181] of the above-mentioned book. For QoS definition, authors used ITU-T recommendation E.800 [77]. According to this recommendation, QoS is: „The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service." The most important distinction of QoS from QoE is that QoS is characteristic of the system, while QoE is the emotional reaction of the user to the system or service.

Building on the above-mentioned definition and distinctions, researchers developed the theoretical foundations of the domain. One of the key concepts in this theoretical advancement was the introduction and taxonomy of Influential Factors (IFs) of QoE.

In "Qualinet White Paper on Definitions of Quality of Experience," Influential Factors are defined as *": Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user"*. This broad spectrum of factors was then systematized in [53]. Originally, factors were organized as follows:

Table 1.1: Classification of Influence Factors based on [53]

Category	**Subcategories**
Human IFs	Low-level Processing and Human IFs
	Higher-level Processing and Human IFs
System IFs	Content-related System IFs
	Media-related System IFs
	Network-related System IFs
	Device-related System IFs
Context IFs	Physical context
	Temporal context
	Social context
	Economic context
	Task context
	Technical and information context

As Table 1.1 shows, there is a plethora of factors that can influence QoE. They consist of latent variables and psychological concepts. Thus, having a guiding theory and a clear taxonomy is essential to measure and control Influential Factors.

Existing theoretical models of QoE vary considerably in complexity and practical applicability, e.g. [52, 151, 131, 154, 159, 163, 53, 44, 73]. Thus, I will describe them shortly to present their diversity.

Brunnstrom et al. [52], building on Jekosch and Raake's work, present a comprehensive model emphasizing Quality Perception. It integrates human and contextual IFs through over 10 conceptual units interconnected by more than 15 directional links, highlighting the intricate interactions between cognitive processes

and system properties. However, the high abstraction level and complex interrelations pose challenges for empirical applications and operational definitions.

The model presented in [151] similarly explores the cognitive mechanisms behind perception. It encompasses approximately nine conceptual units and over 12 directed connections, demonstrating the significant influence of context and internal states on sensory input and anticipatory processes. Despite its descriptive richness, the model lacks clear guidance for the operationalization necessary in empirical studies.

Similarly, authors in [131] propose a model that links QoE to the broader context of content creation and presentation, featuring around eight conceptual units and more than ten connections. This approach effectively bridges creator intentions with user experiences, but remains limited in providing clear methodologies for quantifying specific IFs.

In the paper [154] authors advocate for a multi-perspective QoE model, integrating technical, eudaimonic, and behavioral dimensions. The model, comprising roughly six central units and numerous bidirectional interactions, offers a rich conceptual understanding but struggles to translate abstract factors into measurable empirical constructs.

Robitza et al. [159] developed one of the most elaborate QoE frameworks, combining quality formation, perception, affective states, behavior, and memory. Featuring more than 20 conceptual units and over 30 directional links, it extensively employs latent variables. Although theoretically profound, the complexity and reliance on unobservable constructs significantly limit its applicability in practical research settings.

Schmitt et al.'s model [163] specifically targets multi-party video conferencing, structuring IFs across distinct categories such as system, context, user, and behavior. With over 25 identifiable units and detailed interconnections, it is notably suitable for empirical research. Nevertheless, ambiguity in user-related factors and unclear causal relationships between contextual and system variables restrict its effectiveness for dynamic and longitudinal analyses.

The taxonomy proposed by Egger et al. [44] categorizes QoE into Influence Factors, Interaction Performance, and Perceived Quality. With approximately 15 clearly labeled components, this approach offers simplicity and clarity. Despite its clear structure, its descriptive nature and absence of causal mechanisms limit predictive and hypothesis-testing capabilities.

Lastly, Husic et al. [73] present a quantitative model operationalizing QoE through measurable system, human, and context factors linked to perceptual dimensions such as perceived quality and ease of use. While perceptual aspects remain latent and context somewhat static, this approach provides a statistically grounded, empirically validated framework that facilitates predictive modeling, especially relevant to video streaming contexts.

Overall, those theoretical models are very elaborate and highlight the complexity of a multidimensional perspective on QoE. Nonetheless, researchers in the QoE domain developed multiple research paradigms and standards that aim for control measurement of QoE. Below, I will provide an overview of standardized methods, models, and applications in the context of video QoE studies, as the scope of this thesis focuses on video services.

1.4 Measuring Video Quality of Experience

In this section, I will provide a general overview of methods that are standardized and used not only in academia but also in many Content and Application Providers and Communication Services Provider companies. It is worth mentioning that those methods are constantly developing and being evaluated. One of the key bodies that carries this work is the Video Quality Expert Group [183], which joins the effort of companies and universities to provide new recommendations and domain growth. In this overview, I will focus on methods that are not only established but also subject to ongoing refinement and adaptation. While several of them are noteworthy for the innovations they introduced, I will also indicate which have become the dominant sources of subjective data for QoE model development.

1.4.1 Subjective Assessment Methods for Video Quality in Recognition Tasks

Current QoE methods not only study the relative quality but also its utility, dependent on quality. This idea can already be traced back to early telegraph experiments, where Herman [68] evaluated transmission quality not only through subjective judgments of operators but also through their task performance, measured as the accuracy of message reception. In modern terms, ITU-T Recommendation P.912 [1] summarizes methods that allow for measurement at which quality level the task is no longer possible to conduct. These include:

1. **Multiple choice method** – after viewing a video sequence, the subject selects the recognized target from a predefined list of possible answers. This method reduces ambiguity and is suitable for human, object, and alphanumeric recognition tasks.

2. **Single answer method** – the subject provides a direct response without a list of options, for example, by typing the alphanumeric characters seen in the video or answering Yes/No to the presence of an object. Responses are then evaluated as correct or incorrect.

3. **Timed task method** – the subject indicates the moment they recognize the target (e.g., by pressing a button). This allows the experimenter to measure both the accuracy of recognition and the time required to make the decision, reflecting real-world constraints.

4. **Real-time vs. viewer-controlled viewing** – depending on the application, recognition may be assessed in real-time (without pausing or replaying, as in surveillance scenarios) or in viewer-controlled conditions (allowing pausing and replaying, as in analytical tasks).

These recognition-oriented methods are especially useful in application domains where the utility of video is measured by successful task completion rather than aesthetic impression. Typical scenarios include video surveillance (e.g., license plate or object recognition), telemedicine and remote diagnostics (supporting accurate medical decisions), fire safety (early detection of flames or smoke), and driver assistance systems such as backup cameras (ensuring correct maneuvering). In such contexts, the strength of P.912 methods lies in directly linking video quality to operational effectiveness, even though their applicability is often limited to the specific datasets and conditions under which the tests are conducted [115].

While recognition-based methods are important for task-specific applications, the majority of video quality studies focus on quality in general, not on its utility. Datasets mostly rely on descriptive rating or just noticeable difference paradigms. As summarized in Zheng et al.'s comprehensive survey on video quality assessment [200], subjective studies most commonly use a discrete ACR scale or continuous scales to collect Mean Opinion Scores. In the following sections, I will describe methods for stimulus comparison and scales used in QoE research, which remain the dominant source of subjective datasets.

1.4.2 Stimuli Comparisons

Paired comparison

The method of paired comparisons introduced in [23] provides a statistical framework for analyzing preferences when items are judged two at a time. Each comparison is modeled as a probabilistic choice, where the likelihood of one item being preferred over another depends on underlying latent "treatment ratings" assigned to each item. Using a maximum-likelihood approach, these latent ratings can be estimated from observed outcomes across all pairs, allowing items to be placed on a common interval scale. The model also supports hypothesis testing (e.g., equality of treatments) and can handle incomplete block designs, making it a powerful tool for subjective evaluations where absolute scales are difficult to define.

One of the examples of the paired comparison method to Quality of Experience (QoE) was presented by Chen et al. [30], who use a framework for multimedia content evaluation based on the Bradley–Terry–Luce model. This approach replaces absolute rating scales with relative judgments between pairs of stimuli, thereby reducing scale bias and cultural differences in interpretation. By collecting binary preferences and transforming them into interval-scale estimates, the authors demonstrated that paired comparison can provide more robust and comparable measures of subjective quality, particularly in large-scale or crowdsourced QoE studies.

Pair comparison is part of the ITU-T P.910 standard [5]. This standard recognizes pair comparison as a variant of the comparison category rating (CCR) or double stimulus comparison scale (DSCS) method. In the CCR/DSCS method, stimuli are presented in pairs, with two versions of the same sequence shown in randomized order. Observers are then asked to rate the second stimulus relative to the first using a seven-point comparison scale, ranging from “much worse” (–3) through “the same” (0) to “much better” (+3). Because the presentation order is randomized, this method avoids the systematic bias present in degradation category rating (DCR), where the reference always comes first. CCR/DSCS is particularly useful when comparing two impairments of nearly equal quality, providing sensitivity in cases where absolute category ratings may be less discriminative.

In ITU-T P.910, outcomes from CCR/DSCS and other double-stimulus methods are typically reported as Differential Mean Opinion Scores (DMOS). DMOS captures the perceived change in quality between a reference and a processed sequence, whereas MOS reflects an absolute quality level. By focusing on degradation relative to a reference, DMOS is particularly suitable for pairwise and double-stimulus methodologies where relative judgments are central. The Recommendation defines DMOS through differential viewer

scores (DV) calculated per subject as:

$$DV(\text{PVS}) = V(\text{PVS}) - V(\text{REF}) + 5, \tag{1.4}$$

where $V(\text{PVS})$ is the viewer's score for the processed video sequence and $V(\text{REF})$ the score for the hidden reference. The constant $+5$ ensures that DV values remain aligned with the original 1–5 ACR quality scale. DMOS is then obtained as the average across all N subjects:

$$DMOS(\text{PVS}) = \frac{1}{N}\sum_{i=1}^{N} DV_i(\text{PVS}). \tag{1.5}$$

Just noticeable difference

Building on the idea of relative judgments in subjective quality assessment, another influential method is the just noticeable difference (JND) paradigm. JND is a psychophysical unit that describes the smallest change in quality that people can reliably detect. One of the first proposals to apply JND to video quality came from Watson, in the context of IEEE and VQEG discussions on how to establish an absolute scale for video assessment [188]. Instead of asking viewers to rate videos, he suggested pairwise comparisons where observers simply chose which sequence looked more impaired. Using Thurstone's comparative judgment framework and adaptive methods, these binary decisions were converted into a perceptual scale expressed in JND units. The approach also introduced "blends," where original and distorted videos were mixed to precisely control impairment strength. Pilot experiments showed that JND values aligned with categorical scores but provided a more consistent, context-independent measure of quality.

In recent years, JND has become an important direction in QoE research, moving beyond its original role as a psychophysical threshold. New studies use JND information to train machine learning models that can predict perceived quality on a continuous scale. For example, in [11], authors showed that combining first-JND measurements with deep learning and models of the human visual system allows quality predictions that generalize well across different types and levels of distortion. This illustrates how JND has developed from a basic detection concept into a key component of modern QoE modeling.

1.4.3 Scales for Quality Assessment

One of the dominant tools in QoE assessment is the Absolute Category Rating (ACR) scale. As I showed at the beginning of this chapter, this method was already a common practice in telecommunication studies in the 1950s [75]. The ACR scale is favored in QoE studies for its simplicity and effectiveness in assessing user perceptions of media quality. It offers participants a straightforward method to rate their experience, typically on a scale of 1 to 5, with higher scores indicating better quality. The numbers have the following labels: bad, poor, fair, good, and excellent. This approach reduces complexity and concentrates on the perceptual aspect of quality evaluation. In effect, ACR tests generate reliable and repeatable results [147].

Continuous versions of the scales, such as SSCQE and SAMVIQ were proposed in ITU-T recommendations P.913[3] and BT-500 [27]. These methods are designed for evaluating video across a wide range of resolutions, from low to high-definition television. Unlike discrete-scale methods, those scales use a continuous quality scale from 0 to 100, annotated with five descriptors (Excellent, Good, Fair, Poor, Bad). Test

subjects view and freely navigate among multiple versions of the same video sequence, including a hidden reference, and assign quality ratings to each.

Although continuous scales propose finer discrimination between quality levels, it is criticized for introducing unnecessary cognitive load on raters [146]. The use of a 0–100 continuous scale often leads participants to cluster their responses around prototypical values such as multiples of five, which undermines the intended granularity of the method. Moreover, such wide scales may increase uncertainty and promote careless decisions. Thus, simpler scales like the five-level ACR are equally effective for capturing QoE.

There is also an ongoing effort to propose a QoE scale that will capture not only the perceived level of impairments of video, but also users' response to this impairment. A good example of this work are studies using scales for acceptance and annoyance (e.g., [98, 170, 90]). While traditionally these scales were applied separately, there are advantages in combining them within a single AccAnn framework [116]. In this work, researchers introduced a three-level, single-step methodology that simultaneously measures acceptability and annoyance, reducing cognitive load and testing time while maintaining equivalence with the classical two-step approach.

Nonetheless, while these methods represent an important step forward for the QoE domain, the ACR scale and JND remain the dominant methods for quality metrics development.

1.5 Metrics predicting quality using subjective quality rating

As I described at the end of the first section of this introduction, the goal of quality studies is to derive a model that will serve as a predictive metric for quality. Objective approaches to video quality assessment are generally divided into three categories: full-reference, reduced-reference, and no-reference methods [172].

Full-reference methods rely on direct comparison between an original and a distorted signal, assuming complete access to the source. Reduced-reference methods use only partial information from the original, aiming to balance accuracy with lower overhead. No-reference methods operate without any access to the source, relying instead on statistical or perceptual models to estimate quality from the distorted video alone.

In the context of video quality studies, two objective signal-based methods were the cornerstone for the development of full-reference metrics. These are the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM).

PSNR is derived from the mean squared error between a reference and a distorted image, and has long been used because of its simplicity and strong mathematical grounding. However, it often fails to capture perceptual aspects of quality, since different types of distortions can lead to the same PSNR value. By contrast, SSIM was designed to align more closely with the human visual system by decomposing image similarity into luminance, contrast, and structural components.

Research comparing both measures [71] shows that PSNR and SSIM are mathematically related and often produce consistent trends across degradations. Still, their sensitivity differs: PSNR performs better in detecting additive Gaussian noise, while SSIM is more effective in capturing structural degradations.

These and similar metrics are used in combination with scores from subjective tests to build models that predict real users' perception of quality. One of the most recognizable examples of such a metric is Video Multi-Method Assessment Fusion (VMAF). VMAF was developed by Netflix as a full-reference perceptual

metric designed to better approximate human judgments of quality than traditional signal-based measures. Its construction followed a fusion-based approach: instead of relying on a single indicator such as PSNR or SSIM, VMAF integrates multiple quality-related features and learns how to weight them against subjective opinion scores. Specifically, it combines measures such as anti-noise signal-to-noise ratio (AN-SNR), detail loss measure (DLM), visual information fidelity (VIF), and a temporal feature based on mean co-located pixel differences (MCPD), capturing both spatial fidelity and motion-related distortions. These features are then fused through a supervised machine learning model trained on large-scale datasets of subjective scores. The output is a continuous score ranging from 0 to 100, calibrated to the Absolute Category Rating (ACR) scale, where lower values reflect "bad" or "poor" quality and higher values align with "good" or "excellent." In this way, VMAF was explicitly built to bridge the gap between objective metrics and perceptual reality, enabling scalable, automated assessment while preserving strong correlation with human opinion scores [120, 152, 121].

A similar principle underlies the FovVideoVDP metric [125], which is also trained and validated against subjective tests. The way it connects objective modeling with subjective data, however, differs from VMAF. VMAF combines several hand-crafted metrics and learns their weights from large-scale ACR opinion scores. FovVideoVDP, in contrast, is built on perceptual models of the human visual system, including spatio-temporal contrast sensitivity, cortical magnification, and masking. It is calibrated with pairwise comparison experiments, which produce a perceptual scale expressed in Just-Objectionable-Difference (JOD) units, a variation of the JND concept anchored to the reference video. In short, both metrics rely on subjective data, but VMAF is grounded in signal-based features linked to ACR scores, while FovVideoVDP is derived from vision science models tuned to JND-based judgments.

From evaluation studies, for example [124], we know that modern metrics such as VMAF and FovVideoVDP can achieve correlations of up to 0.87 with subjective quality scores provided by participants. This close fit is possible because these metrics are trained or calibrated with data from subjective experiments, allowing them to link objective measurements to perceived quality. However, this data fit is non-linear. This approach makes these metrics especially dependent on the scope and biases of the data collected in the experiments. This is one of the reasons why Influential Factors of QoE are an inseparable problem from quality studies. Below, I will describe which IFs are taken into account in QoE applications, and later show what is missing.

1.6 Factors Considered in Applied QoE Models and Metrics

Dominant factors considered in applied QoE applications are Quality of Service (QoS) and its measurable Key Performance Indicators (KPIs). According to ITU-T E.804 [78], QoS refers to the end-to-end characteristics of a service that determine its ability to satisfy user needs, and it is typically operationalized through standardized KPIs such as delay, packet loss, or call setup success. These KPIs provide objective and reproducible measurements of service performance, forming the baseline for technical benchmarking.

Based on these baseline measurements, Key Quality Indicators (KQIs) can be inferred as user-oriented metrics derived from KPIs. According to ITU-T E.840 [81], KQIs represent aggregated, service-specific indicators that capture how well a network delivers quality from the end-user's perspective. They are ob-

tained by applying standardized quality-estimation models, such as ITU-T P.863 [82] for speech or ITU-T P.1203 [79] for video streaming, to underlying KPI data. In this way, technical measures like throughput, delay, or packet loss are translated into perceptual scales, often expressed in Mean Opinion Score (MOS) units. The Recommendation E.840 [81] emphasizes that KQIs should account for core service dimensions, including accessibility (whether a service can be reached), retainability (whether it remains active), integrity (the quality during service use), and availability (the likelihood that the service is offered).

In summary, the general approach followed in applied QoE modeling is to begin with QoS as measured through standardized KPIs, which provide objective and reproducible evidence of network performance. These technical measurements are then complemented with contextual service factors such as accessibility, retainability, integrity, and availability, which together capture whether a service can be reached, maintained, and experienced at an acceptable quality level. Based on ITU-T E.840 [81], these combined indicators are subsequently processed using standardized quality-estimation models in order to translate technical performance into perceptual outcomes. In this way, KPI-based measurements are mapped onto subjective quality scales such as MOS, enabling a systematic link between network operation and user-perceived quality. As a result, most applied QoE models remain limited to QoS-based KPIs combined with a narrow set of contextual service factors and mapped to subjective data that is collected in ways designed to minimize the influence of additional variables.

1.7 The Gap Between Theory and Practice in QoE Research

Having discussed QoE theory and QoE practice separately, it is now necessary to consider how they relate to each other. On the one hand, the literature offers rich theoretical models, including taxonomies of QoE Influential Factors (IFs). On the other hand, standardized test protocols (for example, [83]) aim for factor reduction to ensure comparability. Moreover, while QoE definitions describe it as a level of delight or annoyance, QoE metrics typically aim to predict impairment perception using either psychophysical data, such as JND, or descriptive scales such as ACR. Moreover, practical network application of models focuses on utilizing QoS in context to predict perceived distortion. This gap between theory and practice is a known problem in the domain and has been discussed previously e.g., in [57].

As noted by Peroni and Gorinsky [146], many pitfalls in QoE research grow from simplifying this inherent complexity. While QoE theory acknowledges a wide range of influential factors, practical work often narrows them to a limited subset, which risks the validity of resulting models and reinforces the divide between conceptual frameworks and applied methodologies. In response, researchers continue to develop methods for broader IF assessment. However, in the context of video studies, these methods are not yet as standardized or as widely adopted as the paradigms mentioned above, and they are almost never included in QoE metrics or models. Nonetheless, the measurement of IFs remains a crucial frontier in QoE research, aiming to close the gap between theory and practice. Below, I present a brief overview of exploration studies on IFs and experiments on the most relevant IFs to this dissertation.

1.8 Measuring Influential Factors

In exploratory studies, the main goal is to find the most important variables or patterns and, if possible, to identify the latent structure of the phenomenon. Models comprising the most important variables are outcomes of this method. This type of research is often used as a source of hypotheses for the following experimental designs. Below, I present a short overview of the exploration method used in QoE research.

Over time, numerous statistical techniques have been developed to extract insights from data. Approaches such as principal component analysis (PCA) [63, 132, 15] and factor analysis (FA) (e.g. [141, 179]) are commonly applied to uncover latent structures within datasets. In addition, methods like analysis of variance, covariance, and correlation are frequently employed to explore data patterns. More recently, machine learning and network analysis have gained popularity as tools for examining complex, nonlinear relationships [196]. All these techniques are well suited for handling datasets with multiple variables and factors. However, incorporating every possible IF into a single experimental design remains unfeasible. To capture such diversity, questionnaires are often introduced ([15, 63, 132, 41]). An overview of statistical factor analysis applied in QoE studies is presented in Table 1.8..

Besides statistical solutions, typical for an exploratory study, this approach can also be used with qualitative research. Analysis of interviews and open questions in questionnaires are a rich source of information that must be condensed and structured by researchers [50]. Similarly, bibliometric studies provide valuable insights into a vast amount of previous results and conclusions [199, 182, 123, 167, 12, 88]. Moreover, mixed methods, such as open profiling of quality (OPQ) [174], can provide quantitative results based on qualitative data. In other cases, mixed-method approaches have been used to deepen the understanding of quantitative data by means of a qualitative follow-up (see, e.g. [39, 140]). All the abovementioned methods share a similar goal. Their aim is to understand the collected data better. This type of insight is especially useful in the study of complex phenomena such as QoE IFs.

A good example of such exploratory research is a study by Baraković Husić and Baraković on multidimensional modelling of QoE for video streaming [15]. They conducted an experiment with 233 participants who evaluated 90 manipulated video sequences in which system factors such as resolution, coding tree unit, and constant rate factor were varied, contextual conditions such as location and lighting were recorded, and human characteristics such as gender, prior experience, and education were considered. The evaluation relied on a structured questionnaire with quality ratings, overall QoE, and perceived ease of use. Regression analysis revealed that perceived quality and perceived ease of use together explained about 68% of the variance in QoE, with perceived quality being by far the stronger predictor($\beta \approx 0.93$ vs. $\beta \approx 0.09$). ANOVA tests indicated that system, context, and human factors all had statistically significant effects on the perceptual dimensions, although the analyses focused on significance rather than the relative strength of these effects. This suggests that QoE in video streaming is shaped not only by technical parameters but may also be influenced by contextual conditions and user characteristics.

While it is important to include a broad range of factors in exploratory stages of QoE research, later modeling efforts often benefit from reducing the number of variables to those that most strongly predict user perception. An example is the study by Ben Youssef et al. [196], who combined Principal Component Analysis (PCA) with the Analytic Hierarchy Process (AHP) to identify the most influential application-

Table 1.2: Summary of QoE studies applying factor and component-based analysis

Authors / Year	**Method(s) Used**	**Collected Data**	**Number of Subjects**
(Özer, Argan, & Argan, 2013) [141]	PCA Confirmatory Factor Analysis (CFA)	Questionnaire	1000 participants
(Baraković Husić & Baraković, 2022) [15]	PCA Multiple Linear Regression (MLR)	Questionnaire, ACR scale	233 participants
(Ende, 2015) [179]	Exploratory Factor Analysis CFA T-test Correlation analysis	Questionnaire, ACR scale	Exploration: 107 participants Confirmation: 100 participants
(Msakni & Youssef, 2016) [132]	PCA and correlation analysis	Emotional Scale, ACR scale	72 participants
(Youssef, Mellouk, Afif, & Tabbane, 2016) [196]	PCA Machine learning	ACR scale	45 participants
(De Pessemier, Martens, & Joseph, 2013) [41]	MLR	Experience Sampling Diary Questionnaire	29 participants
(Ketykó, De Moor, Joseph, Martens, & De Marez, 2010) [63]	PCA and correlation analysis	Questionnaire, Experience Sampling Method	19 participants

layer factors for video quality assessment. From an initial set of eleven parameters, they found that frame rate, video size, audio rate, resolution, and mean bitrate were the best predictors of user-rated quality. Using these five factors in a Random Forest model yielded higher accuracy than models including all parameters, highlighting the value of statistical selection in building efficient and accurate QoE models.

Besides exploratory studies, there are QoE studies oriented on the direct assessment of a group of IFs in an experimental manner. A good example came in early studies on video on mobile devices [89], researchers evaluated the modulating effect of interest on content and its familiarity. In an audiovisual context, a significant interaction effect between interest and recognition was found ($F(9, 2390) = 19.80, p < 0.001$): unfamiliar but interesting content received higher quality ratings, while familiar and uninteresting content was rated lower. Similarly, in [105], the researchers found an influence on the desirability of the content. The desirability of the content alone explained up to 28% of the variance in quality ratings. However, in a more recent study [190], the effect of content interest was not found to be significant. This might be due to the different quality manipulation approaches used. The original studies measured the bit rate, and the author

of [190] manipulated the video stalling. This difference may also be explained by some human factors, such as the experiences, gender, or cultural and personality differences of the participants.

For example, study [136] found that gender and viewing frequency influence perceived video quality, with women and non-daily viewers giving up to 0.3 higher MOS ratings than men and daily viewers. Although this effect was observed at packet loss levels below 1–3%, beyond which the differences diminished. In [164], researchers measured personality traits and cultural aspects among the participants. Those human factors explained about 9.3% of the variance in quality ratings. Specifically, conscientiousness negatively predicted both perceived quality and enjoyment, while neuroticism positively influenced perceived quality, and openness increased enjoyment. Regarding cultural traits, individualism and pragmatism negatively affected perceived quality, while power distance and uncertainty avoidance negatively influenced enjoyment. Notably, in this experiment, participants were answering questions about satisfaction and enjoyment, not about quality measured with the ACR scale.

In a study in which only the quality rating was measured [137], human factors had a much smaller impact. Specifically, content delight and user mood showed only slight, statistically insignificant effects, while the frequency of watching online videos had practically no influence on quality ratings. This indicates that users can reliably evaluate video quality independently of personal preferences and emotional states when standardized procedures are followed.

Similarly, in [132], the emotional state that participants had before the study had no influence on the rating of the quality rating scale. However, quality influenced the emotional state of the participants after the test.

Based on those results, it is hard to draw an unambiguous conclusion. The general trend suggests that content-related effects were stronger in early studies, but under standardized testing conditions, their influence diminishes, leaving only minor contributions from individual factors such as gender, culture, or personality. One of the drawbacks of those studies is that they mix exploration of IFs with their confirmation. Moreover, the above-mentioned studies' exploratory part is conducted in a laboratory setting, and it is not followed by experimental evaluation. To address this, exploration should ideally be conducted in natural contexts, with subsequent confirmation carried out under controlled laboratory conditions. A purely explorative methodology would first allow the broad identification of candidate factors, which could then be measured and validated with standardized QoE procedures. To find a proper tool for exploration in a natural context, I also considered methods outside of the QoE domain. In effect, I found the memory-recall approach emerged as a potentially useful tool for IF exploration in QoE research.

1.8.1 Memory Recall Procedures and Their Application in Technology Studies

One of the ways to overcome the limitations mentioned above, where typical QoE tests measure participants in experimental conditions that are very different from normal video usage contexts, was proposed in [105]. Researchers sent DVDs with videos which quality had been manipulated, along with measurements for participants. Participants, during their experience with DVD in their homes, had to rate the quality. Similarly, De Pessemier et al. [41] applied a living lab approach, asking participants to watch videos on smartphones

in their everyday environments and rate loading speed and distortions through short questionnaires, while objective measures such as packet loss and startup delay were logged.

Those approaches can be described as the Experience Sampling Method (ESM). ESM is widely used in social sciences for gathering real-time data on behaviors, thoughts, and feelings [19]. In ESM, participants are prompted at random intervals to report their immediate experiences, thus capturing data in naturalistic settings and illuminating the dynamics of daily life. A clear application of this method to QoE was done by my colleague Natalia Cieplińska, who described it in her doctoral thesis [32]. She ran a longitudinal living lab study with a simple mobile app called DailyVideo. Participants watched short, preplanned clips with controlled compression at home. After each clip, they either wrote a brief free report about any problems or gave an ACR score, and at the end of each week, they provided one retrospective ACR for the whole week.

Although effective, ESM is intrusive and may alter participant behavior by disrupting natural routines. It also demands substantial participant commitment, potentially affecting compliance and data reliability.

To mitigate these limitations, the Day Reconstruction Method (DRM) was proposed by Kahneman et al. [51] for quality of life studies. Instead of multiple daily interruptions, DRM involves a single retrospective assessment where participants reconstruct their experiences from memory, answering questions about the contexts and associated emotions. The method combines features of time-budget techniques with the experience sampling method (ESM), aiming to capture the affective quality of everyday episodes while minimizing participant burden. Participants are first asked to divide the previous day into a sequence of episodes, like scenes in a film, and then describe each in terms of activities, social context, and emotional states using structured affect scales. This procedure has been shown to yield results that closely correspond with real-time experience sampling, while allowing for a more comprehensive coverage of the full day and rare or brief events. DRM is thus particularly useful in large-scale studies where intrusiveness, cost, and participant fatigue are a concern.

An example of DRM applied in technological research can be found in the work of Karapanos *et al.* [92], who conducted a five-week ethnographic study of six individuals purchasing the Apple iPhone. DRM was used to collect daily reflections on user experience, capturing both positive and negative episodes and assessing perceived qualities of the product in context. This method enabled the researchers to trace how user experience evolved through distinct phases. Orientation, incorporation, and identification—highlighting shifts in emotional attachment, functionality, and personal meaning over time. The study illustrates how DRM can reveal temporal dynamics in product adoption and help designers understand long-term user engagement beyond initial impressions.

In the context of Quality of Experience (QoE), the study by Wac et al. [185] represents a notable example of integrating the Day Reconstruction Method (DRM) to better understand user experience in natural environments. The research investigated Android users' QoE with mobile applications by combining objective data collection (QoS, sensors, logs) with subjective feedback gathered through Experience Sampling Method (ESM) and DRM-based weekly interviews.

The DRM was employed to support retrospective interviews, allowing participants to reconstruct their day as a sequence of episodes involving app usage, context, and experience. This approach enabled the researchers to combine information about context, system performance, and user perception, grounded in

the real-life routines and choices of mobile users. Through this mixed-method design, the study not only highlighted contextual and behavioral determinants but also showcased how DRM can bridge qualitative insights with system-level data, offering concrete design implications for mobile services.

Both of the above-mentioned applications demonstrate the usefulness of the DRM approach in technology-oriented user studies, yet they diverge from the original method proposed by Kahneman [51], which was designed as a one-day reconstruction of episodic experiences to approximate real-time sampling while minimizing recall bias. Karapanos [92] extended this approach into a longitudinal study, asking participants to reconstruct their daily experiences over five weeks in order to capture the temporal evolution of user experience. By contrast, Wac [185] adapted DRM into weekly retrospective interviews, combining 24-hour reconstructions with Experience Sampling and system log data as a qualitative complement to in-situ measurements. In this work, I return to an application of DRM closer to its original intent: a single-session reconstruction of a past experience aimed at collecting quantitative data while limiting biases that may arise from sustained attention to video quality.

1.9 Motivation, Goals, and Hypothesis

This section outlines the objectives and hypotheses that guided this thesis. The aim is to clearly present the reasoning and research logic that shaped the entire PhD project. Before I describe the hypothesis and objectives, I first describe the general research problem that motivated this thesis in particular.

1.9.1 Motivation

This work was possible thanks to funding from the Norwegian Financial Mechanism 2014–2021 under project 2019/34/H/ST6/00599. The project was titled “Towards Better Understanding of Factors Influencing the QoE by More Ecologically-Valid Evaluation Standards,” acronym TUFIQoE. To understand the goal of this project, a clear definition of ecological validity is needed. In the American Psychological Association dictionary [180, 13], ecological validity is defined as „the degree to which results obtained from research or experimentation are representative of conditions in the wider world. For example, psychological research carried out exclusively among university students might have a low ecological validity when applied to the population as a whole. Ecological validity may be threatened by experimenter bias, oversimplification of a real-world situation, or naive sampling strategies that produce an unrepresentative selection of participants.” During the project, we found with colleagues that the term ecological validity is interpreted differently across the field and often generates confusion. In effect, at the end of the project, a summary of the problems with the usage of the ecological validity term was proposed in paper [148]. Below, I present a justification for this change.

For the first time, the adjective „ecological” was used to describe the experiment by Egon Brunswik in 1944 [26]. In this work, the term was used to describe „intra-environmental physical relationships” that is, statistical interrelations among variables within the environment itself, independent of the organism’s responses. So the original term was very specific and distinct from the broad definitions of ecological validity, such as one described by APA. In [70], the authors trace the history of the term ecological validity back to

Brunswik's work and argue that, although widely adopted, the term has been used inconsistently and without a coherent definition by researchers. Beechey [17] summarizes this problem with the following words: „The use of the term ecological validity within hearing science appears to conflate realism, which is a means to potentially improve generalizability, and validity, which is a method of quantifying generalization." Authors of [148] describe that in effect, researchers use the term ecological validity as a measure of realism and assume that an increase in realism will automatically improve validity, which has not been proven. As the term ecological validity serves multiple purposes in the domain, the authors provide 5 definitions that should be used for a better description of the research problem.

- *Internal validity*: The degree to which a study or experiment is free from flaws in its internal structure, and its results can therefore be taken to represent the true nature of the phenomenon [13].
- *External validity*: The extent to which results obtained in one context can be generalized to another context [17].
- *Mundane realism*: Qualitative similarity between experimental conditions and everyday conditions experienced by the subject [17].
- *Psychological realism*: The extent to which psychological states or processes elicited by an experimental task are similar to those that occur during comparable tasks outside the laboratory [17].
- *Ecological validity* def. #2 from [13]: In perception, the degree to which a proximal stimulus (i.e., the stimulus as it impinges on the receptor) covaries with the distal stimulus (i.e., the actual stimulus in the physical environment).

These definitions may be adopted for a clearer description of the research problem in the TUFIQoE project and, consequently, in this thesis. According to the project proposal, the goal was to better understand the process of interacting with a video service. This was to be achieved by identifying factors influencing QoE through subjective testing methods with varying "degrees of ecological validity." However, in the proposal, the term ecological validity was used ambiguously, referring both to the technical properties of the experiments and to the overall project objective. This ambiguity can be resolved by replacing it with more precise terms: mundane realism and external validity. Thus, the actual goals of the project can be formulated as:

1. Increasing mundane realism by providing research protocols that resemble everyday usage experiences.
2. Investigate the external validity of QoE measures and metrics by verifying their generalizability to real-world service usage.

In this framework, the identification of Influential Factors on QoE can be understood as part of achieving both goals: factors should be studied under conditions of diverse mundane realism and assessed for their external validity.

1.9.2 Research Objectives and Hypotheses

As I argue in the above, QoE Influential Facts (IFs) are very broadly defined and describe various phenomena that might alter video users' experience. So the goal of this thesis is to explore this multifactor problem with a high mundane realism method and then verify the external validity of the outcomes. In this thesis, I use a memory recall procedure to gain data from everyday user experiences. There are a few arguments for that approach. Firstly, during everyday experience, users are not biased to focus only on quality. Moreover, if technical problems occurred, they are natural and not induced by experimenters. Finally, the context is as natural as possible. Thus, one can argue that such data has high mundane realism. There is also a more pragmatic argument for using the memory recall procedure for QoE studies. Such research is less expensive and easier to apply on a massive scale than laboratory studies.

Nevertheless, there are limitations to this method. Firstly, using this procedure will be less accurate than laboratory studies. Moreover, the number of data points from one person is limited. The lack of knowledge about exact technical conditions is yet another limitation. Thus, memory recall methods have some potential strengths but also serious limitations. As a consequence, it should be used for a well-defined purpose and in a clearly described research pipeline. For this reason, I use a memory recall-based questionnaire as a first step in the research. The above-mentioned strength of this method makes it a good fit for a tool to explore the IFs of QoE. By using statistical methods, those factors can be reduced and taxonomized. The most important factors can be used to build a model that consists only of the most important variables. Finally, this model can be applied in laboratory studies, providing a means to confirm or falsify the influence of the most important factors. This step ensures verification of the external validity of the exploration results. This pipeline is represented in Figure 1.1.

This dissertation consists of all those steps described in Figure 1.1. In Chapter 2, I present an exploratory part using a memory recall procedure. Then, in Chapter 3, I present a path model based on conclusions from Chapter 2. Consequently, Chapters 4 and 5 test the Influential Factors from Chapter 2 using the model from Chapter 3. Finally, conclusions and recommendations are described in Chapter 6.

To ensure clarity, a list of the research objectives and hypotheses used throughout the dissertation is presented below.

Hypotheses

- **Hypothesis One:** The influential Factors can be measured and taxonomized by a memory recall-based questionnaire.
- **Hypothesis Two:** Main Influential Factors can be combined in a simple model explaining their relationship.
- **Hypothesis Three** The propose model can help in planning comparable experiments targeting Influential Factors one by one.
- **Hypothesis Four:** Content has a strong influence on QoE ratings *.
- **Hypothesis Five:** Social presence has influence on QoE ratings*.

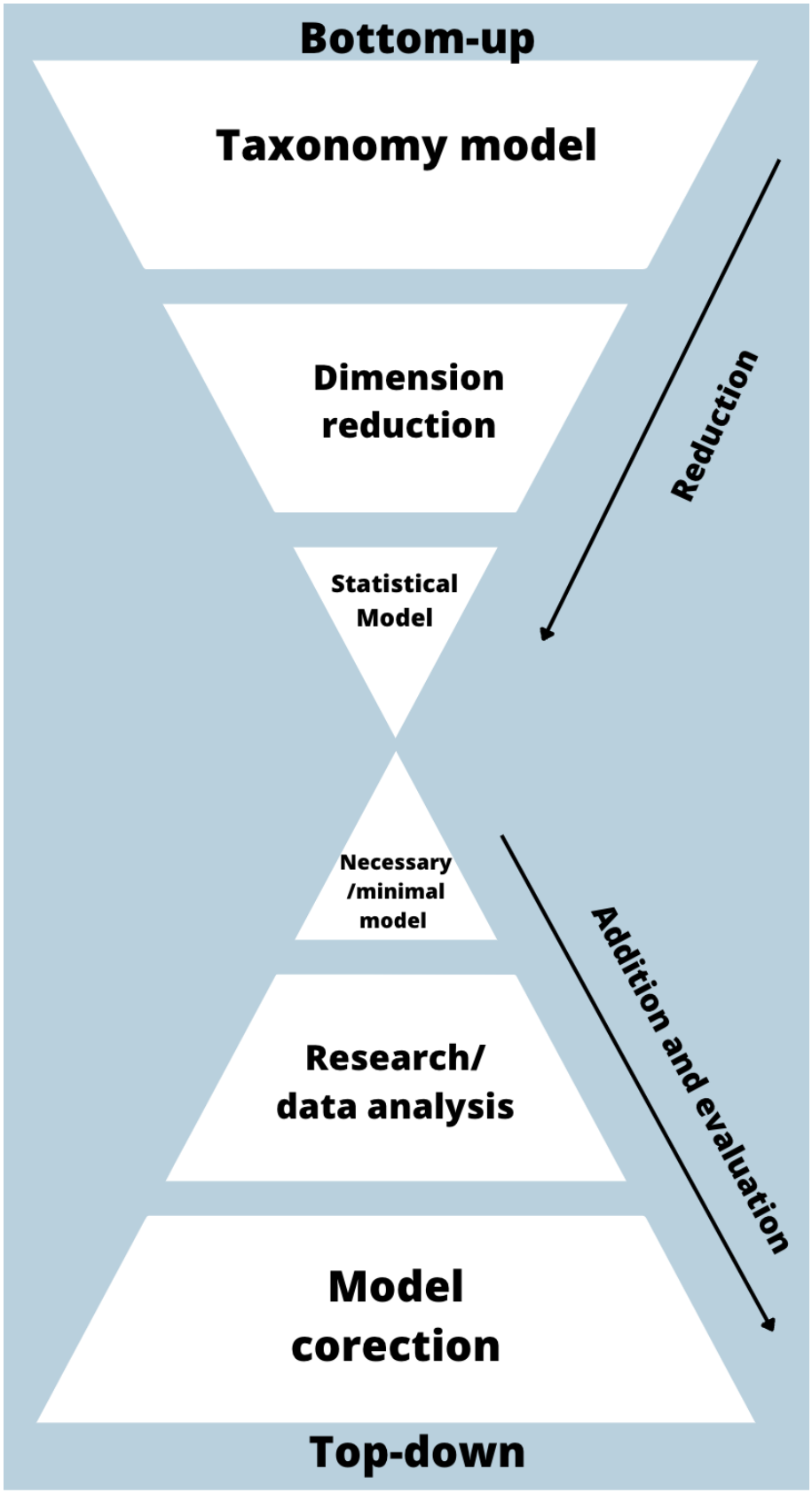

Figure 1.1: Modeling approaches. Figure sizes represent the number of variables in the model and arrows.

Objectives

- **Objective One:** Identify the most important Influential Factors for QoE using memory recall procedure.
- **Objective Two:** Design a parsimonious QoE model based on a minimal number of assumptions.
- **Objective Three:** Design a series of original studies based on one theoretical model for evaluation of Influential Factors.
- **Objective Four:** Provide recommendations for future QoE studies.

* Hypotheses Four and Five were formulated after memory recall studies as the outcome of the factor exploration process.

Chapter 2

Study 1: Exploration of Influential Factors Through a Memory Recall-Based Questionnaire

Parts of the material presented in this chapter are currently being prepared for submission as a research article, and therefore, I would like to acknowledge the contributions of all co-authors. The authors contributing to this part are Kamil Koniuch, Lucjan Janowski, Michał Wierzchoń, and Katrien De Moor. All authors agreed for their work to be included in the dissertation. Author's contributions were as follows: KK: conceptualization, methodology, investigation, formal analysis, writing – original draft, review, and editing; LJ: funding acquisition, formal analysis, supervision, writing – review; MW: consulting design and results, writing – review and editing, supervision; KDM: consulting design and results, writing – review and editing, supervision. This research was supported by the Norwegian Financial Mechanism 2014–2021 under project 2019/34/H/ST6/00599. The project was titled „Towards Better Understanding of Factors Influencing the QoE by More Ecologically-Valid Evaluation Standards", acronym TUFIQoE.

In this chapter, I address the first objective of my PhD research. To identify the most important Influential Factors (IFs) affecting Quality of Experience (QoE). For that purpose, I developed a memory recall-based questionnaire, which will be evaluated in this chapter. As outlined in this thesis introduction, to bridge the gap between QoE theoretical and practical models, there is a need for studies that provide an IFs taxonomy. With such gradation, models that balance the number of factors and predicting power could be established. For exploratory, part I choose to use the memory reconstruction procedure, to ensure the mundane realism of the measurement. In effect, I formulated Hypothesis 1, which postulates that IFs can be measured and taxonomized using a memory recall-based questionnaire. This hypothesis is tested in this chapter.

2.1 Introduction

In classical video Quality of Experience (QoE) studies, the influence of such factors is reduced through strictly controlled experimental protocols (e.g., [27]). These protocols effectively reduce the dimensionality by limiting the factors and variables under investigation. While this approach provides reliable results, typically with high internal validity, it overlooks the potential role of other IFs, which are not considered in the experimental design. This is because it is impossible to incorporate such a broad set of variables in one experimental setup due to their high complexity. However, it also implies that the relative importance of different types of influence factors in the context of QoE is, to date, still poorly understood.

Yet, to facilitate QoE optimization, e.g., across different contexts and user segments, key IFs have to be identified. This exploration and identification of the most important IFs is the first step to understanding the mechanisms of their influence. For this reason, self-report questionnaires are often used as post-experimental supplements to the setup (see e.g. [47]). This method has proven to be useful e.g. for novel, immersive multimedia [48].

However, such questionnaires are often inseparable from laboratory experiments. They are used to provide additional information before or after the subjective test. The limitation of this combined approach is that it requires a lot of resources and fails to represent and capture the everyday life technology experiences of participants, as well as the factors playing a role in that respect.

Given the above-mentioned limitations, I designed a questionnaire based on memory recall as a cost-effective, *stand-alone* method to explore a broad array of factors influencing QoE. The method allows for capturing data about everyday video usage without interrupting the experience itself and could be used on a massive scale. Additionally, it enables insight into goals fulfilled through different services, typical usage contexts, and beliefs about video quality.

As I described in the literature review section 1.8.1, such a reconstructive approach has been successfully applied in the related fields to understand e.g., the fulfillment of psychological needs through the use of social media [91]. However, to the best of my knowledge, it has not been explicitly or systematically used to study QoE IFs. To investigate this approach's applicability and potential value, it is crucial to understand if memory recall can provide valid results.

In this chapter, I present three substudies covering: Video On Demand (VOD), videoconferencing, and live streaming. All of them were designed to evaluate IFs in each use case. Moreover, I compared results before and after the application of a memory recall procedure. In total, 140 participants took part in this evaluation. The General Hypothesis, for this study, postulates that IFs can be measured and taxonomized using a memory recall-based questionnaire. This General Hypothesis can be further operationalized into three working hypotheses. First, IFs ratings will be statistically different before and after the memory reconstruction procedure. Second, the questionnaire will provide a statistically distinctive gradation of Influential Factors (IFs) for each use case. Third, the questionnaire will provide statistical differences for the most dominant factors between use cases. Due to the number of factors, multiple comparisons will be required to verify these hypotheses. While there are methods for adjusting to the number of comparisons, they are designed for confirmatory studies [18]. This study serves exploratory purposes, thus, the hypothesis will be verified with methods unadjusted to the number of comparisons.

Below, I present both the methodology and the results of these experiments.

2.2 Method

For the purpose of this study, I prepared an original questionnaire. Items in the questionnaire were formulated based on the literature described in section 1.8. Moreover, they were discussed with experts from the field. The first version of the questionnaire was evaluated in a pre-test with 14 participants (7 male, 7 female, mean age 29.4, S.D. 8.99) in terms of clarity and completeness. Participants from the pre-test answered items in the questionnaire during a live session with the author. They were commenting on each part of the questionnaire. On top of that, they could ask for clarification each time they need.

Based on the feedback, further improvements were made, and the questionnaire was implemented in Qualtrics [149]. The original study was conducted in the Polish language, and the questionnaire itself was translated for the purposes of this chapter and method publications. Full questionnaire in both language versions, with the data from the experiment, can be found at `https://github.com/TUFIQoE/questionaire`.

For the purpose of each substudy, I implemented three versions of the questionnaire, namely for VOD, live stream, and video chat. This design was used to verify whether the method captures differences across distinct contexts, thereby providing an indication of external validity. Each version had a similar structure of questions, thus enabling comparison between the three use cases. Each questionnaire was divided into five sections: demographics, motives, attitudes, factors in general, and factors after the memory reconstruction procedure. The questionnaire was preceded by a presentation of the purpose of the survey, along with an agreement form for participation.

140 participants (78 female, 54 male, and 5 non-binary individuals) were recruited to this study by means of a convenience sampling approach after participating in an unrelated QoE study. Participants were mostly students (65%), aged between 18-24 (63%). Based on a short survey about their everyday life experience with different types of services, participants were assigned to one of three main topic groups: VOD (N=50), video chat (N=48), or live stream (N=42). The inclusion criterion was time spent with a particular service during the preceding week. Given the lower popularity of live streaming compared to the other services, live streaming was prioritized when categorizing participants who used both VOD and live streaming. I present more details about demographics in section 2.3.

Questionnaires were displayed in the Qualtrics domain, on laptops in the lab. On average, filling out the questionnaire took 15 minutes (Mean=912.1 s SD = 248.3 s). Within the questionnaire, a focus-measuring question was embedded to ensure participants' attentiveness. Those who answered this question incorrectly were excluded. Consequently, during our initial data analysis, 6 of these participants were identified as outliers and removed from the dataset. Table 2.1 contains detailed descriptions of each sample after data cleaning.

Table 2.1: Demographic data categorized for three use cases

	VoD n=49		Video chat n=44		Live Stream n=41	
Sex	females	69.39%	females	61.36%	females	39.02%
	males	24.49%	males	38.64%	males	58.54%
	non-binary	4.08%	non-binary	0%	non-binary	2.44%
	prefer not to say	2.04%	prefer not to say	0%	prefer not to say	0%
Occupation	student	61.22%	student	51.27%	student	60.98%
	unemployed	4.08%	unemployed	2.27%	unemployed	2.44%
	employee	30.61%	employee	31.82%	employee	26.83%
	self-employed	0%	self-employed	4.55%	self-employed	2.44%
	other	4.08%	other	9.09%	other	7.32%
Age	18-24	57.14%	18-24	63.64%	18-24	68.29%
	25-34	34.69%	25-34	22.73%	25-34	26.83%
	35-44	4.08%	35-44	9.09%	35-44	4.88%
	45-54	4.08%	45-54	4.55%	45-54	0%
Time spent	non-user	4.08%	non-user	0%	non-user	2.44%
	occasional user	20.41%	occasional user	40.1%	occasional user	68.29%
	regular user	53.06%	regular user	50.0%	regular user	26.83%
	intensive user	20.41%	intensive user	9.09%	intensive user	2.44%
	overwhelmed user	2.04%	overwhelmed user	0%	overwhelmed user	0%
Competence in technology usage	to some degree	6.12%	to some degree	2.27%	to some degree	0%
	moderately	12.25%	moderately	13.64%	moderately	12.2%
	considerably	42.86%	considerably	40.91%	considerably	41.46%
	very good	38.78%	very good	43.18%	very good	46.34%
Connection type	LAN	2.04%	LAN	6.82%	LAN	12.2%
	Wi-Fi	44.9%	Wi-Fi	54.55%	Wi-Fi	63.42%
	cellular network	8.16%	cellular network	2.27%	cellular network	7.32%
	it depends	44.9%	it depends	36.36%	it depends	17.07%
Home internet satisfaction	not at all	4.08%	not at all	9.09%	not at all	0%
	to some degree	4.08%	to some degree	9.09%	to some degree	9.76%
	moderately	40.82%	moderately	25.0%	moderately	34.15%
	considerably	30.61%	considerably	31.82%	considerably	43.9%
	a lot	20.41%	a lot	25.0%	a lot	12.2%
Mobile internet satisfaction	not at all	2.04%	not at all	0%	not at all	2.44%
	to some degree	2.04%	to some degree	4.55%	to some degree	2.44%
	moderately	32.65%	moderately	11.36%	moderately	14.63%
	considerably	55.65%	considerably	52.27%	considerably	51.22%
	a lot	8.16%	a lot	31.82%	a lot	29.27%

2.2.1 Questionnaire Structure

Demographics and Usage Context

Demographic data often offers valuable insights into human factors. To capture this, the first section of the questionnaire focused on demographics, multimedia experience, and internet connection through a series of 12 closed-ended questions. This section resembles the demographic and socio-economic background part of Human Factors described in [53].

This section began by gathering information on participants' employment status, gender, and age. Subsequently, participants were asked about the multimedia services they utilized in the past month (see 2.1). Leveraging Qualtrics' conditional functions, subsequent sections of the questionnaire were tailored to delve deeper into the specific services participants had previously identified.

Which of the following video on demand (VOD) services have you used during the last month? Multiple choice is possible.

- HBO GO
- Netflix
- Hulu
- Disney
- Amazon Prime
- Showtime
- YouTube (excluding live stream content)
- Tik Tok (excluding for live stream content)
- Instagram (stories and videos only, excluding for live stream content)
- Vimeo
- Dailymotion
- Facebook (only "watch" video section excluding for live stream content)
- Vevo
- Streamable
- TED
- LiveLeak
- Apple Tv
- Twitter (videos only, excluding for live stream content)
- CDA
- Other _____

Figure 2.1: This is an example of services listed in the VOD version of the questionnaire.

Next, participants' expertise was measured with two items asking about the regularity of video service usage and proficiency with electronic devices in general. Subsequently, participants answered six questions about the internet connection that they use for video services, including the type of connection, subjective satisfaction, data plans, and payment.

Motives

After collecting demographic information, the questionnaire continued with questions addressing user motives for using video services. First, participants were asked about their reasons for using video services in the past month, with a multiple-choice question. See the list below.

- Pleasure and entertainment (e.g. not to be bored, to pass some time, etc.)

- Relaxation (e.g. Forgetting my daily duties, worries, distraction from stressful events, etc.)
- Learning and information seeking (e.g., updating information about important events, detailed information about how people do something step by step, etc.)
- Watching with friends or family
- To have a company (e.g. sound in the background or as a second activity, etc.)
- To separate from others (e.g. to have personal space for transportation or to be able to ignore surroundings, etc.)

In addition to these six predefined purposes listed, participants also had the option to provide their own answer. Based on this information together with previously chosen services, participants were asked to map purposes with services (see Figure 2.2.).

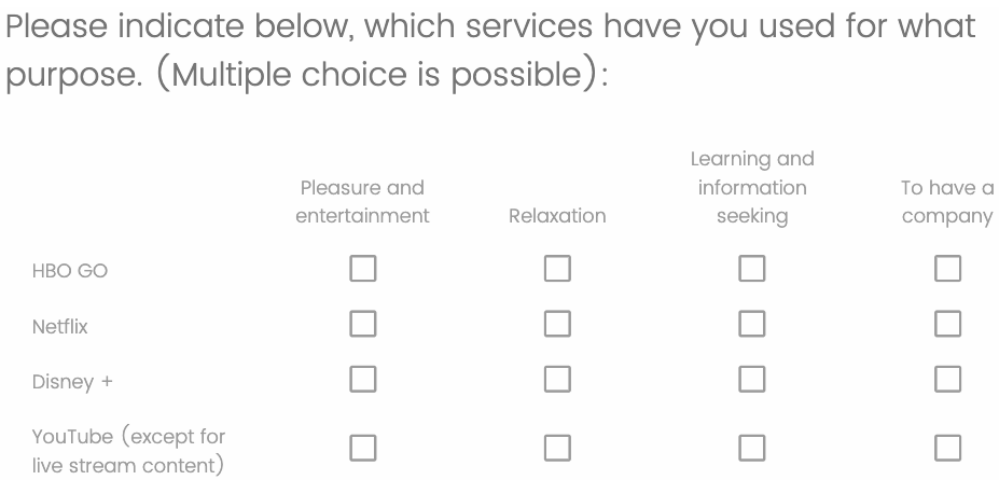

Figure 2.2: This is a preview example for participants selecting four services and four purposes, with multiple-choice answering.

The goal of this part of the questionnaire was to investigate intrinsic and extrinsic factors that drive participant engagement in video services [175]. A similar matrix was used to ask participants which devices they use for particular services. In this matrix, participants were asked about smartphones, tablets, laptops, PCs, projectors, and TVs.

Finally, at the end of this section, there were questions about the physical and social context of video usage. Participants were asked where they tend to use video services and how many people are usually around them during video sessions.

Attitudes Towards Video Quality

This section was designed to measure attitudes toward the quality of video, based on the ABC (affect, behavior, cognition) model of attitudes [24]. This section was built to investigate whether there is a difference between declarations about emotional affect and beliefs about quality and user behaviors. In other words, the goal was to see how behaviorally important video quality is.

In total, this section covered 7 items rated on a 5-point, nominal scale labeled: “Not at all”, “A little”, “Moderately”, “Considerably”, “A lot”. Items were adjusted for three different use cases (VoD, video chat, and live stream). Therefore, in the below list, ’video service’ notation replaces a particular example.

The first item was implemented to measure how video malfunctions affect users emotionally.

- "How annoying do you find video malfunction during using video service (low resolution, lack of smoothness, stalling, etc.)?".

Then I implemented 5 questions covering behavior related to video problems. Actions were chosen to represent diverse demands.

- “How convinced are you that you would call your Internet operator or write an e-mail to the Service Owner if you knew it may help you solve these problems?”.
- “How convinced are you that you would refresh your browser to improve video quality on video service?”.
- “How convinced are you that you would restart your Internet connection to improve video quality video service?”
- “How convinced are you that you would pay more to get better video service quality?”
- “How convinced are you that you would change Internet operator or video service Service Provider due to the poor video quality?”

Lastly, cognitive attitudes were measured by asking about the direct statement on the importance of video quality, as a component of QoE.

- "How important is video quality to you?"

At the end of this section, a trap question was used to measure participants’ attention during the test.

- "Many people experience irritation or satisfaction when viewing a video due to various factors. In this study, we want to try to understand this phenomenon better. For this purpose, we need to check how carefully people read our questions. Please choose the answer "considerably" below. Then, you can move on to the rest of the study related to the remembered example experience."

Below this question, the same scales as previously were displayed. If participants chose an option other than ’considerably,’ they were filtered out as outliers.

Factors Influencing Quality of Experience

This section of the questionnaire was divided into two parts, each containing matrix questions about the Influential Factors (IFs) of QoE (see Fig. 2.3). The same questions were asked twice, with a memory recall task inserted between the two parts. The goal was to investigate whether the memory recall procedure would increase the precision of the measurements.

Before the matrix directly asking about the influence of factors, the following instructions was displayed.

- In this section, we will ask you to evaluate various characteristics and factors in terms of how much you think they affect your video service experience. Imagine how much the given events and characteristics make the experience pleasant or annoying.
- Please mark it on a scale from “Not at all” to “A lot”. If you think an aspect has a big impact, choose “A lot” and if not, select “Not at all”. We encourage you to differentiate as much as possible during the assessment. There are no “right or wrong” answers. If you do not understand an item, please select "I don’t understand".

Then, in a series of matrix questions, participants were asked to rate “how much these factors impact your experience with video service”. VoD and live stream had 29 items (see. tab. 2.2). For video chat, the following items were removed due to their inadequacy for the use case:

1. Your knowledge of the content (how many times you’ve seen it, how much do you know what to expect)
2. The fact that it was only available on one type of service
3. The fact that it was or was not a premiere
4. Content genre (e.g. comedy, sports, talk show ...)
5. The presence of advertisements (playback interruption, screen covering, etc.).

Instead, two following items were add for video chat:

1. The fact that the meeting was recorded or not (possibility to watch it again later)
2. Your interaction with other participants

Table 2.2: Items measuring Factors Influencing QoE

A. Video quality factors	**B. Context factors**	**C. Content factors**
1. Fluency of the video (e.g., Occurrence of stalling events, frame drop, freeze, time jumps, lack of continuity, etc.)	1. Network connection efficiency	1. Your interest in the content
2. Image and sound synchronization	2. Application features (design, appearance, ease of use, ease of access)	2. Content importance/significance
3. The presence of artifacts or distortion in the video (visibility of shapes that are strange and unnatural)	3. Environment (lighting, time of day, comfort, temperature, etc.)	3. Number of crucial details (e.g., presence of small essential elements, little drawings/inscriptions, slides with graphics, etc.)

A. Video quality factors	B. Context factors	C. Content factors
4. Reproduction quality of dark/black parts of the video (visible blocks or other artifacts in the dark part of the video)	4. The presence of family, friends, or a supervisor/boss/teacher	4. Emotions evoked by the content
5. Video resolution (visible pixels, number of details, sharpness, etc.)	5. Your attention (multitasking / other activity, lack of sleep, etc.)	5. Duration of the clip/video/movie
6. Colors quality (reality, diversity, contrast)	6. Previous experiences (screen time, significant events of this day, etc.)	6. Your familiarity with the content (how many times you've seen it, how much do you know what to expect)
7. Device type (resolution, size, quality)	7. Your mood and emotions	7. Your appreciation of the content (e.g., whether you find it interesting or boring)
8. Visibility of details in dark scenes - the quality of these scenes	8. Cost/price (if access was paid, price of the service, how much money you have already spent on that service, etc.)	8. Content genre (e.g. comedy, sports, talk show...)
	9. Purpose of use - work, education, entertainment, etc...	
	10. The fact that it was only available on one type of service	
	11. The fact that it was or wasn't a premiere	
	12. Your expectations regarding the content (e.g., based on reviews, other people's recommendations, etc.)	
	13. The presence of advertisements (playback interruption, screen covering, etc.)	

In the matrix participants were using a similar scale as in the previous part of the questionnaire, namely: "Not at all", "A little", "Moderately", "Considerably", "A lot". The answer "I don't understand" was added to investigate if our questions were understandable for participants.

A matrix consisting of over 20 items would be overwhelming for participants. Thus, questions were divided into a series of matrices consisting of 6 items. Using Qualtrics, logic items were displayed randomly in every matrix (see fig. 2.3). This functionality enabled reduction of the grouping effect [64] on participants' answers.

Figure 2.3: Example of the matrix generated in Qualtrics. Participants were scoring factors in a set of 5 of those matrices. The order was randomly generated for each participant.

Memory Recall

After the first measurement of factors influencing QoE, a memory recall section consisting of 9 questions was presented to participants. At the beginning of this part, I asked participants to recall a unique, unforgettable experience (satisfying or unsatisfying) that they had while using a video service in the past month. Additionally, the services that participants claimed to use in the past month from section 2.2.1 were displayed to help them in memory recall.

Firstly, participants were asked about the time of the event- the day of the week and time of the day. Then, more detailed questions about context were implemented about the place, social presence, device, and service. At the end, participants were asked about the purpose, feelings, and whether the event was present or unpleasant.

After memory recall, questions about the factors influencing were displayed a second time. This time, participants were asked to rate this particular experience that they recalled. Because items describing factors influencing QoE are very precise, one more answer was added. Participants could now choose the option "absent" in the matrix questions. The purpose was to avoid forcing participants to rate events that might not be present in specific memory, such as the presence of advertisements.

2.3 Results

In this section, I present the results of all three substudies side by side to facilitate comparison across the three use cases: video on demand (VoD), live streaming, and video chat.

Results from each part of the questionnaire are presented in the same order as the questionnaire structure. The demographic characteristics are first described and illustrated in Table 2.1, followed by participants' motives, visualized in Figures 2.4 and 2.5. Subsequently, attitudes toward quality measurements are reported. The factor measurements taken before and after the memory recall procedure are then compared. Finally, the results obtained after the memory recall procedure are presented. All results are shown after the removal of outliers - responses from participants who failed the attention verification question were excluded from the analysis.

2.3.1 Demographics

In table 2.1, I present demographic statistics divided by each substudy. Across the surveyed platforms, females predominantly used VoD and video chat, while live streaming attracted more males. The 18–24 age group was the largest across all use cases, with no participants aged 45–54 in the live streaming group. Most respondents were students, though video chat included more participants combining work and study (as reflected in the "other" category). VoD users engaged more regularly, whereas live streaming was mostly used occasionally; video chat fell in between. While technological competence was generally high, live streamers most frequently rated themselves as "very good." Wi-Fi was the most common connection type, but VoD users showed the most variability. Home internet satisfaction was moderate across all groups, whereas mobile internet satisfaction was highest among video chat and live stream users, and lowest for VoD users.

2.3.2 Motives

Motives and usage context data were gathered by matrix-like questions with multiple answers possible. Thus, to analyze them, I prepared heatmaps (see fig. 2.4). This approach enabled the illustration of multiple dimensions of the data. Firstly, it shows which services are generally the most commonly used (the most frequent options are presented at the bottom of the chart). Secondly, the heat maps present the most popular motives for each use case on the left side of Figure 2.4.

While live stream and VoD were primarily reported as means for entertainment and relaxation, video chat users predominantly declared using it for learning and connecting with friends and family. YouTube was declared as the most common platform both for live stream and VoD. Moreover, the video chat heat map clearly shows the differences in the purposes of using various services. Messenger was reported as the most popular platform for connecting with friends and family, while Teams was the most commonly used tool for learning.

Similar heat maps are used to picture device usage per service (see Figure 2.5). VoD and live stream services were reported to be mostly used on smartphones. For the video chat use case, laptops were the most common device, with the exception of Messenger, which was primarily used on smartphones.

2.3.3 Attitudes Towards Video Quality

The importance of video quality was rated by participants, with 81.8% assigning a score of 4 "Quite a bit" (47%) or 5 "Extremely" (34.8%). In contrast, when asked about the likelihood of paying more for

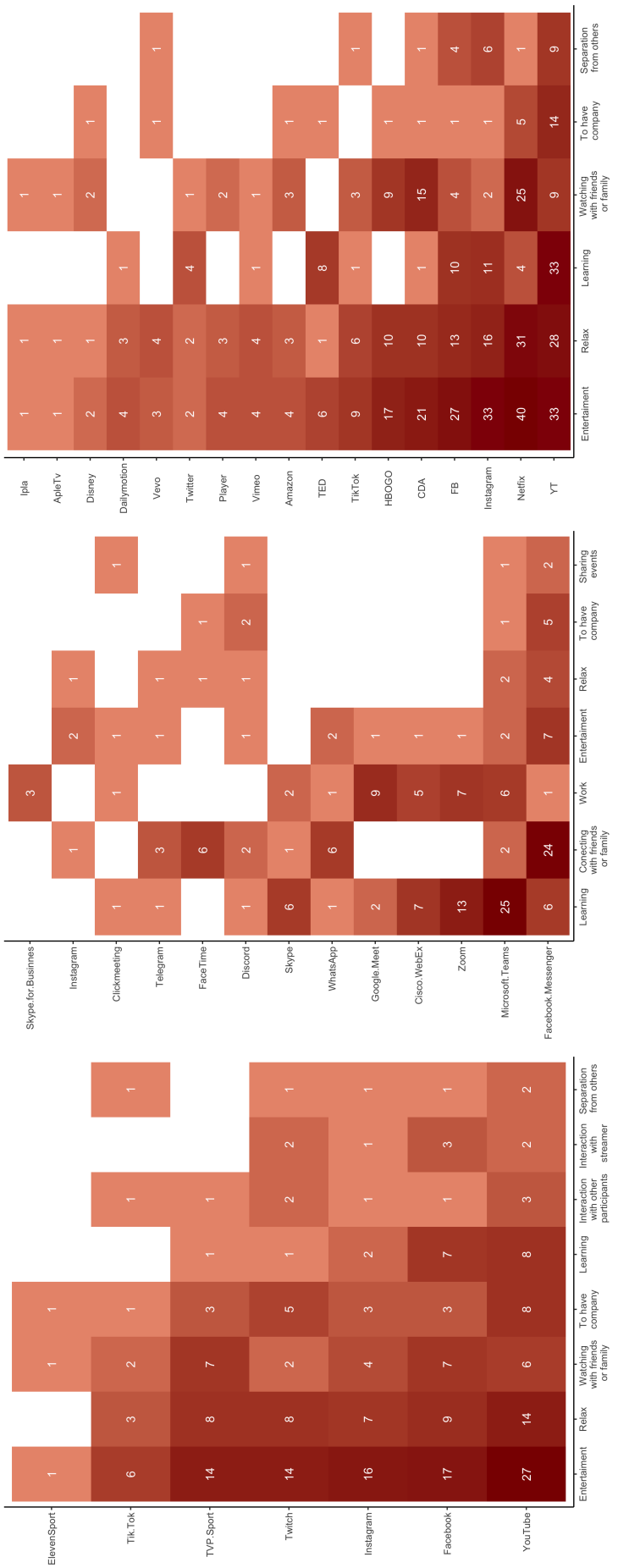

Figure 2.4: Purpose per service. Heat map representing the number of answers. From the left for live stream, video chat, and VoD. TVP Sport is a popular Polish live sports streaming platform. CDA is a popular Polish VoD service.

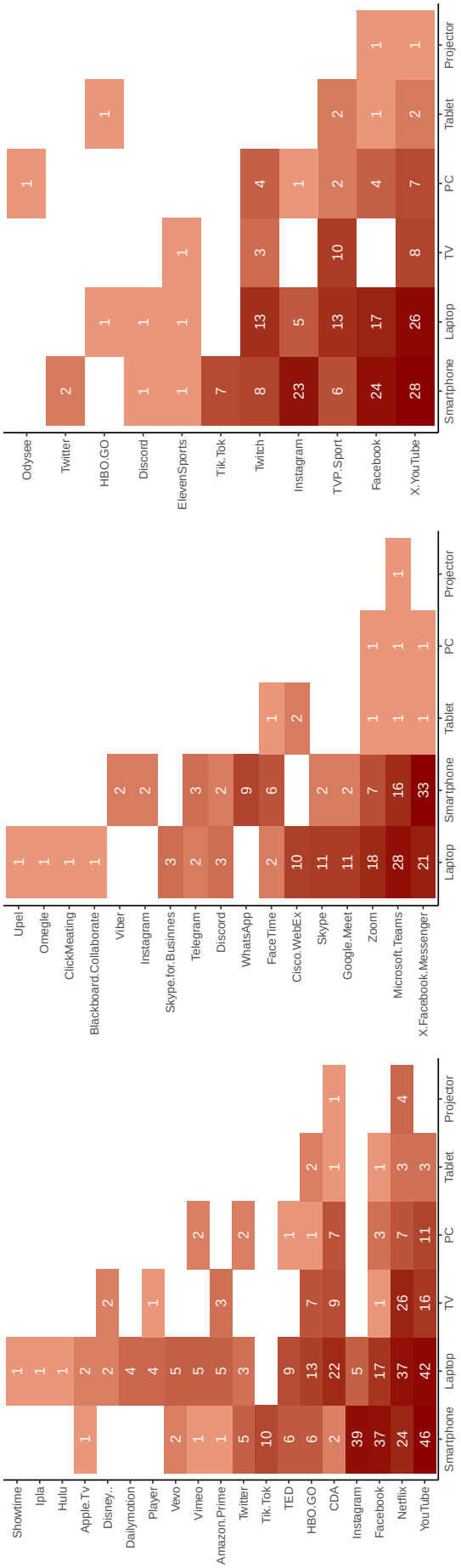

Figure 2.5: Device per service. Heat map representing the number of answers. From the left: VoD, video chat, and live stream.

improved video quality, only 28% selected either 4 “Quite a bit” (22.7%) or 5 “Extremely” (5.3%). These results suggest a distinction between perceived importance and actual willingness to invest financially. A Multivariate Analysis of Variance (MANOVA) was conducted to examine potential differences across the three use cases; no statistically significant effects were found. Thus, here I present combined results from all three substudies on one graph. The distribution of responses is presented in Figure 2.6. To further explore response variability, Wilcoxon signed-rank tests were applied. Pairwise comparisons were conducted in descending order of mean values, and statistically significant differences are marked by a change in color.

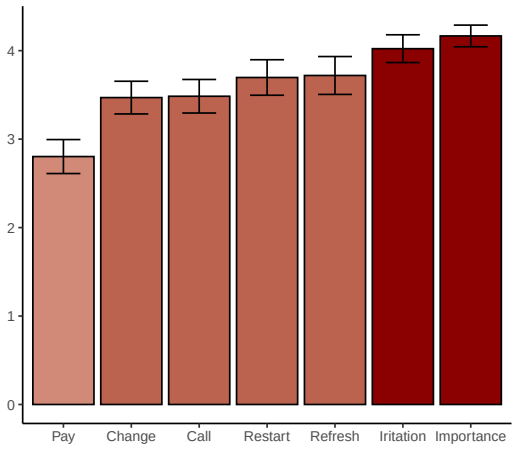

Figure 2.6: Wilcoxon signed-rank tests measured response differences, and significant changes are color-coded. MANOVA showed no significant differences between use cases thus, data is presented accumulatively for all use cases.

2.3.4 Influential Factors Before Memory Reconstruction

This section presents the self-descriptive importance of each Influential Factor of Quality of Experience. Both tables and graphs presented in this subsection analyze the same means. Figure 2.7 presents the outcomes of the three substudies side by side. The average self-reported influence of each factor is indicated by the length of the bars, with 95% confidence intervals included. Paired comparisons of means were conducted using the Wilcoxon test. Whenever a statistically significant difference was found, the color of the bar was changed, and the next factor in the sequence was used as the reference for comparison.

In the same figure 2.7, the first two color groups can be interpreted as the most important. For both the VoD and live stream substudies, three key groups of factors emerge: interest and appreciation of content, technical quality, and advertisement. In contrast, for video chat, only technical factors are reported as most important.

The differences between substudies were measured with the Wilcoxon test pair by pair. In total, 77 comparisons were made. Tabel 2.4 present means compared with statistically significant results marked with bold text and an asterisk. These are unadjusted tests, so at the 0.05 significance level, about 4 of the 77 comparisons may represent chance findings. Among the significant differences, appreciation, details, interest, and previous events show notable variation, particularly in comparisons involving video chat. For

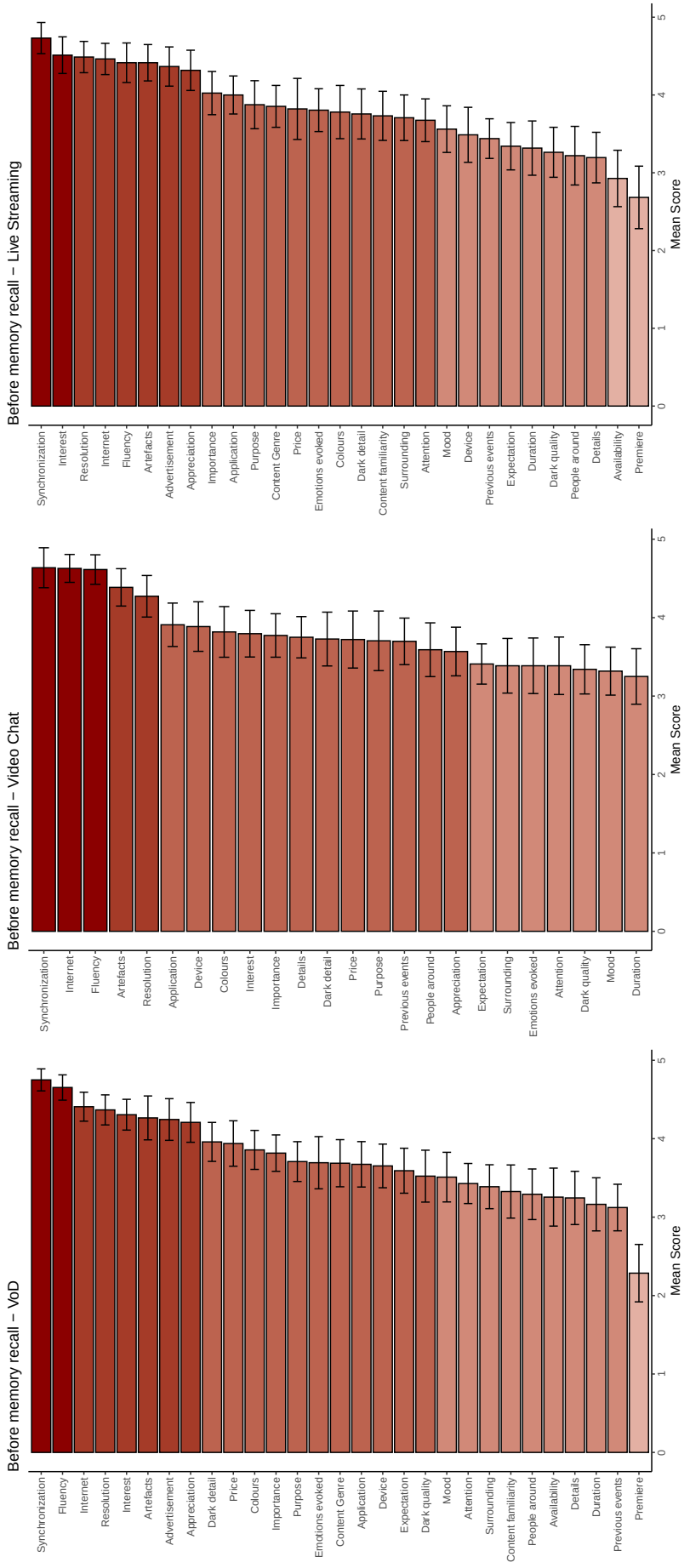

Figure 2.7: Mean scores before memory recall procedure for VoD, video chat, and live stream. Differences between means were measured by the Wilcoxon test. Significant differences are reported as changes in color.

instance, video chat was perceived as less influenced by appreciation and interest than VoD and live stream. This suggests that users attributed their satisfaction or dissatisfaction during video chats less to content-related factors, highlighting a stronger role of interpersonal or technical aspects in shaping their experience. Table 2.3 summarize only statistically significant differences.

Table 2.3: Summary of statistically significant differences

Substudy	VoD	VChat	p (VoD–VChat)	Live	p (VChat–Live)
Appreciation	4.21	3.57	**0.0018***	4.32	**0.0005***
Details	3.24	3.75	**0.0374***	3.20	**0.0166***
Interest	4.31	3.80	**0.0091***	4.51	**0.0003***
Previous events	3.12	3.70	**0.0180***	3.44	0.221

Table 2.4: Mean comparison with Wilcoxon test between each pair of substudies factors.

Label	Mean_VoD	Mean_VChat	p_VoD_VChat	Mean_VoD	Mean_Live	p_VoD_Live	Mean_VChat	Mean_Live	p_VChat_Live
Advertisement	4.24	NA	not present	4.24	4.37	0.615	NA	4.37	not present
Application	3.67	3.91	0.287	3.67	4	0.147	3.91	4	0.722
Appreciation	4.21	3.57	**0.0018***	4.21	4.32	0.570	3.57	4.32	**0.0005***
Artefacts	4.27	4.39	0.768	4.27	4.41	0.726	4.39	4.41	0.941
Attention	3.43	3.39	0.901	3.43	3.67	0.290	3.39	3.67	0.288
Availability	3.26	NA	not present	3.26	2.93	0.164	NA	2.93	not present
Colours	3.86	3.82	0.866	3.86	3.78	0.932	3.82	3.78	0.860
Content familiarity	3.33	NA	not present	3.33	3.73	0.118	NA	3.73	not present
Content Genre	3.69	NA	not present	3.69	3.85	0.665	NA	3.85	not present
Dark detail	3.96	3.73	0.493	3.96	3.76	0.407	3.73	3.76	0.938
Dark quality	3.52	3.34	0.398	3.52	3.26	0.238	3.34	3.26	0.705
Details	3.24	3.75	**0.0374***	3.24	3.2	0.834	3.75	3.2	**0.0166***
Device	3.65	3.89	0.165	3.65	3.49	0.522	3.89	3.49	0.077
Duration	3.16	3.25	0.685	3.16	3.32	0.564	3.25	3.32	0.913
Emotions evoked	3.69	3.39	0.199	3.69	3.8	0.825	3.39	3.8	0.096
Expectation	3.59	3.41	0.272	3.59	3.34	0.249	3.41	3.34	0.952
Fluency	4.65	4.61	0.842	4.65	4.41	0.187	4.61	4.41	0.281
Importance	3.82	3.77	0.794	3.82	4.02	0.185	3.77	4.02	0.174
Interest	4.31	3.8	**0.0091***	4.31	4.51	0.083	3.8	4.51	**0.0003***
Internet	4.41	4.63	0.077	4.41	4.46	0.670	4.63	4.46	0.200
Mood	3.51	3.32	0.366	3.51	3.56	0.889	3.32	3.56	0.244
People around	3.29	3.59	0.228	3.29	3.22	0.692	3.59	3.22	0.143
Premiere	2.29	NA	not present	2.29	2.68	0.121	NA	2.68	not present
Previous events	3.12	3.7	**0.0180***	3.12	3.44	0.209	3.7	3.44	0.221
Price	3.94	3.72	0.470	3.94	3.82	0.846	3.72	3.82	0.643
Purpose	3.71	3.7	0.561	3.71	3.88	0.277	3.7	3.88	0.729
Resolution	4.37	4.27	0.863	4.37	4.49	0.378	4.27	4.49	0.346
Surrounding	3.39	3.39	0.818	3.39	3.71	0.141	3.39	3.71	0.257
Synchronization	4.75	4.64	0.944	4.75	4.73	0.768	4.64	4.73	0.840

2.3.5 Effect of Memory Recall Procedure

In the memory recall section, 75.51% of VoD users recalled pleasant memories, 10.20% remembered unpleasant ones, and 14.29% were uncertain in their categorization. For video chat users, 54.55% recounted pleasant memories, 31.82% brought up unpleasant ones, and 13.64% were ambiguous in their classification. Meanwhile, live stream users reported 78.05% pleasant memories, 9.76% unpleasant memories, and 12.20% that were challenging to categorize. In table 2.5 present which service that were recalled by participants is presented.

Table 2.5: Services recalled by participants divided into 3 substudies.

VoD	%	Video Chat	%	Live stream	%
Netflix	57.1	Teams	34.1	TVP Sport	32.5
Youtube	18.4	Messenger	13.6	YouTube	27.5
CDA	12.3	Skype	11.4	Twitch	22.5
HBO GO	6.1	Telegram	9.1	Facebook	10.0
Facebook	4.1	Google Meet	9.1	Instagram	7.5
Instagram	2.0	WhatsApp	6.8		
Other	2.0	Zoom	4.6		
		Face Time	4.6		
		Instagram	2.3		
		Cisco	2.3		
		Other	2.3		

Respondents were able to relate to certain factors better when having a specific experience in mind. For example, in the section before memory recall, participants claimed in all 3 substudies 22 times that they did not understand the question. On the other hand, after recall, "I don't understand" response was used only 11 times, cumulatively for all use cases. A similar effect was observed in the pre-test phase, in which participants answered with more ease after memory recall.

Another argument in favor of using memory recall comes from reliability analysis. Cronbach's alpha [37] is widely used as a measure of the internal consistency of questionnaires and the reliability index. For all three substudies, alpha was higher in the part after memory recall. (VOD: 0.75 to 0.82, video chat: 0.78 to 0.87, live stream: 0.72 to 0.82).

Table 2.6 presents the change in mean before and after the recall procedure. Means were compared using the Wilcoxon test. Several statistically significant differences emerged, particularly in the VoD and live streaming conditions. Notably, advertisement and price consistently showed a significant decrease in perceived influence after recall across all use cases, suggesting that these aspects may be initially overestimated during immediate evaluations.

Technical factors such as synchronization, internet quality, resolution, dark quality, and fluency also showed significant declines, especially in the VoD and live stream contexts, indicating a reduced importance of system performance after reflective recall.

In contrast, emotional factors like emotions evoked and mood either remained stable or slightly increased in influence after recall, particularly in live streaming and video chat.

The video chat condition showed fewer changes overall, but still exhibited significant drops in perceived influence of dark detail, price, and details, indicating that some technical and visual factors are also re-evaluated for scoring one particular memory.

Table 2.6: Comparison of mean factor ratings before and after memory recall, substudy using the Wilcoxon test. Statistically significant differences ($p < .05$) are shown in bold.

Label	**VoD**			**Video Chat**			**Live Stream**		
	Before	After	p	Before	After	p	Before	After	p
Advertisement	4.24	3.59	**p = 0.0344**	–	–	–	4.37	3.43	**p = 0.0098**
Application	3.67	3.00	**p = 0.0201**	3.91	3.37	p = 0.0799	4.00	3.21	**p = 0.0053**
Appreciation	4.21	4.16	p = 0.8430	3.57	3.86	p = 0.1190	4.32	4.38	p = 0.7540
Artefacts	4.27	3.79	p = 0.1020	4.39	4.50	p = 0.4960	4.41	3.52	**p = 0.0097**
Attention	3.43	2.67	**p = 0.0006**	3.39	3.67	p = 0.2450	3.67	3.43	p = 0.2930
Availability	3.26	2.45	**p = 0.0067**	–	–	–	2.93	2.43	p = 0.0708
Colours	3.86	3.59	p = 0.3890	3.82	3.27	**p = 0.0326**	3.78	3.33	p = 0.1130
Content Genre	3.69	3.62	p = 0.8820	–	–	–	3.85	3.90	p = 0.4980
Content familiarity	3.33	3.09	p = 0.3990	–	–	–	3.73	3.22	p = 0.1070
Dark detail	3.96	3.43	p = 0.0519	3.73	2.82	**p = 0.0041**	3.76	3.13	p = 0.0809
Dark quality	3.52	3.18	p = 0.2560	3.34	2.65	**p = 0.0239**	3.26	2.50	**p = 0.0091**
Details	3.24	2.88	p = 0.1740	3.75	3.11	**p = 0.0247**	3.20	2.53	**p = 0.0126**
Device	3.65	3.59	p = 0.8660	3.89	3.51	p = 0.1540	3.49	3.28	p = 0.6450
Duration	3.16	2.90	p = 0.3800	3.25	3.47	p = 0.3530	3.32	2.74	**p = 0.0464**
Emotions evoked	3.69	3.86	p = 0.5150	3.39	3.81	p = 0.0794	3.80	4.28	**p = 0.0072**
Expectation	3.59	3.34	p = 0.5720	3.41	3.71	p = 0.1520	3.34	3.18	p = 0.8760
Fluency	4.65	4.02	**p = 0.0025**	4.61	4.55	p = 0.5740	4.41	4.18	p = 0.4580
Importance	3.82	3.78	p = 0.7400	3.77	3.77	p = 0.7500	4.02	4.05	p = 0.6960
Interest	4.31	4.33	p = 0.6510	3.80	4.02	p = 0.1270	4.51	4.55	p = 0.8180
Internet	4.41	3.84	**p = 0.0151**	4.63	4.50	p = 0.4480	4.46	4.05	p = 0.0931
Mood	3.51	3.84	p = 0.1410	3.32	3.86	**p = 0.0136**	3.56	3.83	p = 0.2260
People around	3.29	3.22	p = 0.9420	3.59	3.53	p = 0.9080	3.22	2.97	p = 0.4770
Premiere	2.29	1.87	**p = 0.0481**	–	–	–	2.68	3.00	p = 0.4230
Previous events	3.12	2.76	p = 0.0831	3.70	3.32	p = 0.2230	3.44	2.95	**p = 0.0210**
Price	3.94	2.60	**p = 0.0000**	3.72	2.90	**p = 0.0047**	3.82	2.95	**p = 0.0213**
Purpose	3.71	3.60	p = 0.9220	3.70	4.05	p = 0.1620	3.88	3.67	p = 0.4410
Resolution	4.37	4.04	p = 0.1360	4.27	3.95	p = 0.1970	4.49	3.83	**p = 0.0033**
Surrounding	3.39	3.17	p = 0.5090	3.39	3.37	p = 0.9680	3.71	3.15	**p = 0.0318**
Synchronization	4.75	4.20	**p = 0.0099**	4.64	4.33	p = 0.1170	4.73	4.20	**p = 0.0037**

2.3.6 Influential Factors After Memory Reconstruction

In the matrix question, participants had the opportunity to choose “not present” if some of the factors were irrelevant to their recollection. This design served two purposes. Firstly, participants were not forced to choose a response on the ordinal scale when a specific factor was not part of their experience. Secondly, it gives an overview of which factors are most common. Participants chose the option “not present” 89 times in VOD, and 75 and 99 times in video chat and live stream, respectively. Figure 2.8 presents the percentage of missing values for each factor in the VOD substudy. For this use case, the most commonly missing factor was the influence of advertisements. Figure 2.9 represents the number of missing factors in the video chat substudy. In this substudy, there were no questions about the advertisement, as for the time of the study, there were no communicators using advertising in the service. Consequently, the price factor was most commonly reported as "not present." Figure 2.10 depicts missing values for the live stream use case. Similarly to video chat, price was the rarest factor. Moreover, advertisements were reported more commonly as not present in live streams compared to VoD.

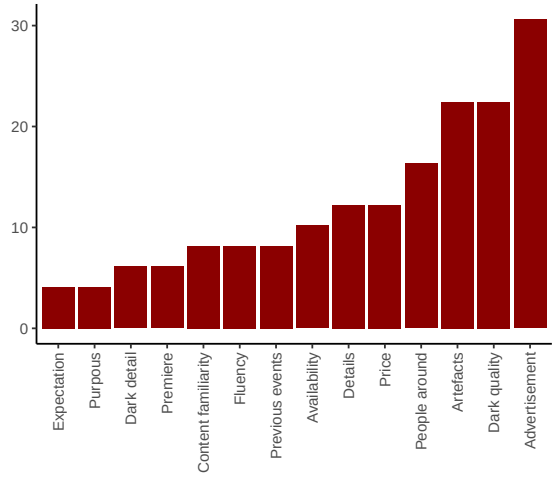

Figure 2.8: Percentage of "not present" answers for factors influencing VOD.

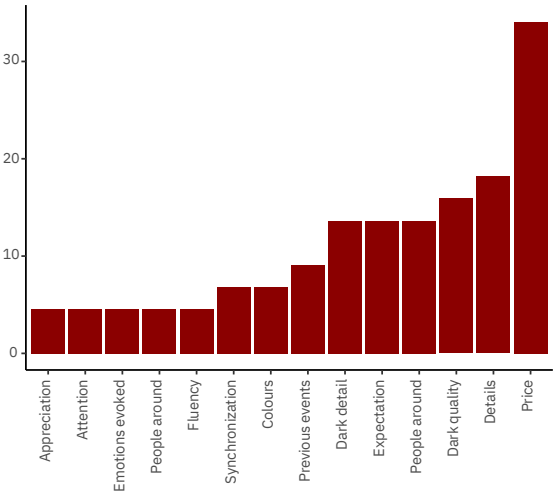

Figure 2.9: Percentage of "not present" answers for factors influencing video chat.

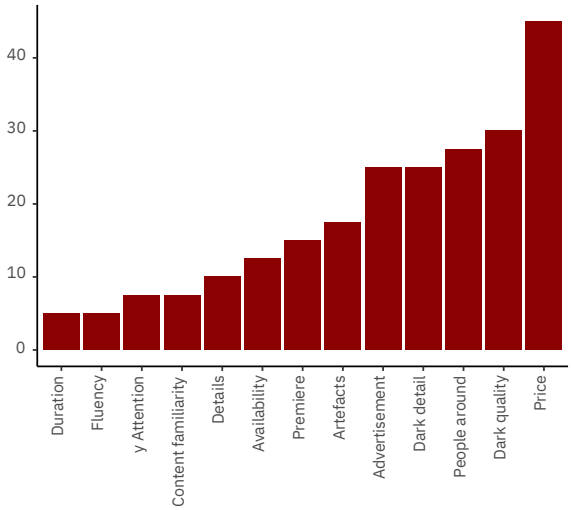

Figure 2.10: Percentage of "not present" answers for factors influencing live stream.

To analyze the differences between use cases, Ordered Logistic Regression (OLR) was used to compare the influence of each factor across service types (VoD, Video Chat, Live), while controlling for whether the recalled memory was pleasant or not. For each factor, a separate OLR model was fitted using the `polr()` function from the `MASS` package in R. Three pairwise comparisons were extracted from each model using Wald tests based on the estimated coefficients and their standard errors: VoD vs. Video Chat, VoD vs. Live, and Video Chat vs. Live. The resulting p-values are reported in Table 2.8. A total of 77 such comparisons were conducted across all included influential factors. These are unadjusted tests, so at the 0.05 significance level, about 4 out of the 77 comparisons may occur due to chance.

The model estimated the probability of a response level y_i for each factor using the cumulative logit formulation:

$$\log\left(\frac{P(Y_i \leq k)}{P(Y_i > k)}\right) = \theta_k - (\beta_1 \cdot \text{UseCase}_i + \beta_2 \cdot \text{Pleasure}_i) \tag{2.1}$$

where θ_k are the intercepts (thresholds) for category k, and β_1, β_2 are coefficients for the predictors.

The table 2.8 compares Impact Factors across three video service substudies. In the table 2.7, all statistically significant effects are reported.

Notably, differences were most pronounced for technical aspects such as artefacts, and the Internet, where video chat was often rated higher than VoD. Emotional and contextual factors, including emotions evoked, Interest, and people around, showed significantly higher scores for live streaming compared to video chat. A few effects were also observed between VoD and live, including dark quality and premiere.

Table 2.7: Summary of statistically significant differences (after recall)

Substudy	VoD	VChat	p (VoD–VChat)	Live	p (VoD–Live)	p (VChat–Live)
Appreciation	4.16	3.86	0.206	4.38	0.260	**0.0259***
Artefacts	3.79	4.50	**0.0326***	3.52	0.364	**0.0041***
Attention	2.67	3.67	**0.0002***	3.43	**0.0031***	0.374
Dark detail	3.43	2.82	**0.0356***	3.13	0.402	0.269
Dark quality	3.18	2.65	**0.0400***	2.50	**0.0404***	0.979
Duration	2.90	3.47	0.076	2.74	0.640	**0.0329***
Emotions evoked	3.86	3.81	0.907	4.28	**0.0346***	**0.0371***
Interest	4.33	4.02	0.086	4.55	0.246	**0.0071***
Internet	3.84	4.50	**0.0053***	4.05	0.327	0.078
People around	3.22	3.53	0.140	2.97	0.370	**0.0252***
Premiere	1.87	–	–	3.00	**0.0019***	–
Previous events	2.76	3.32	**0.0359***	2.95	0.453	0.171

Table 2.8: Comparison of mean scores and p-values between conditions

Label	Mean VoD	Mean VChat	p (VoD–VChat)	Mean VoD	Mean Live	p (VoD–Live)	Mean VChat	Mean Live	p (VChat–Live)
Advertisement	3.59	Not present	not present	3.59	3.43	0.747	Not present	3.43	not present
Application	3	3.37	0.388	3	3.21	0.489	3.37	3.21	0.851
Appreciation	4.16	3.86	0.206	4.16	4.38	0.260	3.86	4.38	**0.0259***
Artefacts	3.79	4.5	**0.0326***	3.79	3.52	0.364	4.5	3.52	**0.0041***
Attention	2.67	3.67	**0.0002***	2.67	3.43	**0.0031***	3.67	3.43	0.374
Availability	2.45	Not present	not present	2.45	2.43	0.885	Not present	2.43	not present
Colours	3.59	3.27	0.253	3.59	3.33	0.281	3.27	3.33	0.959
Content familiarity	3.09	Not present	not present	3.09	3.22	0.672	Not present	3.22	not present
Content Genre	3.62	Not present	not present	3.62	3.9	0.225	Not present	3.9	not present
Dark detail	3.43	2.82	**0.0356***	3.43	3.13	0.402	2.82	3.13	0.269
Dark quality	3.18	2.65	**0.0400***	3.18	2.5	**0.0404***	2.65	2.5	0.979
Details	2.88	3.11	0.452	2.88	2.53	0.234	3.11	2.53	0.071
Device	3.59	3.51	0.920	3.59	3.28	0.447	3.51	3.28	0.526
Duration	2.9	3.47	0.076	2.9	2.74	0.640	3.47	2.74	**0.0329***
Emotions evoked	3.86	3.81	0.907	3.86	4.28	**0.0346***	3.81	4.28	**0.0371***
Expectation	3.34	3.71	0.245	3.34	3.18	0.570	3.71	3.18	0.106
Fluency	4.02	4.55	0.077	4.02	4.18	0.418	4.55	4.18	0.366
Importance	3.78	3.77	0.637	3.78	4.05	0.232	3.77	4.05	0.110
Interest	4.33	4.02	0.086	4.33	4.55	0.246	4.02	4.55	**0.0071***
Internet	3.84	4.5	**0.0053***	3.84	4.05	0.327	4.5	4.05	0.078
Mood	3.84	3.86	0.923	3.84	3.83	0.935	3.86	3.83	0.985
People around	3.22	3.53	0.140	3.22	2.97	0.370	3.53	2.97	**0.0252***
Premiere	1.87	Not present	not present	1.87	3	**0.0019***	Not present	3	not present
Previous events	2.76	3.32	**0.0359***	2.76	2.95	0.453	3.32	2.95	0.171
Price	2.6	2.9	0.340	2.6	2.95	0.314	2.9	2.95	0.902
Purpose	3.6	4.05	0.101	3.6	3.67	0.954	4.05	3.67	0.123
Resolution	4.04	3.95	0.602	4.04	3.83	0.402	3.95	3.83	0.770
Surrounding	3.17	3.37	0.477	3.17	3.15	0.845	3.37	3.15	0.387
Synchronization	4.2	4.33	0.720	4.2	4.2	0.789	4.33	4.2	0.553

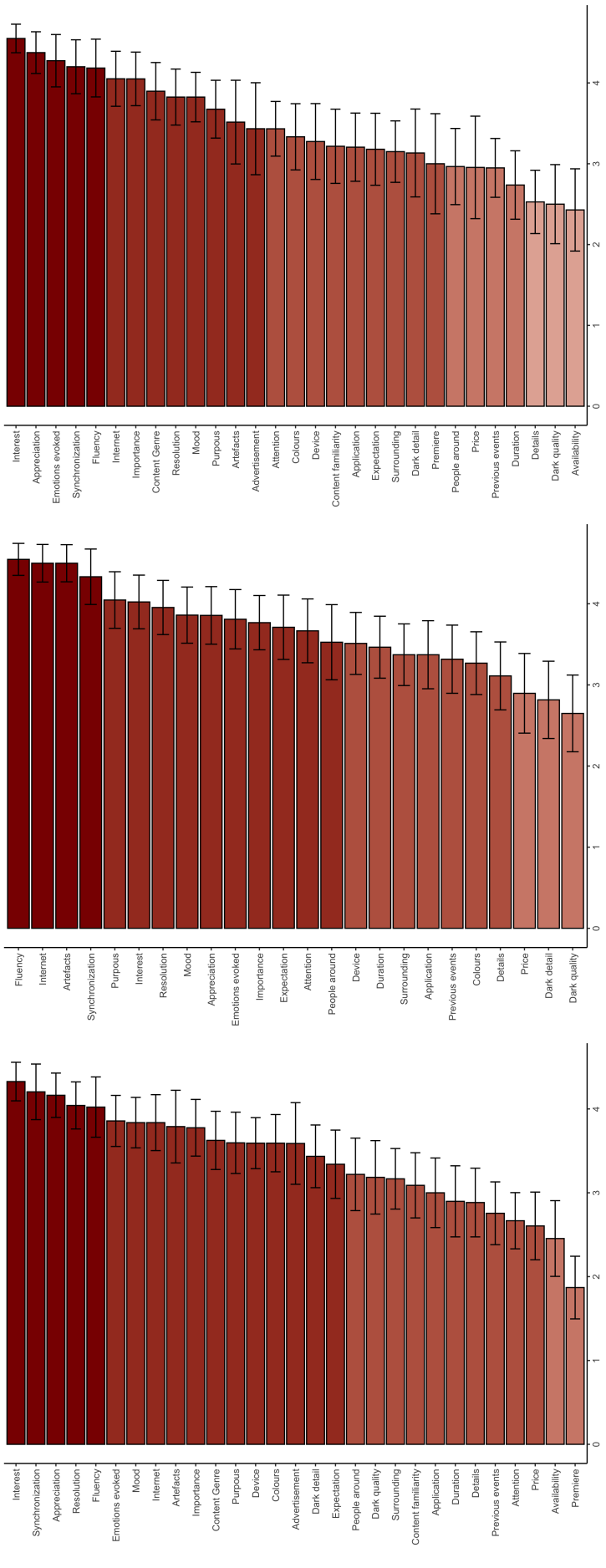

Figure 2.11: Mean scores after memory recall procedure for VoD, video chat, and live stream. Differences between means were measured by the Wilcoxon test. Significant differences are reported as changes in color.

2.4 Results Discussion

In this chapter's discussion, firstly, I describe the demographics of participants in subsection 2.4.1, highlighting the role of expectations depending on context. Then address participants' attitudes towards video quality in subsection 2.4.3, emphasizing the gap between cognition and behavior. In subsection 2.4.4 first working hypothesis (IFs ratings will be statistically different before and after the memory reconstruction procedure) is discussed along with effect of memory recall. Finally, I present the factors influencing the participants' Quality of Experience (QoE) in subsection 2.4.5, with a relation to second and third working hypothesis. Those hypothesis were stated as: the questionnaire will provide a statistically distinctive gradation of Influential Factors (IFs) for each use case, and differences between IFs will be statistically significant across use cases. Discussion highlights the content's role in shaping users' experiences.

2.4.1 Demographics

Demographics were captured through an extensive section of questions (see tab. 2.1). This allows for the assessment of multiple characteristics of samples assigned to a specific substudy. The groups were relatively balanced in terms of occupation, age, competence in technology usage, and internet satisfaction. A relatively high level of Internet satisfaction resembles good access to the Internet in Poland. According to the Central Statistical Office of Poland, at the time of the study, 93.3% of Poles had access to broadband internet [29]. This data includes 69.5% with fixed-line broadband internet access and 70.6% with access to broadband mobile internet. Substudies were less comparable in terms of sex, time spent on video usage, and preferred type of internet connection. Notably, live stream users, including occasional users, were more often male and utilized more LAN and Wi-Fi connections. As the sample was not extensive enough, additional subsampling analysis of the demographics is limited. I will discuss this issue in more detail in subsection 2.4.6

2.4.2 Motives

The heat maps (fig. 2.4) distinguish between services in terms of fulfilled goals. For instance, entertainment is the most popular application for VOD and live streaming, but not for video chat. Additionally, learning is the most common purpose for using video chat. This trend is likely influenced by the COVID-19 pandemic, which took place during the study. The pandemic disrupted the education of 90% of students worldwide at the time of our study [178, 54]. Moreover, for "watching with relatives," participants are more likely to choose Netflix over YouTube, while the opposite is true for learning. Similarly, participants prefer using smartphones and laptops over other video devices when utilizing video services (see fig. 2.5).

These findings not only offer insights into consumer behavior but also contribute to the theoretical understanding of QoE by highlighting the role of user expectations and prior experience. Such expectations may vary depending on the device and use case. Therefore, in the future, the proposed questionnaire could be employed to select individuals with specific multimedia usage experience for QoE subjective experiments. Comparing such groups with different usage profiles might assist us in operationalizing and quantifying the influence of user experience and expectations on QoE. With the insights gained from the questionnaire study, researchers can establish which factors are most crucial to include in future studies.

2.4.3 Attitudes Towards Video Quality

Referring to Figure 2.6, a discernible discrepancy emerges between the perceived importance of video quality and the willingness to pay. These results suggest that attitudes toward quality can be influenced by context (e.g., socio-economic factors). Moreover, behavioral inclinations can vary. While a substantial 81.8% emphasized the importance of video quality, only a mere 28% exhibited a willingness to pay more for superior quality. This divergence, explored through the ABC (Affect-Behavior-Cognition) theory of attitudes [24], suggests that despite the affective component (satisfaction/dissatisfaction with video quality), it does not straightforwardly translate into a corresponding behavioral response (willingness to improve quality via payment). The cognitive component, or beliefs about video quality, might be influenced by various factors, such as perceived value or financial disposition, which could be explored in future research. The findings highlight a gap between feelings and actions regarding video quality and user experience. It's essential to understand these underlying thought processes better. Most QoE research relies on self-reports, like the Absolute Category Rating scale [83]. It's worth noting that the differences between cognition, affect, and behavior in this context might be more pronounced than currently understood, due to the limitations in the traditional test designs. These insights are fundamental for the development of the author's theoretical model described in chapter 3.

2.4.4 Recall Procedure

To decide if the influential Factors can be measured and taxonomized by a memory recall-based questionnaire, first, the differences before and after the memory recall procedure have to be considered. As I stated in the first working hypothesis for this chapter, IFs ratings will be statistically different before and after the memory reconstruction procedure. Based on the above-presented results, I found this hypothesis supported by data.

The results indicate a consistent decrease in the perceived importance of technical factors after the memory recall procedure. Synchronization, resolution, internet quality, and dark detail, after the memory recall procedure dropped, particularly in the VoD and live stream substudies 2.8. This shift may be attributed to the tendency of participants to recall more pleasant experiences, as previously shown in the distribution of memory valence. When reflecting on positive moments, users might focus less on system-level performance and instead emphasize the content or emotional engagement aspects of the experience. This trend suggests that recall-based evaluations may increase the importance of content-related and emotional factors in shaping perceived quality, aligning more closely with users' actual affective and experiential priorities in everyday use.

Based on the memory recall data, video chat appears to be more frequently associated with unpleasant or ambiguous recollections compared to VoD and live stream services. This suggests that users may encounter more challenges or negative experiences when using video chat. At the same time, across all service types, pleasant memories were overrepresented, indicating a general tendency for users to recall positive experiences when reflecting on past usage. Moreover, this result might imply that technical problems with quality are relatively rare in natural conditions. As I indicated above, participants reported a relatively high level of Internet satisfaction, which reflects a good Internet availability in Poland [29].

Taken together, the increase in internal consistency (as indicated by higher Cronbach's alpha after recall), the reduction in "I don't understand" responses, and qualitative feedback from participants all support the inclusion of a memory recall procedure prior to the questionnaire. This step appears to enhance both comprehension and reliability of the responses.

However, the current results also suggest that participants predominantly recalled pleasant memories, potentially introducing a positivity bias in retrospective evaluations. It is also possible that experiences where the network is not sufficient are rarer than positive ones. To address this, future studies should consider experimentally balancing memory valence by instructing half of the participants to recall a pleasant memory and the other half an unpleasant one. Such an approach would allow for more systematic investigation of how memory valence influences perceived quality and help draw clearer conclusions about the relative importance of different Influencing Factors.

2.4.5 Factors Influencing QoE After Recall

Before discussing a statistically distinctive gradation of Influential Factors (IFs) and differences between IFs across use cases, I will provide information about service usage and missing values in this part of the questionnaire.

In this section, users rated Factors Influencing QoE only based on one particular memory. Table 2.5 presents the distribution of services that participants remembered. For the VoD use case, the Netflix platform was predominant, as 57.14% of participants recalled using this service. CDA and HBO were remembered by 12.25% and 6.12% of participants, respectively. As a result, the majority of responses in this segment pertained to platforms that offer professionally produced content and have limited interactivity. Only 24.5% of participants mentioned social media platforms known for user-generated content and interactive interfaces. The video chat category was more evenly distributed, though Microsoft Teams was still the most recalled. The Live stream category was the most balanced of all, with TVP Sport being notably high at 32.5%, likely due to the concurrent European Football Championship.

Having established which services were most recalled, it's also important to note the Influence Factors that were often absent from participants' memories. The factors that were often underrepresented, such as price, advertisement, and the presence of people around, highlighted the usefulness of the "not present" option for participants (see figs. 2.8, 2.9, and 2.10). Omitting this option from the questionnaire would have resulted in a loss of valuable insights. For instance, while VoD users frequently reported the absence of advertisements in their recalled memories, both video chat and live stream users commonly indicated that the price factor was not present.

While some factors were less frequently recalled, their importance shouldn't be underestimated. A factor's infrequent occurrence did not always translate to its diminished importance. In the VoD scenario, nearly 25% of participants did not encounter any artifacts. However, among those who did, the perceived impact of these artifacts was considerable, with the mean influence ranking among the top 9 Impact Factors (see fig. 2.11). This highlights that the scale gauges two distinct dimensions simultaneously. Although this method does not force participants to rate absent factors, it does introduce methodological challenges, no-

tably in producing a significant amount of missing, non-random data. The implications of this approach are further discussed in subsection 2.4.6.

Having established the frequency of these factors, next I discuss mean scores for their influence (see fig. 2.11). This figure support, second working hypothesis for this chapter providing statistically distinctive gradation of Influential Factors (IFs) for each use case

In the video chat use case, the most important factors are related to internet efficacy. Factors such as fluency, internet quality, occurrence of artifacts, and synchronization of sound and video were most important for participants. On the other hand, for VoD and live stream use cases not only technical aspects were important, but also the value of the content itself. Factors such as interest, appreciation of the content, and emotions evoked by the content were at the top of the most important factors. This shows that for live stream or VoD QoE studies, the content itself is a crucial factor. This is an important insight because many existing theoretical models of QoE (e.g. [52, 151, 131, 163, 154, 53, 44]) often downplay its role, emphasizing instead technical parameters. On the other hand, price factors were surprisingly low-scored. It is especially interesting because in some QoE studies (e.g. [117, 118]), price is part of experimental manipulation as an important influential factor. This might suggest that in real-life usage scenarios, price might be less influential than in laboratory experiments.

.

Furthermore, the working hypothesis stated that the questionnaire would provide statistical differences for the most dominant factors between use cases. The Table 2.8 provided support for this hypothesis. Scores in this Table contains Ordered Logistic Regression, adjusted for the pleasantness of memories. First of all, the reproduction of details in dark scenes was reported as more influential for VOD. This might be the result of low-light aesthetics, which is frequently used in professionally generated content and which makes the encoding process more challenging [46]. Services based on this type of content were frequently used by our participants (see fig. 2.4 and tab. 2.5). Moreover, Internet stability was significantly more important for video chat, which is in line with previous studies [140]. Additionally, duration as a temporal IF had a stronger impact on video chat, which might be explained by fatigue associated with longer online meetings [49].

Interest in content was found to be most influential in both the VOD and live stream use cases. However, only the difference between video chat and live stream is statistically significant. Similarly, the role of appreciation of the content was found to be significantly different between the use cases, but again, only when comparing live stream and video chat. One possible explanation is that the three use cases contribute to different goals (see fig. 2.4). While both live stream and VOD are mostly used for entertainment and relaxation, VOD additionally could be used for educational purposes, which is the most common goal for video chat reported in the sample. On top of that, at the time of the study, the European Football Championship took place, which might also explain the higher importance of the “evoked emotion” and “premiere” factors in the live stream.

2.4.6 Limitations

Despite having an extensive demographic section in the questionnaire, it was impossible to leverage it fully. With a larger participant pool, I could have conducted more between-group comparisons. Additionally,

more data points would have allowed to use the information from the section 2.2.1 to model the influence of participants' experience and expectations. A larger sample would also likely provide a more diverse range of recalled memories. A particular concern was the overrepresentation of Netflix usage in the VoD use case sample, as seen in table 2.5. Moreover, the majority of the memories recalled were positive, as indicated in 2.3.5. In a larger study, this could be addressed by instructing half of the participants to recall a positive memory and the other half a negative one.

Secondly, while the recalling procedure simplifies the task for participants and potentially provides more precise data, it does introduce some methodological challenges. Some factors, such as the occurrence of advertisements, might be entirely absent in recalled memories, leading participants to answer "did not occur". This method inherently measures two separate aspects of the experience within one scale: whether a factor was present and, if it was, its level of influence. For exploratory purposes, like mean measurements and comparisons, this is not a major concern. However, methods such as Exploratory Factor Analysis, Confirmatory Factor Analysis, or Structural Equation Modeling are sensitive to missing data. Furthermore, the data is not missing at random. The absence of certain factors in recalled memories carries significant information. Therefore, I caution against using this method for such analyses, even if the sample size is large. Nonetheless, surveying users of similar services or applying this method after a subjective experiment could address this issue.

2.5 Conclusions

One of the key conclusions drawn from the results presented in this chapter is derived from the attitudes section of the questionnaire. The notable discrepancy between the perceived importance of video quality and potential user behavior underscores the need for new QoE measurement approaches. Even if participants in subjective tests recognize and declare differences in quality as important, it doesn't necessarily imply that they would alter their consumer behaviors based on quality alone. To face that challenge, in the next section, I describe a theoretical framework that distinguishes between affective and behavioral outcomes of video service usage. Having a comparison between users' declarations and actions would shed more light on that phenomenon. This insight is also a strong recommendation for future QoE studies to distinguish between the perceptual, affective, and behavioral layers of QoE, which is the first step to fulfill Objective Four of this dissertation. I discuss this further in Chapter 6.

Secondly, the assessment of Influential Factors across three different use cases indicates that participants' needs vary based on the use case. Notably, the significant role of content in VoD and live stream experiences stands out both before and after the memory recall procedure. As highlighted in the discussion section 2.4.5, the role of content is often overlooked, even in established theoretical models of QoE. Furthermore, many QoE studies reduce the importance of content by repeatedly displaying the same videos during the study. These findings suggest that in real-life scenarios, content can play a crucial role in evoking delight or annoyance. This implies that content can act as a confounding factor, influencing both compression and subjective QoE assessments. These insights were also incorporated in the next chapter and in the recommendation part 6.

Answers for factors after memory recall for VoD were used to formulate hypotheses four and five of this dissertation. Besides technical factors, lots of properties of content itself were reported as most influential. Among contextual factors, social presence was reported as the highest. With these insights, I covered first objective of this dissertation and identified the most important Influential Factors. Moreover, the hypothesis one is backed by the data above. In both cases, with and without the memory recall procedure, the Influential Factors were measured and taxonomized based on participants' answers.

The analysis provided in this chapter highlights the potential of the questionnaire and memory recall procedure as a stand-alone method for gathering diverse and extensive data about Quality of Experience. Using the memory recall procedure within the questionnaire offers insights into factors influencing QoE in everyday experiences. Moreover, this method can be applied on a large scale at a relatively low cost since it doesn't require conducting subjective experiments. Such a large-scale study would allow for the full utilization of the method, including the analysis of influential human factors. Furthermore, this kind of study could help in understanding the market by categorizing users and their consumption behaviors.

Moreover, the questionnaire could complement traditional subjective QoE tests. Firstly, the demographic and attitudes sections of the questionnaire can facilitate purposive sampling, allowing for an examination of whether expectations, habits, or attitudes toward quality influence subjective tests. This can be achieved by screening participants for their consumer behaviors and attitudes before recruitment. Furthermore, the demographic information gathered can provide a detailed description of the sample, potentially enhancing the comparability of QoE studies and simplifying the replication process [104]. Consequently, this unified tool could address the issue of overlooking human factors in QoE studies [33].

The part of our questionnaire that measures influential factors can also be used as an add-on to traditional QoE tests. Most QoE studies primarily focus on technical influential factors, as they are often the independent variable. By using select items from the questionnaire, researchers can gain insights into factors outside of experimental manipulation. This could help in understanding, for example, what participants prioritize when scoring a video on the ACR scale. Such insights can enhance the conclusions drawn from a study. Additionally, if influential factors are measured using the same questionnaire, it allows for comparisons between different QoE studies on new dimensions.

The full Video Quality of Experience Questionnaire, along with the study results, can be found at `https://github.com/TUFIQoE/questionaire`. Additionally, the GitHub repository includes the original method, a stand-alone memory recall questionnaire, and its post-experiment version.

In summary, this chapter highlights the effectiveness of the questionnaire in understanding Quality of Experience, emphasizing the importance of content and the distinction between perceived video quality, affective response, and user behavior. Using the memory recall procedure and analyzing influential factors, the questionnaire can be used both independently and alongside traditional QoE tests. The strong argument in favor of Hypothesis 1, is that the questionnaire provided differentiation between factors in various use cases. In the future, this approach might be widely adopted, contributing to greater comparability of QoE studies across the field. With a better understanding of Factors Influencing the perceptual, emotional, and behavioral part of QoE, new research paradigms and quality metrics can be developed. With this approach, metrics with greater external validity can be built. I discuss this conclusion more in Chapter 6.

Chapter 3

Theoretical Model of Video Quality of Experience

This chapter is an enriched version of previously published articles; thus, I would like to acknowledge the contributions of all co-authors.

Parts of the material presented in this chapter were first introduced in my single-authored doctoral consortium papers [100, 101], which presented the initial proposal. A subsequent practical application of the model was showcased at QoMEX 2023 [103]. The authors contributing to this part are Kamil Koniuch, Lucjan Janowski, Katrien De Moor, Michał Wierzchoń, and Sruti Subramanian. All authors agreed for their work to be included in the dissertation. Author's contributions were as follows: KK: conceptualization, methodology, writing – original draft, review and editing; LJ: funding acquisition, supervision, writing – review and editing; KDM: supervision, writing – review and editing; MW: supervision, writing – review and editing; SS: writing – review and editing.

The full theoretical model was later published in Frontiers in Computer Science [102]. The authors contributing to this publication were Kamil Koniuch, Sabina Barakovic, Jasmina Barakovic Husić, Sruti Subramanian, Katrien De Moor, Lucjan Janowski, and Michał Wierzchoń. All authors agreed for their work to be included in the dissertation. Author's contributions were as follows: KK: conceptualization, writing – original draft, writing – review and editing; SB: conceptualization, writing – review and editing; JH: conceptualization, writing – review and editing; SS: conceptualization, writing – review and editing; KDM: conceptualization, writing – review and editing; LJ: conceptualization, writing – review and editing; MW: conceptualization, writing – review and editing. Following the clarification of the legal framework for doctoral dissertations issued by the Rada Doskonałości Naukowej [135], Subsection 3.3.1 and 3.3.2 include direct quotes from [102].

This research was supported by the Norwegian Financial Mechanism 2014–2021 under project 2019/34/H/ST6/00599, titled „Towards Better Understanding of Factors Influencing the QoE by More Ecologically-Valid Evaluation Standards".

Moreover, I would like to thank Narciso García, Pablo Pérez, Jesús Gutiérrez, Marta Orduna, and Carlos Cortés for their valuable insights and constructive feedback provided during the IMG subgroup meeting of VQEG held in Madrid in 2023, which helped shape and refine the proposed model.

In this chapter, I address the second objective of the dissertation to provide a parsimonious model of video QoE. I also verify the second hypothesis, that most important factors from the exploratory part of the research can be organized in a simple, parsimonious model. To operationalize this hypothesis, I propose working hypothesis that most important IFs can be describe in form of a Direct Acyclic Graph that explain causal relations between factors.

3.1 The Role of Theoretical Models

Theoretical models serve multiple purposes in science, such as organizing and structuring complex factors, guiding experimental design and interpretation, and increasing comparability. They are also efficient tools for communication between researchers [103]. In the context of implementation research, Nilsen emphasizes three overarching aims for theoretical approaches: describing and guiding processes, explaining influencing factors, and evaluating outcomes, thereby clarifying and systematizing otherwise complex phenomena [139].

Building on this general role, theoretical models also play a crucial part in statistical inference. Statistical models act as a bridge between empirical data and theoretical concepts [36]. Theoretical models' value lies in providing a framework in which theoretical assumptions can be translated into testable statistical structures, enabling both interpretation and inference.

Since statistical inference is not only driven by data but also shaped by analytical judgment [36], the same dataset can yield different conclusions depending on the statistical model employed. As McElreath illustrates [128], varying model specifications may generate misleading or unstable conclusions if not carefully considered. Some of the common pitfalls include:

1. Spurious associations: These arise when a predictor appears important only because it is correlated with an unmeasured confound. Such correlations can create the illusion of causality, leading to false conclusions about relationships that do not exist in reality.

2. Masked relationships: When two predictors have opposite influences on an outcome but are themselves correlated, their effects can cancel one another out. As a result, simple bivariate models may hide genuine associations that only become visible in a multivariate model.

3. Multicollinearity: When predictors are highly correlated, the model cannot distinguish their separate effects. This leads to unstable parameter estimates that may even flip signs, creating confusion in interpretation, even though overall predictions may still perform well.

4. Post-treatment bias: This occurs when controlling for variables that are consequences rather than causes. By statistically adjusting for such "post-treatment" variables, the model removes part of the causal pathway and distorts the estimated effects.

5. Overfitting: Adding too many predictors can lead the model to capture random noise rather than a meaningful signal. While this may improve apparent fit to the training data, it reduces generalizability and produces misleading conclusions about true underlying relationships.

Together, these examples illustrate how the choice of statistical model can create, obscure, or distort relationships in the data, underscoring the importance of careful model construction and interpretation.

Many of these pitfalls arise when statistical models are specified without sufficient theoretical grounding. A theoretical model that is explicitly structured and easily translated into statistical terms can help avoid such problems by clarifying causal pathways, identifying confounders, and constraining arbitrary choices in model specification [34]. Thus, it is advantageous to have a theoretical model that can be easily translated into a statistical model. Such models constrain the freedom of interpretation, provide structure that enhances comparability between studies, and facilitate the design of experiments [103]. As I present in Chapter 1, many existing QoE theoretical models aim to provide taxonomies or comprehensive descriptions of the phenomenon. However, there is still a lack of theoretical models that are sufficiently simple and structured to be directly translated into statistical models. To address this gap, in [102], together with colleagues, we proposed a theoretical model based on a Directed Acyclic Graph, designed to facilitate its straightforward application in future statistical studies.

3.2 Graph-Based Approaches to Modeling

Modern statistical methods allow for representing the assumption of the model in the form of a graph. Two common approaches are structural equation model (SEM) diagrams and causal directed acyclic graphs (DAGs). SEM diagrams are both conceptual and statistical tools: they depict hypothesized relationships, including latent variables, and are then formally tested against observed data. In contrast, causal DAGs are purely conceptual tools for clarifying causal assumptions, identifying confounders, and guiding analytic strategies. While visually similar, SEM diagrams focus on modeling associations, whereas DAGs are explicitly designed for causal inference [106]. One of the biggest strengths of the DAG approach is that the authors provide a clear description of when and how to control variables in statistical models [34]. This is directly relevant to the problems outlined by McElreath [128]: DAGs explicitly address spurious associations by revealing confounding paths that must be blocked, and they prevent post-treatment bias by showing when a variable is a descendant of the treatment and thus inappropriate for adjustment. However, issues such as multicollinearity and overfitting remain outside the scope of DAGs, as they concern statistical estimation rather than causal identification. Masked relationships, while sometimes clarified through a causal graph, are not the primary focus of this approach. Thus, DAGs provide a principled framework for avoiding two of the most consequential sources of model misspecification—spurious associations and post-treatment bias. This makes them a particularly useful foundation for structuring QoE IFs research.

A Directed Acyclic Graph (DAG) is constructed by representing variables as nodes and drawing arrows to denote assumed direct causal influences between them, ensuring that no cycles are formed. Building such a graph relies on domain knowledge to decide which variables to include and how they are connected, making the DAG an explicit way of formulating and presenting hypotheses about the underlying causal structure. Crucially, the absence of an arrow represents a stronger assumption than its presence, since it states that no direct causal link exists between two variables, whereas an arrow merely hypothesizes the possibility of such an effect [142]. In other words, when constructing a graph, arrows should be drawn between all nodes unless there is a strong reason to assume the absence of a causal link. So in practice,

the process of formulating such a graph is to try to get rid of as many arrows as possible, by looking for justification for the lack of a causal link, to make the graph useful and acyclic.

Thus, to be able to build DAG, the number of variables in the model has to be balanced. As I showed in Chapter 1, current QoE models were not built for that purpose. The trade-off between the complexity of models and the amount of information they provide is one of the key problems researchers must address when building a model. A good model predicts as much as possible with as few assumptions and variables as possible [166, 60] . This property of the model is called parsimony. Only the most important variables must be determined to achieve parsimony. Parsimony is also directly linked to the overfitting problem mentioned above [129]. Probably one of the strongest advocates for simpler models was George Box [21].

Thus, the theoretical model presented in this chapter consists only of the most important Influential Factors (IFs) obtained in the questionnaire. The idea is to provide a minimal model that describes QoE. This model can then be extended and verified by adding variables in experiments [103]. To enable practical application across diverse research scenarios, a theoretical model should be formulated at an abstract level [60]. In other words, it should capture broader mechanisms or phenomena rather than narrowly defined variables, which makes the model generalizable. Therefore, the described model contains general components, with operationalization and examples of measurements that allow for adaptation to a specific research scenario.

Below, I present the most important factors influencing VoD QoE from Chapter 2 in the form of a structural graph using principles from DAG. This part is direct quote from [102].

3.3 Proposed Video QoE Model Based on the Path Diagram

We based our model on the interpretation of the first part of the general definition of QoE: “Quality of Experience is the degree of delight or annoyance of the user of an application or service.” For this reason, we did not represent QoE as a separate unit in the Figure 3.1. Instead, we included a “delight or annoyance” unit as an outcome of the interaction of variables typical for video service experience. In real life, it is hard to imagine a scenario in which the delight or annoyance of the user is not generated by the video content. This result is backed by the outcomes from Chapter 2. Depending on internet efficacy, this content-dependent experience might be moderated by drops in video quality. With this reasoning, we built a path model of QoE.

Following previous QoE models, we distinguished Influential Factors IFs as predictors of general user satisfaction. Moreover, we described these variables as latent variables and provided examples of their measurements. Below, I briefly present the operationalization of each variable in our model and their connections, quoting the paper [102].

3.3.1 Components Operationalization

QoS

Quality of Service (QoS) is a measure of the overall performance of a network. It is often used to describe the ability of a network to provide a consistent level of service to its users. ITU-T defines QoS as *“The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied*

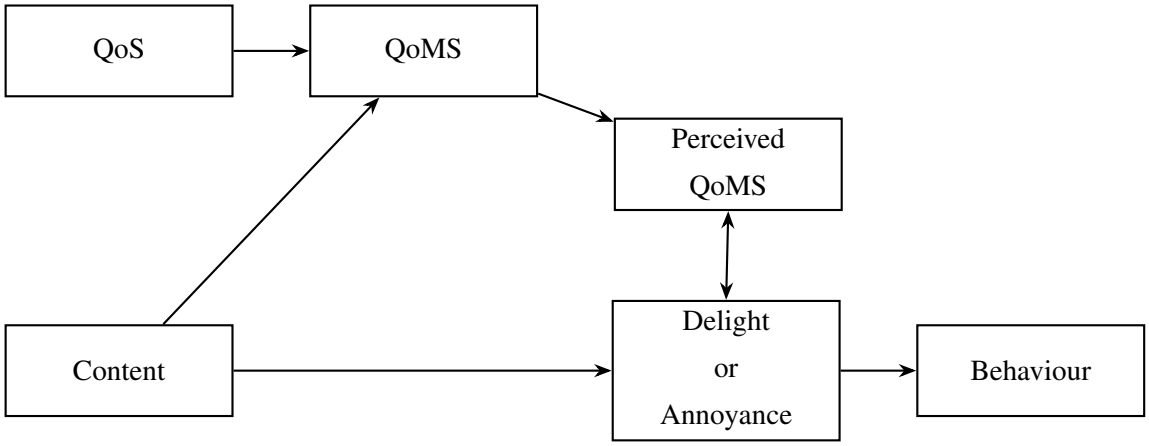

Figure 3.1: Theoretical path model for video QoE. Conceptual model links QoS, content, perceived QoMS, affect, and behaviour.

needs of the user of the service" [153]. QoS can be expressed in terms of performance indicators such as throughput, latency, jitter, and packet loss. Special metrics for QoS assessment are constantly being developed. There are also models describing the influence of QoS on QoE [56, 31, 97]. QoS is a well-known QoE IF. Particularly, real-world scenarios emphasize the primacy of instantaneous throughput and latency for multimedia services, where fluctuating network conditions are common [165]. These factors should be prioritized in QoE theoretical models for their pronounced impact on the streaming experience.

Content

Content characteristics are multidimensional. Depending on the research question, different variables might be taken into account to analyze their influence. Due to its complex character, content can be identified as a system or a context IF (see 1.1). For example, characteristics such as motion, number of details, brightness, and computation complexity are well recognized in QoE studies [171, 43, 96]. They are categorized as system factors influencing QoE. On the other hand, the content type chosen by the user can be classified as a context IF and can be used for predicting user satisfaction [8]. Moreover, influential service providers such as YouTube have their own metrics for classifying and quantifying content. One of the crucial statistics for distinguishing videos in the service is "engagement," operationalized as the mean percentage of watching time. It can be used for the operationalization of the content such as perceived engagement or interest in the video. Furthermore, self-description methods can be used for assessment of participant level of interest [169] or motivation [99]. Moreover, there are standardized data sets of visual stimuli that can be used to evaluate the influence of emotions evoked by content[126],[42], [16].

Quality of Multimedia Signal

In our model, the quality of the multimedia signal represents the objective properties of visual stimuli. In the context of video streaming, it is the video displayed on the user's device. There are many methods for assessing the quality of video, some of which are objective and some subjective. Thus, we propose a division into the objective quality of media signal (QoMS) and its perceptual dimensions – the perceived

QoMS described below. In this approach, QoMS is the IF that represents the quality of reproduction of the source signal (content). It has properties of the source content that are moderated by the efficacy of the network and user hardware. It can be assessed with objective metrics e.g. signal-to-noise ratio [62, 61]. In current models [52, 151, 160], QoMS is described as the physical representation of the signal.

Perceived Quality of Multimedia Signal

We use the term "perceived QoMS" (PQoMS) to emphasize the role of perception in video QoE studies where subjective assessments of quality made by users are in the spotlight. Typically, researchers estimate QoMS with a 5-point Absolute Category Rating (ACR) scale [86]. Additionally, there is a new effort to build matrices that take into account both QoMS and PQoMS to predict user satisfaction [120]. In previous models [151, 52], these descriptions are the outcome of the quality formation process.

Degree of Delight or Annoyance

We assume that the user's state of delight or annoyance (DoA) is the outcome of both quality and content properties. According to the general definition, this is, in fact, the measure of QoE. It can be assessed, for example, with an adapted Differential Emotions Scale [40]. In the natural context, both technical [134] and content [109], related factors may lead to a change in the QoE. This might cause a shift in user behavior [160].

Behavior

Depending on the scope of the study, behavior might be a short-term reaction to quality-related events [159, 108, 58], habit evaluation [158], or even consumer attitude [154, 145] predictors. Generally in the commercial context, pleasure and arousal are good predictors of approach–avoidance behavior [95]. As long-term behavior toward network providers might be influenced by a set of additional important variables (e.g., pricing), we focus on short-term behavior. In our model, behavior is an outcome of DoA and can be observed as a change in interaction with service (e.g., change of the video).

3.3.2 Relations Between Variables and Model Assumptions

We described our generalized QoE model in the form of a diagram (see Figure 3.1). Path model of video QoE with causal paths between variables. Following the causal path analysis approach, arrows represent the assumption that variable A may be a direct cause of B. For example, in Figure 3.1, DoA might be caused by perceived QoMS and content. Additionally, arrow representation does not imply that the process is simple. Relationships between A and B can be complex, multi-staged, and nonlinear, what we assume is the direction of the relationship. Moreover, by drawing an arrow from A to B, we do not assume that A is the only thing that causes B. In fact, this approach assumes that every variable can be influenced by unspecified factors not represented in a graph. This enables including measurement error and the influence of unknown factors in the model, as well as adding new variables in future updates of the model. However, if we know that units in the model have a common cause, this must be expressed in the graph.

On the other hand, the absence of an arrow represents a stronger assumption, namely, the lack of relationship between variables represented by nodes [143]. For example in Figure 3.1. Path model of video QoE, the amount of delight or irritation does not change the QoMS; as such, the direction of influence seems impossible.

In our model, we described the physical representation of the video signal (QoMS) as an outcome of the interaction between network efficacy (QoS) and Content. We assumed that from the user's perspective, there is no relationship between QoS and Content other than QoMS. In other words, if network efficacy has some influence on video content, it can only be observed via the QoMS. In consequence, QoS cannot directly influence DoA or behavior. Users must see the change in QoMS to conclude that there are network efficiency problems. Furthermore, the arrow from QoMS to Perceived QoMS represents human perception. In fact, this process could be described using quality formation models such as [52, 151, 131]. Moreover, it could be influenced by cognitive factors such as visual sensitivity [14]. In addition, we cannot assume that the perception of quality is not moderated by DoA. That is why the arrow between PQoMS and DoA is bidirectional.

Most importantly, we assume that in real-life scenarios, DoA is a function of both content and PQoMS. Users might react differently to the same QoMS depending on the type of content.

Finally, users' behavior is the consequence of that DoA [165, 72]. One can be dissatisfied both due to the content and the PQoMS. This can result in behavior directed to enhance network efficacy, change of content, compensation, or abandonment of activity.

The model described above is general. This means that we treat it as a framework that can be specified for particular use cases and experimental setups. As already mentioned, in such cases, the model can be extended with additional variables. This procedure requires the operationalization and inclusion of new units in the graph. For example, if we want to add participant interest to the model, we can place it on the arrow from content to DoA. In that case, we assume that interest is caused by content but not by other variables. Moreover, content can only influence the general satisfaction by the level of interest. If one's hypothesis is that interest does not always influence the causal path between content and DoA, interest should be included with an additional path. Another hypothesis might be that the perceived QoMS is influenced by visual sensitivity. In such a scenario, we must add visual sensitivity as a new unit outside the graph and draw the line to the perceived QoMS. Other influential factors from the model ([53]) could be added in a similar manner.

3.4 Theoretical Implications and Taxonomy

Further discussion is beyond the scope of the [102] paper. The above-described model can be used not only for study design and statistical inference but also has strong theoretical implications for the domain. Based on this path structure, we can propose an operational definition of QoE limited to video services: „QoE is the amount of behaviorally relevant user Delight or Annoyance toward a video service evoked by content and moderated by the perceived quality of the video." This definition covers not only units but also the causal relations. In this form, the definition might be useful for further development of the domain.

Moreover, this theoretical structure can provide a taxonomy of QoE models and methods in the form of layers. Table 3.1 presents this approach.

While influencing factors accumulate across layers and continue to shape outcomes, the measurements and metrics remain distinct for each stage, reflecting the specificity of how QoE is assessed at different levels. The table, therefore, makes visible both the strong foundations of the field and current frontiers and needs for development. This clear distinction might help to fill the gap between QoE theory and practice I was describing in the Introduction Chapter 1.7.

Model Unit	**Metrics Examples**	**Measurements Examples**	**Influencing Factors**
Quality of Service		Network-level KPIs [74]	QoS
Quality of Multimedia Signal	PSNR	Full reference analysis of signal	
	SSIM	Full reference analysis of signal aligned with HVS	Human Visual System
Perceived Quality of Multimedia Signal	Metrics trained to align with subjective test (e.g. VMAF)	Scales like ACR or perceptual studies like JND	Properties of content*
		AccAnn scale	Expectations
Delight or annoyance		Emotional measurement e.g., : Self Assessment Manikins (SAM) Scales [67]	Human and Context IFs
Short term behavior		Engagement measurement [158]	UX properties of service
Long term behavior		User satisfaction surveys [4]	Economical factors and alternative availability

Table 3.1: Overview of model units, metrics, measurements, and influencing factors in the form of layers. *To be verified in the next two chapters.

Moreover, Table 3.1 provides a clear and unambiguous framework for defining the scope of a QoE study. For instance, when researchers introduce a new quality metric, they can explicitly situate it within the model and specify that their work addresses the Perceived Quality of Multimedia Signal layer of QoE. Similarly, if a metric aims to predict emotional responses, researchers can position their model at the Delight or Annoyance layer. In addition, the table offers explicit guidance on which Influencing Factors should be considered at each level, helping to align study design with the theoretical structure. Such consistent terminology helps to prevent ambiguity and may also strengthen bibliographic analyses in the field.

Some of these distinctions have also resonated with ongoing community work. The forthcoming VQEG White Paper on Quality of Experience-Aware Management for Collaboration Between Network and Application Providers [184] adopts a layered perspective that partially overlaps with the approach presented here. While the white paper proposes its own framework, certain elements were informed by discussions in which I took part as a contributor, reflecting a broader recognition of the value of clearer distinctions within QoE research.

3.5 Conclusions

As I showed above, it is possible to represent the outcome of an exploration study in a relatively simple and parsimonious diagram. Nevertheless, the model presented in Figure 3.1 has one crucial drawback. We can not exclude the possibility that Percived Quality of Multimedia Signal is influenced by Delight or Annoyance. Thus, the relation between those units remains bidirectional. This loop makes our model cyclic. Due to that, the working hypothesis for this chapter is not met. Thus, the general second hypothesis of this thesis, that QoE can be conceptualized with a minimalistic model, is only partially supported by this chapter. To face this problem, the relation between PQoMS and DoA needs to be verified in an experimental manner. The next two chapters are focused on this problem.

Nevertheless, this chapter highlights an important issue with a multifactor approach to QoE. With each factor added to the QoE model, it is necessary to consider all causal relations associated with this variable. As I show above, path diagrams can be a useful tool for both inference and operationalization. This contributes to Objective Four of this dissertation. The recommendation itself was already presented at QoMEX 2023 [103].

Chapter 4

Study 2: Influence of Emotionally Evoking Content on ACR Scores

Parts of the material presented in this chapter are currently being prepared for submission as a research article, and therefore, I would like to acknowledge the contributions of all co-authors. The authors contributing to this part are Kamil Koniuch, Lucjan Janowski, Michał Wierzchoń, Mikołaj Leszczuk, and Katrien De Moor. All authors agreed for their work to be included in the dissertation. Author's contributions were as follows: KK: Conceptualization, Methodology, Formal analysis, Writing – original draft; LJ: Supervision, Funding acquisition, Formal analysis, Writing – review and editing; MW: Conceptualization, Formal analysis, Supervision, Writing – review and editing; ML: Software, Data Investigation, Writing – review; KDM: Writing – review and editing.

This and the following chapter describe two experiments conducted to fulfill the Third Objective of this dissertation. Both studies are based on the theoretical model introduced in the previous chapter and aim to verify the Influential Factors identified in Chapter 2, in the context of applicable QoE tests. In this chapter, I examine whether emotionally evocative content can influence QoE ratings, in line with Hypothesis Four.

The Absolute Category Rating (ACR) scale is used as the primary method for measuring QoE. The objective is to determine whether factors affecting user satisfaction can alter scores in a standardized QoE assessment. In other words, the study aims to evaluate whether the outcomes of the Influential Factors questionnaire have practical implications for everyday QoE practice.

4.1 Introduction

Perception of stimuli is influenced by their emotional evocativeness [25]. Emotionally charged content has been shown to affect various cognitive processes, including perception [176]. Consequently, perceived video quality is also likely modulated by the viewer's emotional state, making emotion an important factor to consider in quality assessment studies. As it was stated in the Quality of Experience Advanced Concepts, Applications and Methods book [162]: *"The most obvious reason for QoE researchers to pay attention*

to the emotional content of the material they are presenting is that it may affect the quality assessment.”. The authors note that the direction of the effect remains unknown and propose using the emotional stimuli datasets in future studies. However, I found no image or video QoE studies using controlled emotional content.

Studies outside the QoE community utilize standardized datasets with measured, controlled emotional impact and they do this for a similar purpose (e.g.[112, 111, 110, 192, 38, 126, 156, 194, 130, 186]). The key method used for the evaluation of emotional influence is using databases with emotionally evoking content (e.g., images, videos). They are created by gathering a broad scope of stimuli, e.g., pictures that may trigger a broad scope of emotions. Then, the stimuli are presented to participants who rate different aspects of their emotional response to the stimuli. Finally, the database is published with both the stimuli and the averaged emotional response for each stimulus.

In this research, I used one of those databases, namely the Nencki Affective Picture System (NAPS) [126]. NAPS was created by selecting 1,356 high-quality, realistic photographs categorized into five content groups: people, faces, animals, objects, and landscapes. These images were rated by 204 participants along three emotional dimensions: valence, arousal, and approach–avoidance; using continuous slider scales. The final database includes both the stimuli and normative emotional ratings, as well as physical image properties such as luminance and contrast. This database has a wide range of applications from neuroscience [107] to sexology [194]. Similar databases were published for other types of stimuli: text [22], sound [195], and video [7].

As noticed, the book Quality of Experience: Advanced Concepts, Applications and Methods [162] does not propose the direction of effect of using such datasets. Moreover, the empirical evidence discussed in the introduction to this thesis highlights this ambiguity: while some studies suggest that content interest or desirability can increase perceived quality [89, 105], others report no significant effect [190]. This inconsistency indicates that the influence of emotionally evoking content on QoE cannot be assumed to be either positive or negative in a straightforward manner.

The General Hypothesis, for this study, postulates that emotionally evoking content will have a strong influence on ACR ratings. For the purpose of this study, I derive two working hypotheses. As the above-mentioned studies and theory do not provide a clear prediction of influence, the first hypothesis is non-directional: in the model, quality rating will be statistically predicted by approach-avoidance, valence, and arousal of the stimulus. Two, this effect will, on average, change the Mean Opinion Score at least by 0.3 points.

4.2 Method

Path Model

For the purpose of this study, the model introduced in Chapter 3 was adapted (see Figure 4.1). As I mentioned, the theoretical model described above is generalized. Thus, the original framework has to be adjust by replacing model units from 3.1 with variables that will be tested in this chapter.

The Content unit is represented by NAPS, the name of the dataset used in this experiment. Instead of QoS and QoMS, the participants' devices and the method of quality manipulation are indicated. In this study, ACR is used as a measure of the Perceived Quality of the Multimedia Signal. The images were subjected to distortions based on a Hypothetical Reference Circuit (HRC), with seven predefined quality levels determined using the FovVideoVDP metric [125].

The model explicitly represents the variables and their causal relations. The red arrow illustrates the influence of emotionally evocative content on participants' emotional state. Since emotions were not directly measured in this study, the corresponding unit is marked in gray. The relationship between the FovVideoVDP metric and ACR scores can be inferred from previous work (e.g., [124]). Accordingly, Hypothesis Four is illustrated as an arrow from Emotions to ACR. Thus, in this experiment, I investigated whether emotionally evoking content would change ACR scores. Based on [162], it is not possible to assume the direction of this influence. The minimal effect that should be observed would be an increase in the variance of the scores depending on the emotions evoked by the content.

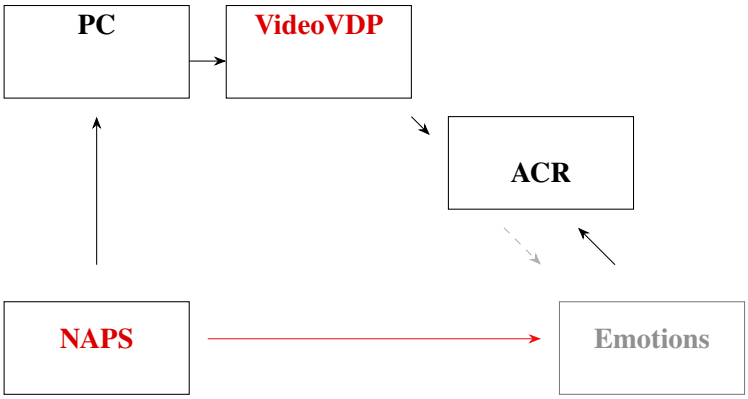

Figure 4.1: This causal graph shows the design of the experiment in line with the theoretical framework presented in the chapter 3. Red text represents variables. Arrows represent the causal influence. The red arrow depicts emotional manipulation. Gray color represents unobserved variables and influences.

Participants and Sampling

122 participants were recruited through the crowdsourcing platform "Prolific". The sample size followed the norms of the ITU Recommendations [85]. Gender balance was emphasized and achieved for a wider demographic representation. The study adhered to ethical guidelines approved by the university ethics committee (KE/10_2021).

The participants who provided demographic information had a mean age of 26.4 years (SD = 6.5, range 20–53). Gender distribution among those who disclosed it was nearly equal, with 51% identifying as female and 49% as male. The majority of respondents who reported their ethnicity identified as White (86%), with smaller proportions identifying as Mixed (6%), Asian (4%), or Black (4%). More than half of those who provided data were students (58%), and nearly half reported full-time employment (48%), while others were unemployed (17%), in part-time jobs (11%), or in other categories. Nationality data revealed that the largest groups came from Portugal (33%) and Poland (29%), followed by smaller groups from South Africa,

Italy, Spain, Hungary, Mexico, Greece, and several other countries, reflecting a diverse but Europe-centered sample.

Design of the Experiment

Participants accessed the experiment using desktop PCs or laptops, evaluating image quality in a standard browsing environment. As stimuli, images from the “people” category of the Nencki Affective Picture System (NAPS) were used [126]. This database is a widely recognized resource in affective science, with close to a thousand citations, and provides high-quality, naturalistic photographs standardized for experimental use. The “people” category contains pictures of living, injured, or deceased human bodies, as well as isolated body parts, excluding close-up facial expressions. Each image has been normatively rated on three separate emotional dimensions—valence, arousal, and approach–avoidance—ensuring a controlled and validated selection of stimuli for psychological research. In the authors’ framework, valence reflects the degree of positivity or negativity, arousal captures the level of excitement or calmness, and approach–avoidance describes the motivational tendency to engage with or withdraw from the stimulus. The images were processed using a hypothetical reference circuit (HRC) procedure to introduce controlled distortions. Seven distinct quality levels (HRC A–G) were created, with gradations determined according to the FovVideoVDP metric [125], which provides a perceptually validated measure of visual quality differences.

Variables and Measurements

Besides the ACR scale, control questions were integrated into the experiment. This procedure was implemented to enhance the robustness of the data. These questions, based on randomly selected images, asked participants about the image content. Each was a simple yes-or-no question, e.g., "Did the photo show a market?" In total, there were 20 such questions.

Procedure

Before beginning the experiment, participants were required to complete a screen quality test. This step was crucial to ensure that variations in monitor quality did not affect the experiment’s outcomes. The results of the screen test, including various parameters such as screen brightness and contrast levels, were recorded and linked to each participant’s data.

The voluntary nature of participation was emphasized, and participants were informed of their right to withdraw at any time. This ethical consideration was vital, especially given the potentially sensitive nature of the images used in the study. Before the experiment, participants were warned about the possible emotional impact of images.

The first five images shown to participants were from the categories ‘animals’ and ‘landscapes’, serving as an introductory phase to familiarise them with the process. The main part of the study then involved the presentation of 250 images from the ’people’ category. After viewing each image, participants provided their quality ratings, and the experiment was concluded once all images were assessed. This process was designed to take approximately 30 minutes per participant.

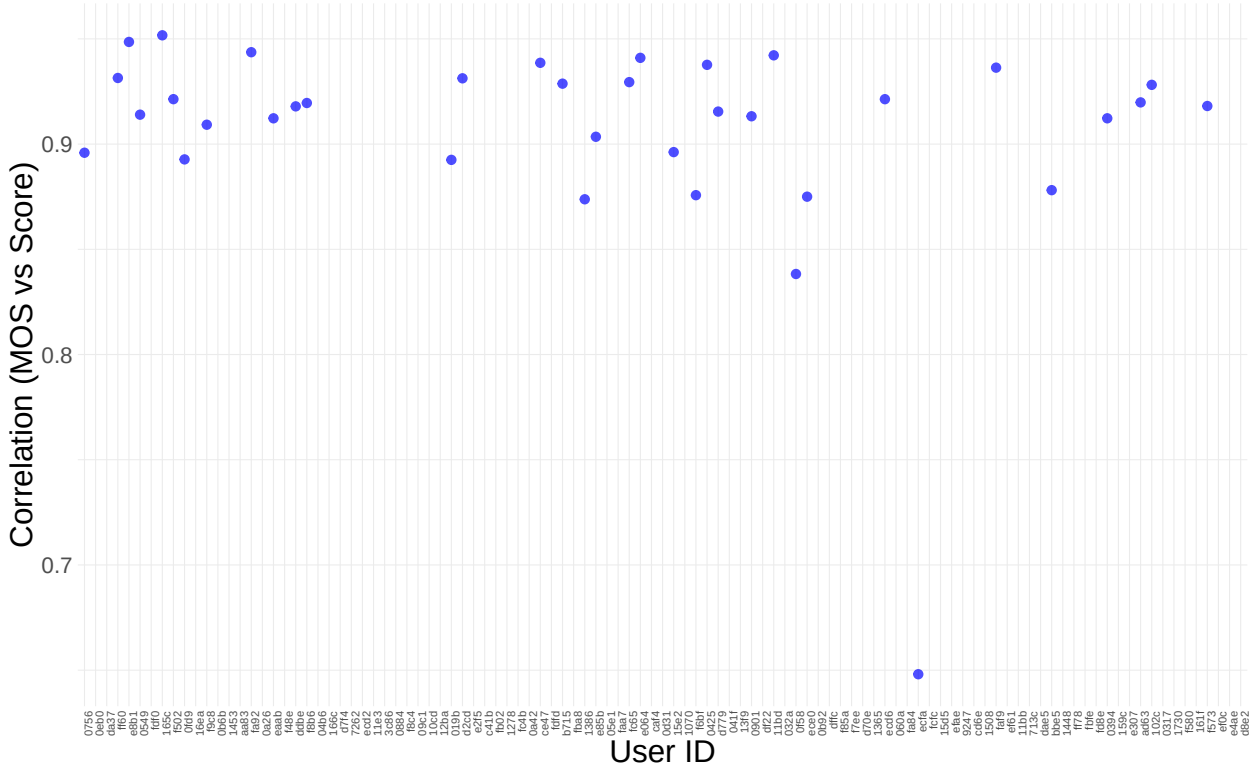

Figure 4.2: Correlation between participants' answers.

4.3 Results

4.3.1 Data Cleaning

Initial analyses showed that 86.07 % of the participants completed the experiment. For further analysis, participants who gave at least 200 votes out of 250 were included. In total, 17 participants did not provide the required minimum of 200 votes, and their results were discarded in the further analyses.

Furthermore, the general correlation between subjects is high (see Figure 4.2). ITU-T recommendation for classical subjective experiments calls for removing testers with a correlation lower than 0.75. In this experiment, a lower threshold of 0.7 was used, since the experiment is more difficult, due to the emotionally evoking content. Moreover, there was a difference in the pull of stimuli between participants. Therefore, a lower correlation was expected. With a threshold of 0.7, only 4 participants were removed. As a result, further analyses were run on a sample of 101 participants.

Finally, responses to the trapping questions were examined to identify inattentive participants. As previously noted, 20 such questions were included for this purpose. Prior to the analysis, a threshold of a maximum of three errors was established, accounting for the possibility that poor image quality could lower the response accuracy. All participants met this criterion, making no more than three mistakes, and were therefore retained for further analysis.

Coefficient	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	4.235973	0.160644	26.369	$< 2 \times 10^{-16}$
hrcB	-0.531805	0.043655	-12.182	$< 2 \times 10^{-16}$
hrcC	-1.150593	0.043612	-26.383	$< 2 \times 10^{-16}$
hrcD	-1.716192	0.043655	-39.312	$< 2 \times 10^{-16}$
hrcE	-2.186445	0.043655	-50.084	$< 2 \times 10^{-16}$
hrcF	-2.625516	0.043612	-60.202	$< 2 \times 10^{-16}$
hrcG	-3.063554	0.043612	-70.246	$< 2 \times 10^{-16}$
av_ap	0.208015	0.028535	7.290	4.69×10^{-13}
arousal	-0.006058	0.019230	-0.315	0.753
valence	-0.182118	0.028376	-6.418	1.78×10^{-10}

Table 4.1: Regression coefficients for the model predicting MOS based on quality levels (HRC) and approach–avoidance, valence, and arousal values from NAPS.

4.3.2 Cumulative Influence of Emotional Dimensions on ACR Scores

To verify the working hypothesis for this study, a simple regression model was investigated first. The model was specified in `R` as:

```
model_psych_all <- lm(mos ~ hrc + av_ap + arousal + valence, data = mos_data)
```

which corresponds to the following linear formulation:

$$\text{MOS}_i = \beta_0 + \beta_1 \cdot \text{HRC}_i + \beta_2 \cdot \text{AV\_AP}_i + \beta_3 \cdot \text{Arousal}_i + \beta_4 \cdot \text{Valence}_i + \varepsilon_i, \tag{4.1}$$

where β_0 is the intercept, $\beta_1, \beta_2, \beta_3, \beta_4$ are regression coefficients, and ε_i is the error term. In this specification, HRC represents quality levels A–G.

Results from this analysis are presented in Table 4.1. While approach–avoidance and valence showed statistically significant effects, the directions of the effects were opposite. This finding is difficult to interpret because the lowest score on the valence scale indicates unpleasantness, whereas the lowest score on the avoidance scale indicates that the picture elicits a repulsive response. Therefore, the influence of these variables on MOS would be expected to align in the same direction. Thus, a simpler model, consisting only of statistical factors, was tested.

Thus, a simpler model, consisting only of statistical factors, was tested. Thus, a simpler model, consisting only of statistical factors, was tested (equation 4.3.2).

The model was specified in `R` as:

```
model_psych_ap_av_valence <- lm(mos ~ hrc + av_ap + valence, data = mos_data)
```

which corresponds to the following linear formulation:

$$\text{MOS}_i = \beta_0 + \beta_1 \cdot \text{HRC}_i + \beta_2 \cdot \text{AV\_AP}_i + \beta_3 \cdot \text{Valence}_i + \varepsilon_i \tag{4.2}$$

where β_0 is the intercept, $\beta_1, \beta_2, \beta_3$ are regression coefficients, and ε_i is the error term. In this specification, HRC again represents the coding condition and is treated as a categorical factor with quality levels A–G.

Coefficient	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	4.18741	0.04518	92.689	$< 2 \times 10^{-16}$
hrcB	-0.53183	0.04364	-12.186	$< 2 \times 10^{-16}$
hrcC	-1.15061	0.04360	-26.390	$< 2 \times 10^{-16}$
hrcD	-1.71620	0.04364	-39.323	$< 2 \times 10^{-16}$
hrcE	-2.18648	0.04364	-50.098	$< 2 \times 10^{-16}$
hrcF	-2.62554	0.04360	-60.218	$< 2 \times 10^{-16}$
hrcG	-3.06357	0.04360	-70.265	$< 2 \times 10^{-16}$
av_ap	0.20766	0.02850	7.285	4.86×10^{-13}
valence	-0.17873	0.02626	-6.807	1.37×10^{-11}

Table 4.2: Regression coefficients for simpler model predicting MOS based on quality levels (HRC) and approach–avoidance and valence values from NAPS.

Table 4.2 shows that this inconsistency in the direction of influence remains even in a simpler model. At this point, one of the remaining explanations for this effect is multicollinearity [128]; therefore, the database was investigated in terms of correlation.

4.3.3 Correlation Analysis of Emotional Dimensions

To exclude the possibility of multicollinearity, emotional dimensions from the NAPS dataset were examined. Relating to each other was checked. The results show very high correlations: valence and approach–avoidance are correlated at 0.973, arousal and valence at -0.848, while arousal and approach–avoidance are correlated at -0.820. These strong relationships raise concerns. From a theoretical point of view, such high correlations may indicate that the variables are not truly independent and may measure overlapping aspects of emotional response. From a statistical perspective, including all of them in the same model could cause multicollinearity, which reduces the reliability of the model estimates and increases the chance of drawing misleading conclusions. This suggests the need to rethink how these variables are used in both our theoretical framework and statistical modeling. Figure 4.3 presents these correlations. Thus, the first working hypothesis of this chapter, postulating that in the model quality rating would be statistically predicted by approach–avoidance, valence, and arousal of the stimulus, cannot be tested by a single model including all variables. Thus, in the next sections, those variables are analyzed separately.

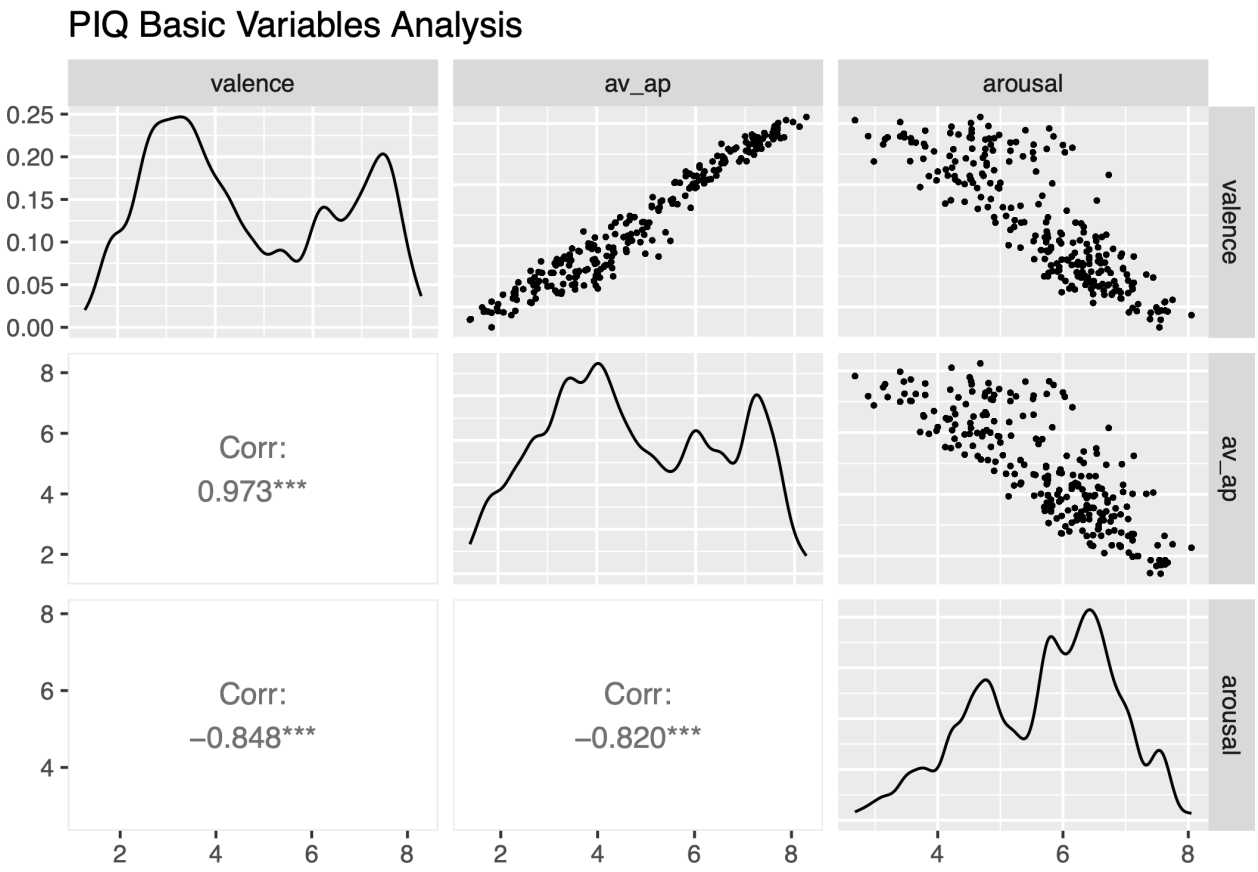

Figure 4.3: Correlation analysis for the faces subset of the NAPS database.

4.3.4 Separated Influence of Emotional Dimensions on ACR Scores

For each emotional dimension, a series of linear mixed-effects models was used to test whether emotional dimension scores predicted subjective quality ratings across different quality conditions (HRC levels). Separate models were fitted for each group, with user ID included as a random intercept to account for individual differences. Additionally, linear trends were visualized, and confidence intervals were computed to assess the consistency and strength of the effect.

Approach–avoidance

As shown in Table 4.4 and Figure 4.5, the results indicate that the influence of approach–avoidance on ACR scores is limited to the highest quality levels. Significant positive effects were observed only in HRC groups A, B, and C. In lower quality conditions, the slopes were close to zero and not statistically significant, suggesting that emotional response plays a negligible role when technical degradation dominates user perception.

Moreover, two types of mixed-effects models were compared: a *Baseline Model* and an *Extended Model* that included the approach–avoidance (`av_ap`) variable. The Baseline Model estimated ACR scores as a function of the quality level (`hrc`) with a random intercept for each user, whereas the Extended Model additionally included the predictor `av_ap`:

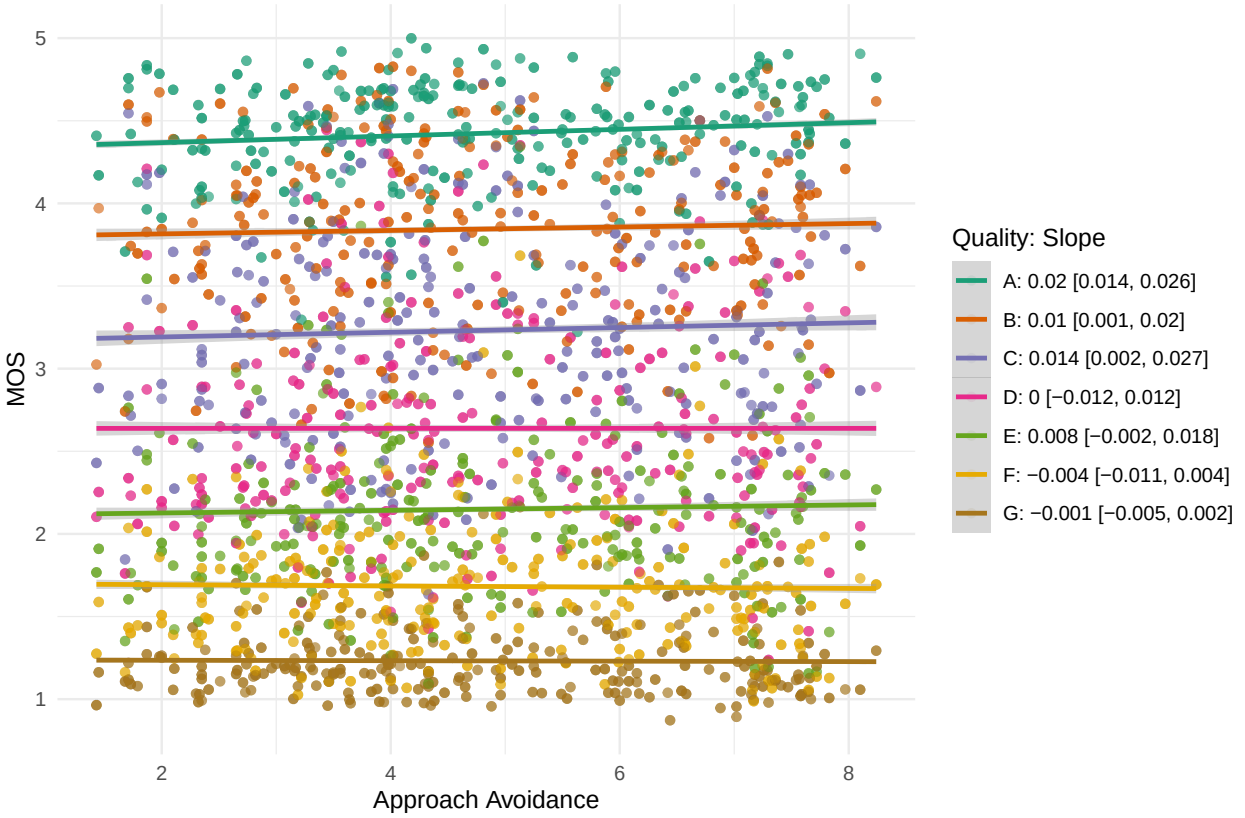

Figure 4.4: Linear regression lines showing the relationship between approach–avoidance scores and MOS across HRC groups. Statistically significant positive trends are observed only in the highest quality conditions (A–C).

Table 4.3: Effect of approach–avoidance scores on MOS across HRC levels. Statistically significant slopes are marked in bold.

HRC	Slope	95% CI	p-value	Significant
A	**0.020**	[0.014, 0.026]	1.62×10^{-11}	Yes
B	**0.010**	[0.001, 0.020]	0.038	Yes
C	**0.014**	[0.002, 0.027]	0.026	Yes
D	0.000	[-0.012, 0.012]	0.997	No
E	0.008	[-0.002, 0.018]	0.103	No
F	-0.004	[-0.011, 0.004]	0.320	No
G	-0.001	[-0.005, 0.002]	0.487	No

$$\text{score}_{ij} = \beta_0 + \beta_1(\text{hrc}_{ij}) + \beta_2(\text{av\_ap}_{ij}) + u_{0j} + \varepsilon_{ij} \tag{4.3}$$

where $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ represents the random intercept for user j, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ denotes the residual error term.

Adding the approach–avoidance variable improved model fit, with the AIC decreasing from 57640.0 to 57628.3, resulting in a ΔAIC of 11.7 in favor of the extended model. This improvement was statistically significant, as confirmed by a likelihood ratio test ($\chi^2(1) = 13.68$, $p = 0.0002$). Although the effect of `av_ap`

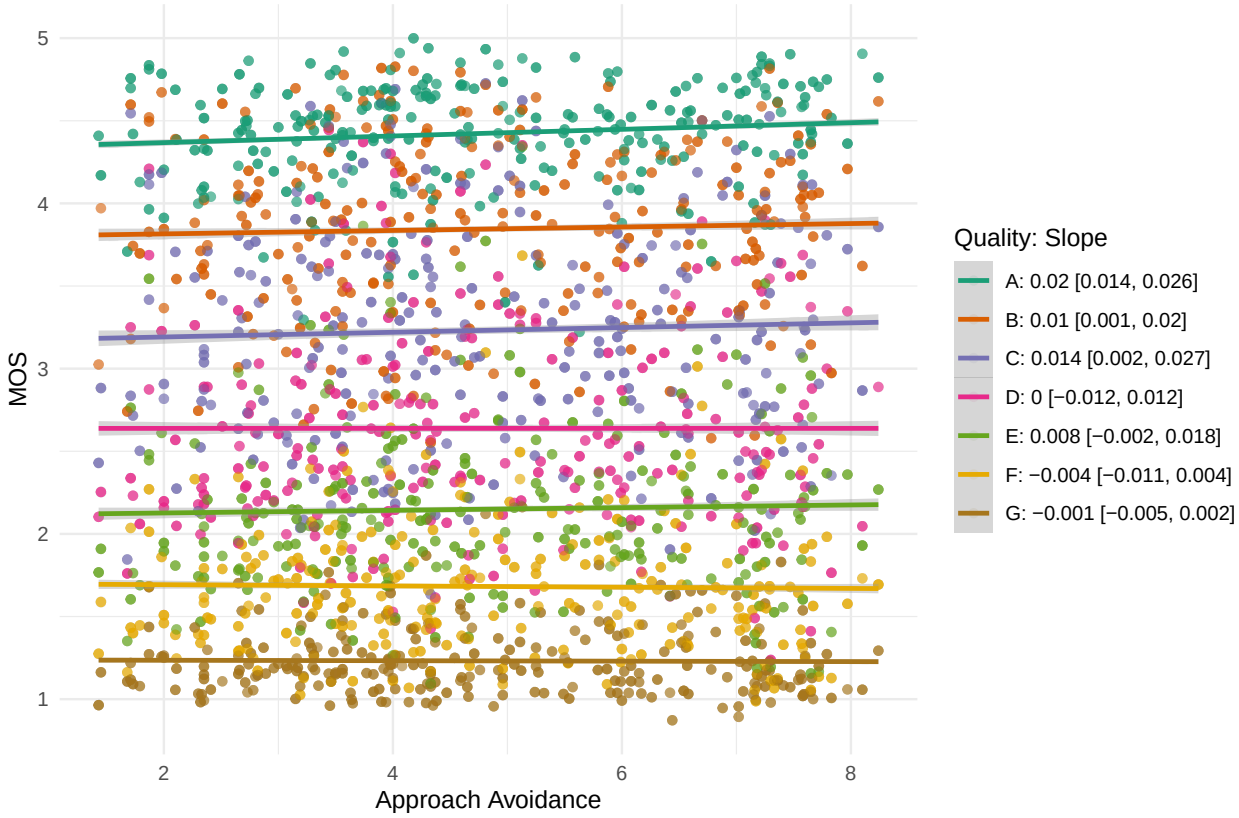

Figure 4.5: Linear regression lines showing the relationship between approach–avoidance scores and MOS across HRC groups. Statistically significant positive trends are observed only in the highest quality conditions (A–C).

Table 4.4: Effect of approach–avoidance scores on MOS across HRC levels. Statistically significant slopes are marked in bold.

HRC	Slope	95% CI	p-value	Significant
A	**0.020**	[0.014, 0.026]	1.62×10^{-11}	Yes
B	**0.010**	[0.001, 0.020]	0.038	Yes
C	**0.014**	[0.002, 0.027]	0.026	Yes
D	0.000	[-0.012, 0.012]	0.997	No
E	0.008	[-0.002, 0.018]	0.103	No
F	-0.004	[-0.011, 0.004]	0.320	No
G	-0.001	[-0.005, 0.002]	0.487	No

was statistically robust, its practical impact was modest: the estimated coefficient was 0.00994, meaning that even across the full range of `av_ap` (1–10), the predicted increase in ACR scores was less than 0.10 points on a 5-point scale. This suggests that approach–avoidance tendencies have a real but limited influence on perceived quality across quality conditions.

Valence

In Table 4.5 and Figure 4.6, the influence of valence on ACR scores for each HRC is presented. Results show that valence had a significantly positive effect on HRC A and a negative effect on HRC D, F, and G. Thus, the influence of valence is inconsistent across HRC levels. Moreover, the slopes are up to -0.020, which means quite a small effect.

Table 4.5: Effect of valence scores on MOS across HRC levels. Statistically significant slopes are marked in bold.

HRC	Slope	95% CI	p-value	Significant
A	**0.016**	[0.010, 0.021]	8.36×10^{-9}	Yes
B	0.000	[-0.009, 0.009]	0.989	No
C	-0.004	[-0.015, 0.008]	0.540	No
D	**-0.020**	[-0.031, -0.009]	0.00028	Yes
E	-0.004	[-0.013, 0.005]	0.415	No
F	**-0.010**	[-0.017, -0.004]	0.00291	Yes
G	**-0.005**	[-0.008, -0.001]	0.0124	Yes

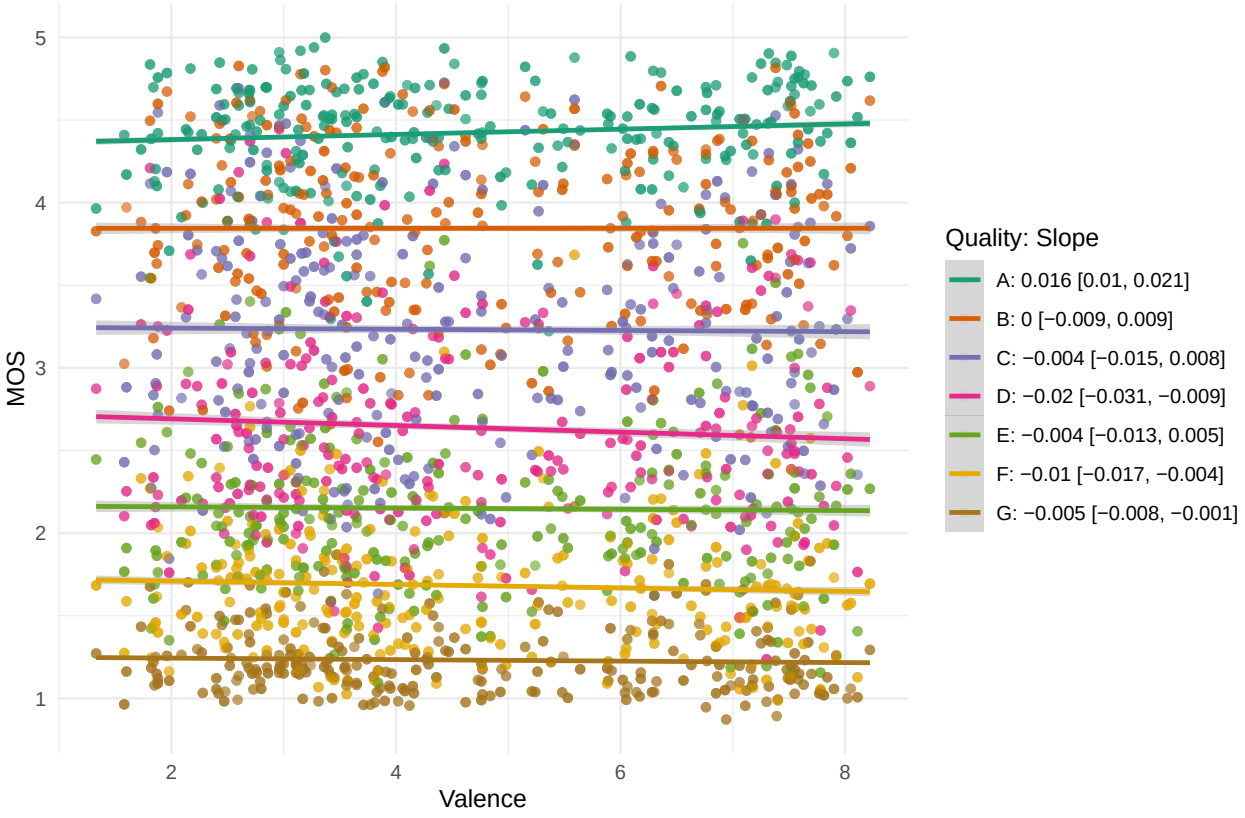

Figure 4.6: Linear regression lines showing the relationship between valence scores and MOS across HRC groups. Statistically significant positive trends are observed only at the highest quality A. Negative effects can be observed for quality D, F, and G.

Similarly to approach–avoidance, two mixed-effects models were estimated: the *Baseline Model* and the *Extended Model* that included the `valence` variable. The Extended Model can be formally expressed as:

$$\text{score}_{ij} = \beta_0 + \beta_1(\text{hrc}_{ij}) + \beta_2(\text{valence}_{ij}) + u_{0j} + \varepsilon_{ij}$$

where $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ represents the random intercept for user j, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ denotes the residual error term. In this case, the addition of the `valence` variable did not improve model fit: the AIC increased slightly from 57640.0 to 57641.8, yielding a ΔAIC of -1.8, which indicates a worse fit. The likelihood ratio test further confirmed that the difference between models was not statistically significant ($\chi^2(1) = 0.2$, $p = 0.675$). The estimated effect of `valence` was very close to zero (–0.00104), and thus, even across the full range of `valence` (1–10), the predicted change in ACR scores would be negligible and statistically indistinguishable from noise. This suggests that valence had no measurable influence on perceived video quality in this dataset when estimated for all HRC levels. Valence showed significant effects in a few individual HRCs (e.g., A, D, F), but these effects did not generalize when modeled across all quality levels simultaneously.

Arousal

In Table 4.6 and Figure 4.7, the relationship between arousal and ACR scores is shown separately for each HRC. The results indicate a significant positive association for HRC B and D, while a negative effect is observed for HRC A. This suggests that the impact of valence varies depending on the HRC. Additionally, the effect sizes are relatively small, with slopes reaching up to -0.033.

Table 4.6: Effect of arousal scores on MOS across HRC levels. Statistically significant slopes are marked in bold.

HRC	Slope	95% CI	p-value	Significant
A	**-0.033**	[-0.042, -0.024]	5.70×10^{-13}	Yes
B	**0.028**	[0.013, 0.043]	0.00033	Yes
C	0.009	[-0.011, 0.028]	0.383	No
D	**0.022**	[0.004, 0.040]	0.0162	Yes
E	-0.003	[-0.018, 0.013]	0.741	No
F	0.004	[-0.007, 0.016]	0.448	No
G	0.005	[-0.001, 0.011]	0.0854	No

Similarly to before, two types of mixed-effects models were compared: the *Baseline Model* and the *Extended Model* that included the `arousal` variable. The Extended Model can be expressed as:

$$\text{score}_{ij} = \beta_0 + \beta_1(\text{hrc}_{ij}) + \beta_2(\text{arousal}_{ij}) + u_{0j} + \varepsilon_{ij}$$

where $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ represents the random intercept for user j, and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ denotes the residual error term. where $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ is the user-specific random intercept and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual error.

Adding the `arousal` variable did not improve model fit, with the AIC increasing from 57640.0 to 57641.8, resulting in a ΔAIC of -1.8 in favor of the baseline model. The improvement was not statistically

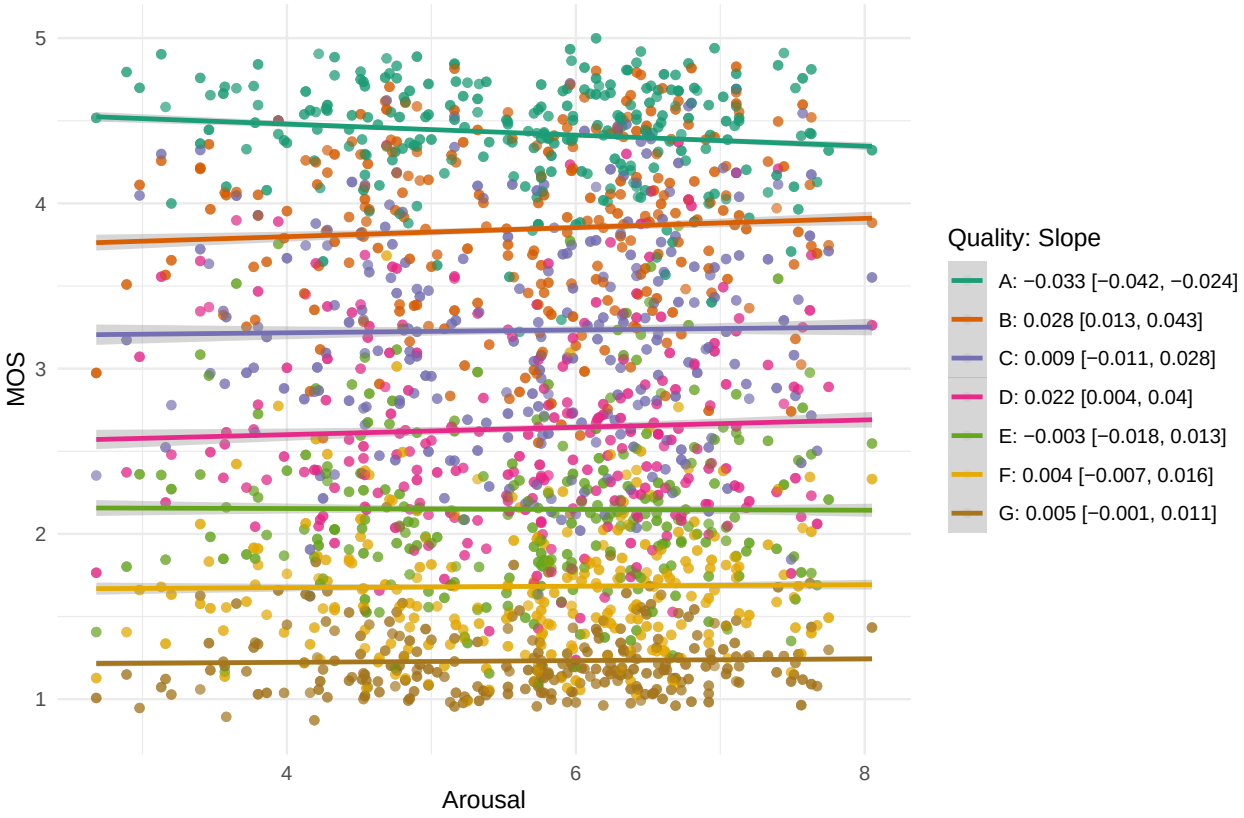

Figure 4.7: Linear regression lines showing the relationship between arousal scores and MOS across HRC groups. Statistically significant positive trends are observed for quality B and D. A Negative effect was observed for quality A.

significant, as confirmed by a likelihood ratio test ($\chi^2(1) = 0.20$, $p = 0.736$). The effect size was negligible: the fixed-effect estimate for `arousal` was 0.0014, implying that even across the full range of `arousal` (1–10), the predicted change in ACR score would be less than 0.014 points on a 5-point scale. This suggests that arousal had no practical or statistical impact on the perceived quality ratings.

4.3.5 Qualitative Comments

At the end of the experiment, participants had the option to leave a comment about the study. In total, 26 participants decided to comment on the study. 2 comments were positive: „all fine", and „thank you". 3 participants commented that images were loading too long. One participant commented on the quality and the understanding of the content, probably in response to trapping questions: „Majority of the images had fair to good quality but you could still make out what the image is depicting even with the poor and bad images." The rest 20 comments were very strong and described the content and how images influenced participants' emotions. As one participant stated „that was so graphic some of the most horrific images I've seen".

Among those 20 answers about the content influence on emotion, one participant directly described its influence on scoring. „I noticed something in my evaluation of the materials. If the image showed something

more disturbing, I would have difficulty to attibute the value good to the quality of the material, because I would be associating the image with the bad thing. Despite, I did not let that influence my response".

4.4 Discussion

The working hypothesis one: in the model, quality rating will be statistically predicted by approach-avoidance, valence, and arousal of the stimulus was tested in series of separate models. As the effect is inconsistent over quality levels and emotional dimensions, I find this hypothesis as not supported by data. Moreover, those effects remain below 0.3 change in MOS, thus secound working hypothesis is not supported by data either.

Overall, the results are against the Hypothesis Four, which claims that emotionally evoking content will have an influence on ACR scores. While there are some statistically significant results for single HRC levels and for approach–avoidance, they are inconsistent across emotional dimensions and relatively small. Based on the modeling approach–avoidance, and arousal on single HRC levels, one may argue that the effect of emotions is significant for the highest qualities. But the same effect is not observed for valence. Moreover, for arousal, the direction of change is different between HRC A and HRC B. Thus, there is no consistent effect that could be backed by this data.

What is more, adding psychological variables only improved the model for approach–avoidance. This fact might be explained by arguing that this emotional dimension is the closest to actual human behaviour. Nevertheless, the overall effects of approach avoidance are lower than expected; see 4.5. In the NAPS dataset, the images with the lowest and highest approach avoidance are drastically different. To generate high avoidance, strong images that contain dead bodies, surgical operations, traffic accidents, and violence are used. The high approach is guaranteed by the application of smiling faces and landscapes. These results show that changing gore pictures to smiling faces will provide up to 0.2 MOS shifts, only for the best quality. Moreover, when all qualities are counted in one model, the estimated coefficient was 0.00994, meaning that even across the full range of avoidance (1–10), the predicted increase in ACR scores was less than 0.10 points on a 5-point scale.

One may argue that those small effects are still important. But it is worth mentioning that such variability in emotional evocativeness is rarely present in QoE studies. Therefore, the stimuli used in this study differ substantially from those typically employed in the development and validation of quality metrics. Despite this, a model based solely on HRC levels mapped by FovVideoVDP was able to predict 81.6% of the variance in subjective scores collected by crowd-sourcing. This result highlights the robustness of modern quality metrics, demonstrating their ability to predict perceptual quality even under emotionally variable and non-standard content conditions.

The lack of effect might be caused by the mechanism described by one of the participants in the comment at the end of the experiment. Although there is a tendency to rate negatively associated content with lower scores, participants were able to overcome this in the final rating. This shows that rating quality is an easy task for participants. They can make good quality assessments even when constantly exposed to emotional manipulation. This result is against the common consensus that participants in QoE studies are prone to be biased by emotions [162].

Moreover, this outcome is also contradictory to the questionnaire results described in Chapter 2. This result may be due to the fact that, in the questionnaire, participants were asked what influenced their satisfaction, whereas ACR scale scores appear to be robust against variations in emotional state. What is more, participants in the questionnaire study recalled memories where they did not have a specific task in mind to rate the quality, and they watched a video in a natural context. To face one of these limitations, in the next chapter, I present a study where the context was designed to be as close as possible to everyday video service usage.

Further conclusions and recommendations for future studies are described in chapter 6, backed by result from the next chapter.

4.4.1 Limitations

In the discussed experiment, the crowd-sourced task was not completed by 17 participants (13,93% of the total sample). Although a significant sample of participants was retained for the presented analysis, there remains a risk that too much information may have been lost due to dropout.

In the study [55], the authors compared different crowd-source platforms. In their experiments % of dropouts in Prolific was within the range between 2.5% and 5. 5%. Both experiments carried out by the authors were shorter than the experiment in this chapter. On the other hand, in a more recent study [187], the dropout of prolific users reached 26%. These big differences lead to the conclusion that dropout on the Prolific platform might be caused by the type of task. Unfortunately, most QoE studies using the Prolific platform are not providing information about dropout % [10, 45, 198, 65, 138, 9, 122, 177, 191]. I found only one study where one out of 40 participants was ruled out due to incompleteness [45]. However, it was not a typical ACR QoE study, and the total duration of the experiment is not clearly stated.

Thus, it is impossible to exclude the possibility that some participants resign from the study due to the drastic nature of some pictures in the NAPS database. This effect might potentially skew our results. However, overall data quality was high, since only four participants were removed due to low correlation with others (below the 0.7 threshold), and all remaining participants were positively verified by the trapping questions, making no more than three mistakes. This suggests that most subjects provided reliable and attentive responses throughout the experiment.

This experiment was conducted in conditions with relatively low mundane realism. Participants were rating 250 pictures from the NAPS dataset on the ACR scale. While providing control on IFs, this design is very far from real-world applications of video QoE metrics. In realistic scenarios, users are watching professionally produced movies or TV series on available services. As this difference might potentially limit external validity, in the next chapter I will present an experiment with greater mundane realism.

Chapter 5

Study 3: Social Influence on ACR Scores

Parts of the material presented in this chapter are currently being prepared for submission as a research article, and therefore, I would like to acknowledge the contributions of all co-authors. The authors contributing to this part are Kamil Koniuch, Lucjan Janowski, Katrien De Moor, Michał Wierzchoń, Mikołaj Leszczuk, Rafał Figlus, and Mateusz Zduński. All authors agreed for their work to be included in the dissertation. Author's contributions were as follows: KK: Conceptualization, Methodology, Formal analysis, Writing – original draft, Investigation; LJ: Supervision, Funding acquisition, Formal analysis, Writing – review and editing; KDM: Writing – review and editing; MW: Supervision, Writing – review and editing; ML: Writing – review and editing; RF: Software; MZ: Investigation.

This chapter continues to fulfill the Third Objective of this dissertation. In this case, the influence of social presence is investigated following Hypothesis Five. This hypothesis claims that ACR scores will be influenced by social presence. In this chapter, social presence is manipulated by comparing scores between conditions, watching alone and with a second participant. Thus, the working hypothesis for this chapter is that the relation between VMAF and MOS will be statistically significantly different between conditions alone and together.

As before, the theoretical model presented in Chapter 3 serves as the tool for conceptualizing the experiment. In the previous chapter, the absence of a clear effect may be attributed to the use of a strictly controlled experimental design; therefore, the present study adopts a more naturalistic setup that mimics everyday video usage. The ACR scale is employed once again to provide comparability.

5.1 Introduction

The presence of other people might be an important QoE Influential Factor [53]. In psychology, the process of assigning cognitive focus in social settings is called social attention. In their review article [173], the authors explored how social attention influences perceptions, experiences, and behaviors in the presence of others. They argue that being observed increases arousal, self-awareness, and concern for reputation, which can lead to improved performance on simple tasks and more desirable social behavior. Even subtle cues of

being watched, such as images of the eyes, can unconsciously trigger behavioral changes. Moreover, social attention can alter subjective experiences by intensifying emotions and perceptions through mechanisms of shared reality [173]. These findings highlight the deep cognitive and motivational impact of social presence on the user experience. This influence extends to the realm of media consumption, where the presence of others can alter the way video content is perceived, potentially affecting viewers' assessment of video quality.

In a 2015 QoE study [201], researchers explored the impact of the presence of others on various aspects of QoE, such as enjoyment, satisfaction, and perceived video quality. The research revealed that watching videos in a group can improve enjoyment and satisfaction, suggesting a significant social influence on the overall viewing experience. However, the study does not find a substantial impact of social context on the perceived technical quality of the video. What is important is the fact that perceived quality was measured with the ACR scale. However, this study has some limitations. Above all, it provides only manipulation of quality on two-bit rate levels: high (2000 kbps) and low (600 kbps). Moreover, it used multiple QoE-related questions, which might have biased the participants' answers. In the experiment described in this chapter, I tried to face those limitations. The goal was to determine whether these results were caused by insufficient quality manipulation or if ACR is indeed robust against social influence.

5.2 Method

5.2.1 Path Model

For the purpose of this study, the model introduced in Chapter 3 was adapted (see Figure 5.1). The Content unit is represented by Netflix, indicating the streaming platform utilized in this experiment. Instead of QoS and QoMS, participants' viewing device (TV) and the method of quality assessment (VMAF metric [120]) are explicitly represented. As before, the Absolute Category Rating (ACR) scale is utilized to measure the Perceived Quality of the Multimedia Signal. Additionally, Social Presence is included to illustrate the experiment's social viewing context. This unit is operationalized as two conditions in the experiment: watching alone and watching together.

The adapted model explicitly depicts the variables and their causal relations. The green arrow illustrates the experimental manipulation. Hypothesis Five is represented as an arrow from Social Presence to ACR, examining whether social presence modifies perceived multimedia quality measured on the ACR scale. As participants could watch any TV series from Netflix's portfolio, the influence of content on the emotional state of participants cannot be excluded. This influence is represented by the black arrow at the bottom of the graph. Since emotions were not directly measured, the corresponding unit is marked in gray. Moreover, the relationship between the VMAF metric and ACR scores can be inferred from previous research (e.g., [120]).

Participants and Sampling

In this study, 24 couples (48 individuals in total) were recruited through Facebook advertisements. Participants were asked to take part together with an already known partner (e.g., a friend or a relative), with the aim of enhancing mundane realism [17]. The mean age of participants was 23.3 years (SD = 6.4, range

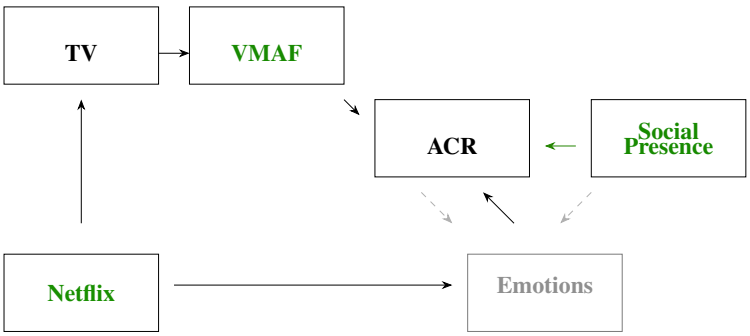

Figure 5.1: This causal graph shows the design of the experiment in line with the theoretical framework presented in chapter 3. Green text and arrows represent variables and manipulations specific to WWF. Arrows represent causal influences. Gray boxes and arrows indicate unobserved variables and latent paths.

18–55). Gender distribution was relatively balanced among those who disclosed it, with 54% identifying as female and 40% as male, while 6% preferred not to report their gender.

Design of the Experiment

During this experiment, participants watched full episodes of TV series available on Netflix, which they chose themselves. The experiment took place in a lab setup, but one that was furnished to feel like a real living room. A sofa was placed in front of a TV on a stand. The sofa was firm, so using a mouse on it was easy, and there was also a small table available for the mouse. The lights were dim, but the lighting was still good enough to see clearly, which matches the BT.500 lab standards [83]. Gray curtains were used around the room. Subjects could adjust the TV volume but not the TV brightness. The viewing distance from the TV was the standard 3H. The setup of the experiment is presented in Figure 5.2.

Figure 5.2: Laboratory viewing room.

The self-selected Netflix series was played using a Chrome extension [66] that performed the following functions:

1. Change the bitrate according to a preset script, covering all available bitrates.

2. It paused the movie every 2.5 minutes to ask the viewer to rate the video quality.

3. It saved a VMAF score every second.

Variables and Measurements

During episode playback, every 2.5 minutes, with a few seconds of jitter, the movie was automatically paused, and a pop-up appeared with the standard 5-point Absolute Category Rating (ACR) scale. The subject could no longer see the paused movie, only the question, “Please rate the quality,” along with a scale from “Excellent” at the top to “Bad” at the bottom. Both the question and the scale were written in Polish. After selecting an answer, the movie was automatically resumed.

Because the Chrome extension relied on the standard compression options produced by Netflix, the available bitrates differed among the chosen series. For example, the worst quality could be VMAF = 60 for one series and VMAF = 30 for another.

Procedure

Before the experiment, the participants watched the recorded standardized instructions. Each couple could choose a TV series from the Netflix portfolio. The requirement was to choose a series in which each episode lasted between 40 and 70 minutes. There were two conditions in this experiment: alone and together. In each condition, the participants watched one episode of the same series. The order of the conditions was randomly assigned between couples. Thus, half of the couples begin with the alone condition and half with the together condition.

5.3 Results

For each subjective rating, the VMAF values were aggregated in the playback segment prior to the response. This window spanned from the video start or previous rating to the timestamp marking the start of the current rating. The resulting mean VMAF per segment reflects the quality experienced during the actual viewing leading up to each score.

Data Cleaning

As limited scale usage may indicate noncompliance or a lack of understanding of the task, participants who used fewer than three distinct values on the ACR scale were marked as outliers in the study. For each participant, the distribution of scores across sessions was computed and visualized separately for the "alone" and "together" conditions.

Figure 5.3 shows the percentage of ACR scores used by each participant in both conditions. Although most of the participants used a wide range of scales, some showed restricted response patterns. Following standard practice, subjects 100, 119, and 222 were marked as outliers due to limited scale usage.

To account for the fact that participants in the "together" condition were exposed to a different TV series than those in the "alone" condition, there is a need for a comparison of both the objective video

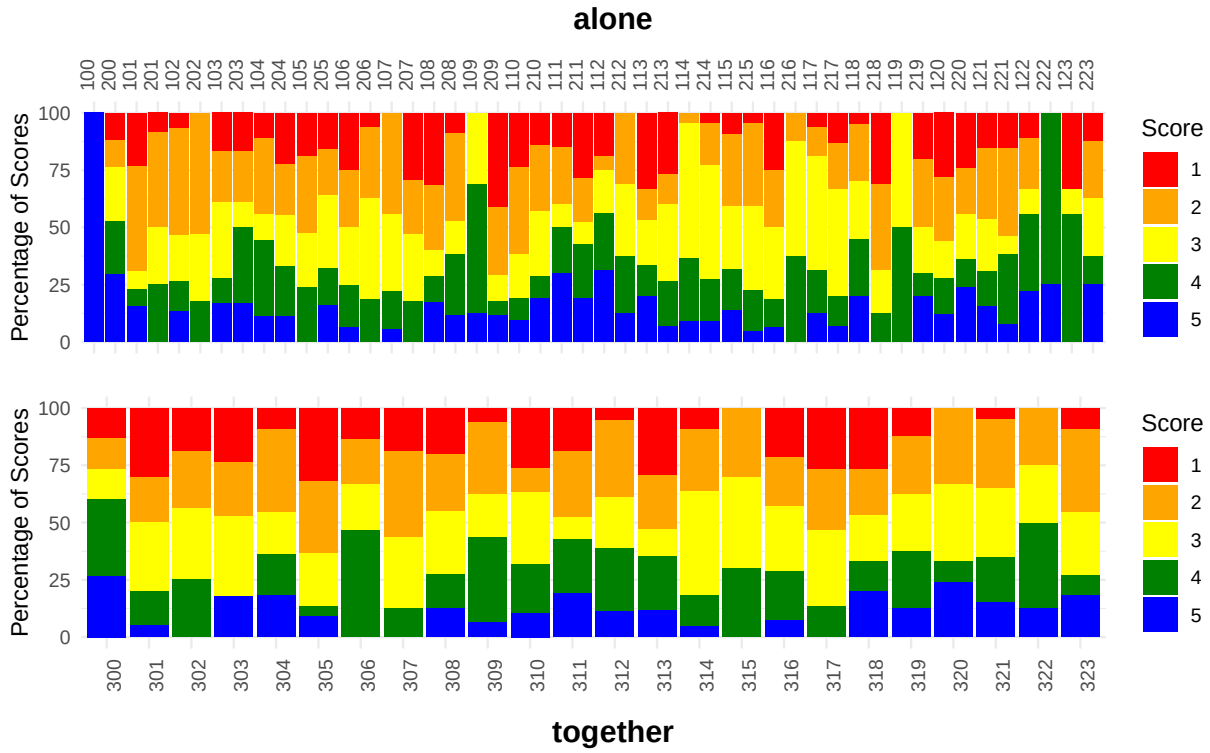

Figure 5.3: Distribution of ACR scores for each participant in the "alone" (top) and "together" (bottom) conditions. Each bar represents a single subject, with color-coded segments indicating the proportion of ratings from 1 (red) to 5 (blue). Participant IDs reflect individual or paired sessions. IDs below 300 indicate individual participants, while IDs 300 and above correspond to subjects from paired sessions. E.g., pair 300 is 100 and 200 individuals together. Participants 100, 119, and 222 were flagged as outliers due to limited use of the rating scale.

quality (VMAF) and the subjective ratings (MOS) across conditions. As shown in Figure 5.4, the distribution of VMAF scores is generally similar under all conditions, although slight changes are visible. Figure 5.5 shows that subjective scores also followed comparable trends, with minor differences in scores proportions, suggesting that the difference in content did not drastically alter perceived quality.

The Influence of Social Presence on ACR Scores

Figure 5.6 illustrates the distribution of VMAF scores across subjective MOS levels, separately for the "alone" and "together" conditions. Each pair of violins corresponds to a specific MOS rating (1 to 5), allowing a comparison of objective video quality (VMAF) associated with each subjective score in different social viewing contexts.

Figure 5.7 shows the relationship between objective video quality (VMAF) and subjective mean opinion score (MOS) under the two viewing conditions ("alone" and "together"). Each point represents a group average, with the color intensity indicating the sample size. In both conditions, a strong positive correlation is observed between VMAF and MOS (alone: r = 0.94; together: r = 0.93), indicating that the perceived quality of the participants aligns well with objective metrics (in this case, VMAF). Linearity tests (Harvtest p = 0.7) confirm the robustness of this linear relationship.

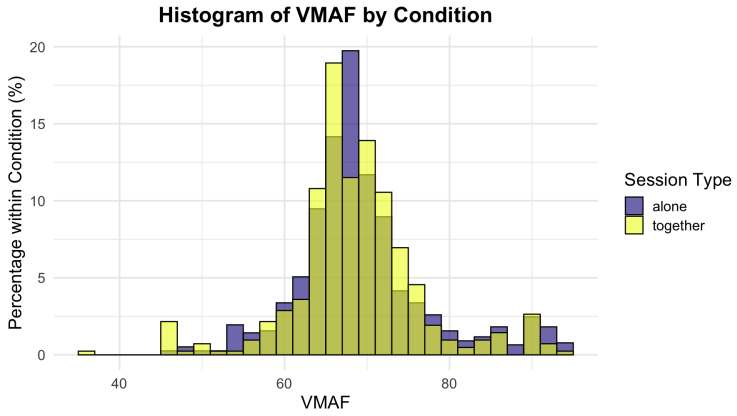

Figure 5.4: Histogram of VMAF scores by session type.

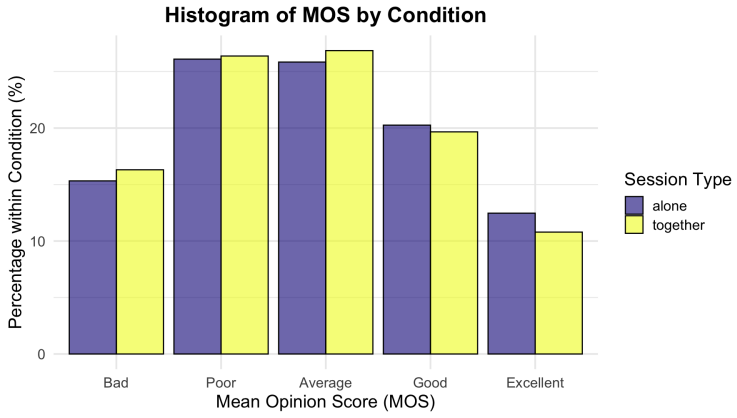

Figure 5.5: Histogram of MOS ratings by session type.

To complement the visual analysis, several statistical models were estimated. A linear regression using aggregated data showed that the mean VMAF strongly predicts subjective quality (MOS), which is approximately 88% of the variance ($\beta \approx 0.05$ per VMAF unit, $p < 0.001$). Including the social condition (alone vs. together) did not significantly improve the model ($\Delta R^2 \approx 0$, $p \approx 0.91$), indicating that the relationship between objective (VMAF) and subjective (MOS) quality is consistent between viewing conditions. Logistic regression analysis further confirmed that the viewing condition could not be reliably predicted from VMAF or MOS scores, reinforcing the finding that the VMAF–MOS relationship holds regardless of social context.

5.4 Discussion

During the initial screening, 3 individual participants were marked as outliers. As Figure 5.3 presents, participants 100, 119, and 222 used fewer than 3 scores on the ACR scale. These same participants, when paired (pairs 300, 319, and 322), used the full scale. It is important to note that the order of the conditions alone and together was randomized. Half of the participants began with the together condition, and half with the alone condition. All excluded participants started with the condition alone. This is important because, in addition to the instruction at the beginning of the experiment, the participants did not receive specific training. This data suggests that scoring quality with the second person might be easier, especially at the beginning of the experiment.

Figure 5.4 presents the distribution of VMAF between conditions. This histogram shows that our manipulation was correct and that our software provided a normal distribution of objective quality between TV series. Thus, exposure to quality was comparable between cases. Figure 5.5 shows that this manipulation

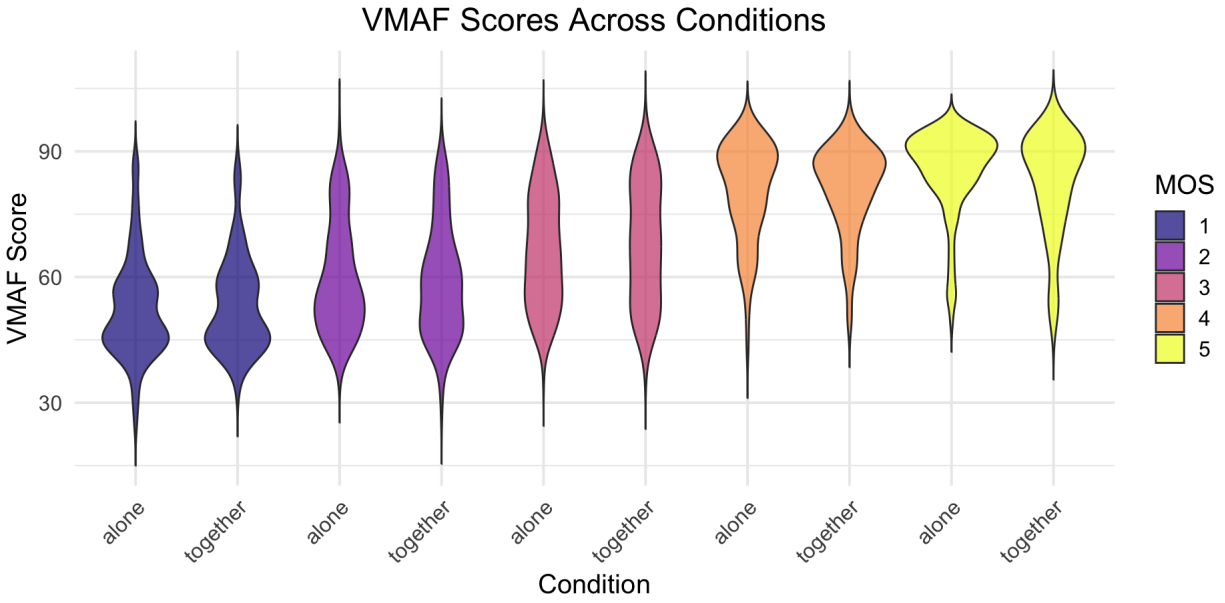

Figure 5.6: Distribution of VMAF scores by Mean Opinion Score (MOS) levels and condition. Each pair of violins corresponds to a subjective MOS level (1–5) in the "alone" and "together" viewing conditions. Higher MOS ratings are associated with higher VMAF scores, consistently across both conditions.

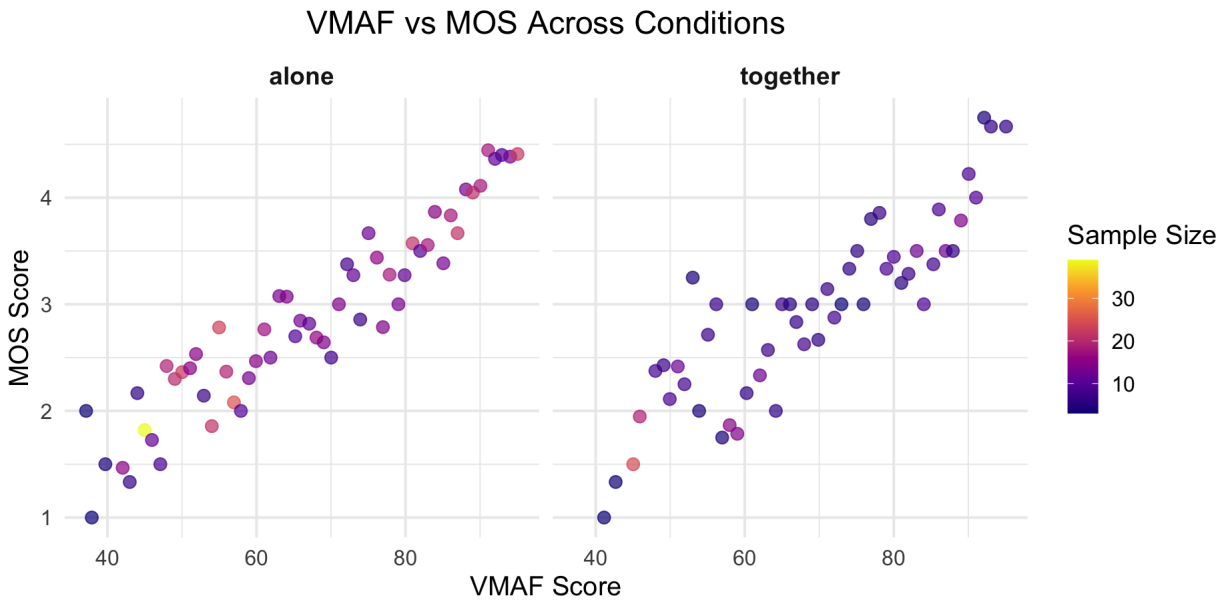

Figure 5.7: Relationship between objective video quality (VMAF) and subjective Mean Opinion Score (MOS) across the "alone" and "together" conditions. VMAF values were binned by rounding to the nearest integer, and the corresponding MOS was computed as the average score within each bin. Only bins with more than two responses were included. Strong linear correlations were observed in both conditions (Alone: $r = 0.94$, Together: $r = 0.93$).

provided comparable results between cases. In fact, we can see that the differences between conditions are very small.

Figure 5.6 allows for better comparison between conditions. Surprisingly, on this graph, there are no significant changes in voting patterns between conditions. This observation is supported by linear models, which did not show a statistically significant influence of social presence on ACR scores. The dominant factor, which explains 88% of the variance in both cases, is VMAF. Figure 5.7 shows an extremely strong correlation between VMAF and MOS under both conditions.

This result is especially important because it shows a strong external validity of VMAF, at least for scores gathered with the ACR scale. In the experiment, participants were watching a full episode of a series on Netflix, under two social conditions, scoring the quality in 2.5-minute intervals. These conditions are different from the data that were used for the development and evaluation of VMAF. However, VMAF predicted 88% of the variance in our dataset and reached 93-94% of correlation with MOS. This shows how robust modern metrics and the ACR scale are against external factors.

Similarly, to study from the previous chapter, Hypothesis Five is not backed by the data. Working hypothesis assumed that the relation between VMAF and MOS will be statistically significantly different between conditions alone and together. This effect was not only not present in model comparison, but also Figures 5.6 and 5.7 present no change in the distribution of scores between conditions. Compared to the previous study, mundane realism was increased, and this explanation of null results can be ruled out. It seems that even if factors are reported as influential in everyday experience, the usage of the ACR scale together with quality manipulation can hide this effect. This conclusion can be drawn, thanks to the usage of the

same theoretical model to conceptualize comparable experiments, with diverse levels of variable control. This inside has several consequences that I will further describe in the next chapter.

Chapter 6

General Discussion

6.1 Objectives and Hypotheses

Objective One: Identify the most important Influential Factors for QoE using the memory recall procedure. This objective was fulfilled by running 3 separate studies on a total sample of 140 participants described in chapter 2. As a memory recall procedure, improved participants' understanding of questions and leveraged the Cronbach's alpha scores, I would argue that the most important scores can be found after the memory recall procedure. In the figure 2.11, the most important Influential Factors can be found aligned with statistical tests. The method provided clear distinctions between IFs depending on the type of service. Due to that, I found Hypothesis One: the influential Factors can be measured and taxonomized by a memory recall-based questionnaire is supported.

Objective Two: Design a parsimonious QoE model based on a minimal number of assumptions. This objective was achieved with the work described in Chapter 3. This model was built on insights provided by the memory recall questionnaire. It is also a strong argument for Hypothesis Two: main Influential Factors can be combined in a simple model explaining their relationship. Further, Chapters 4 and 5 showed that the arrow from Delight or Annoyance to Perceived Quality of Multimedia Signal can be ruled out. Thanks to this result, the model no longer contains a loop, which makes it easier to use for causal inference. This is another argument in favor of Hypothesis Two.

Objective Three: Design a series of original studies based on one theoretical model for the evaluation of Influential Factors. This objective was met by studies described in Chapters 4 and 5. Both studies can provide more general conclusions thanks to the comparability leveraged by the application of the model described in chapter 3. As those studies provide strong arguments for the robustness of the ACR scale on both Influential Factors, this can support Hypothesis Three: The proposed model can help in planning comparable experiments targeting Influential Factors one by one.

Hypothesis Four: Content has a strong influence on QoE ratings, and Hypothesis Five Social presence has an influence on QoE ratings are not supported by the data from ACR-based experiments.

Objective Four: Provide recommendations for future QoE studies. Based on the findings described in this work, strong recommendations can be provided for future studies. Below in highlight them in Section 6.3. Thanks to that, I Objective Four is met.

6.2 Main Findings and Contributions

6.2.1 Factors Influencing QoE Measured With Memory Recall-Based Questionnaire

One important outcome of the questionnaire study is the distinction between cognitive, affective, and behavioral components of attitudes toward video quality. The cognitive dimension was expressed in how strongly participants valued quality, with 81.8% rating it as "quite a bit" or "extremely." The affective dimension appeared in irritation with stalling or low resolution, showing that delight or annoyance emerges as a direct reaction to quality problems. The behavioral dimension was more complex. Short-term actions such as refreshing the browser or restarting the connection were common, while only 28% of participants declared they would pay more for better quality. This indicates that behavior includes both immediate problem-solving and long-term commitments, but these are not equally probable and may depend on external factors such as economic conditions.

This differentiation observed in the questionnaire is directly reflected in the structure of the model proposed in Chapter 3. By treating perceived quality, affective responses, and user behavior as distinct but related units, the model incorporates the complexity of attitudes revealed in the study. This contributes to a clearer understanding of what QoE metrics and scales measure, showing that they address specific yet interconnected dimensions of the experience.

The comparison of mean scores after memory recall shows clear differences between use cases in terms of the most influential factors. For video chat, the most important elements were technical: fluency, synchronization, internet stability, and artifacts, which all directly determine the smoothness of communication. In contrast, VoD and live streaming placed strong emphasis not only on technical parameters such as resolution or synchronization but also on the value of the content itself. Interest, appreciation, and emotions evoked by the content consistently ranked among the top factors, with live streaming further influenced by event-related aspects such as premieres. The table also highlights how some technical details, like dark scene reproduction, were more relevant for VoD, reflecting the aesthetics of professional content, while temporal aspects such as duration played a stronger role in video chat, likely due to fatigue in longer sessions. Importantly, price and advertisements were consistently rated as low in influence, suggesting that in natural usage scenarios, these aspects matter less than often assumed in controlled experiments. Overall, these findings point to a clear division: video chat quality is mainly driven by technical reliability, while VoD and live streaming quality depend on both technical performance and the richness of the content.

The importance of content-related factors was included in the path model described in Chapter 3. In the model, content is represented as a confounding factor that can influence both the quality of the multimedia signal (QoMS) and the user's level of delight or annoyance. In practice, this means that the same technical quality may evoke different emotional responses depending on the type of content and the level of appreciation. By placing content alongside QoS, QoMS, and PQoMS, the model reflects the finding that satisfaction in VoD and live streaming cannot be explained by technical parameters alone. This is an important insight, as many existing QoE models (e.g., [52, 151, 131, 163, 154, 53, 44]) reduce the role of content in favor of technical aspects. By explicitly integrating content, the model shows both how technical characteristics of the material shape perceived quality and how emotional reactions to content alter delight or annoyance.

Another important finding concerns the role of the recall procedure. Results showed that recalling a specific experience made the task easier for participants, as they provided more precise answers and used fewer "I don't understand" responses after recall. However, the balance between positive and negative memories varied strongly across use cases, with the majority of VoD and live stream users recalling pleasant experiences, while video chat produced more negative ones. For future studies, this means that the recall task should be designed with a clear instruction to focus either on positive or negative experiences, depending on the research goals. Researchers may also decide to control the distribution of recalled memories by pushing for a balanced sample (e.g., 50/50 positive to negative) or by deliberately focusing only on negative experiences if the aim is to study dissatisfaction. This adjustment would make the method more systematic and ensure that the data collected directly corresponds to the intended scope of the study.

Taken together, these results lead to an important contribution of the dissertation: the development and validation of a memory recall-based questionnaire for QoE studies. The tool is freely available at `https://github.com/TUFIQoE/questionaire`, making it accessible for reuse in both academic and applied contexts. Three versions were developed, each tailored to a different use case: VoD, live streaming, and video chat, making it possible to compare user experiences across services.

The questionnaire is both cheap and scalable, which makes it suitable not only for stand-alone research but also as a flexible extension to other methods. It can be used to pre-select participants for an experiment or to add context-related questions at the end of a QoE test. Importantly, the tool captures not just attitudes toward quality and influential factors but also contextual factors, such as internet satisfaction, usage goals, and social or physical context. By combining these perspectives, it contributes a versatile and open-source method to the QoE field. This is a direct answer to the problem of overlooking factors in QoE studies [33].

6.2.2 Video QoE Theoretical Model

The proposed path model offers a structured way to represent Quality of Experience. Organizing Influential Factors in a directed graph provides a comparable framework that can be directly applied in study design and statistical modeling. Importantly, the model distinguishes between three layers of QoE: the cognitive level, represented by perceived multimedia quality, the affective level, expressed as delight or annoyance, and the behavioral level, reflected in user actions. This separation clarifies the scope of metrics and scales and allows researchers to design studies that address specific parts of the user experience. The model was already showcased at QoMEX 2023 as a tool for increasing comparability between studies and supporting the design of QoE experiments [103]. Building on this structure, I also proposed a new definition of video QoE: *„Quality of Experience is the amount of behaviorally relevant user Delight or Annoyance toward a video service evoked by content and moderated by the perceived quality of the video"*. With this addition, the model not only integrates existing approaches but also extends them, explicitly linking quality judgments, emotions, and actions. As such, the path model constitutes a theoretical advancement for the domain, providing both a parsimonious framework for research and a foundation for refining the conceptual definition of QoE.

6.2.3 Lack of influence of emotionally evoking content and social presence

Figure 6.1 describes the causal model consisting of Hypothesis Four and Five along with the experimental manipulation. With emotional evocation of content and social presence, I tried to change the emotional state of the participants. Then, investigate whether the change in emotional state can influence ACR scores. None of those manipulations influenced ACR scores. These results showed that the arrow from emotions to ACR can be neglected. These are surprising results that show that when people score quality with the ACR scale, neither content nor social presence influences scores in a significant way. This makes the proposed theoretical model simpler and acyclic. As Figure 6.2 shows, with differentiation on Perceived QoMS, delight or annoyance, and behaviour, it is possible to capture Quality of Experience in the form of a direct action graph (DAG).

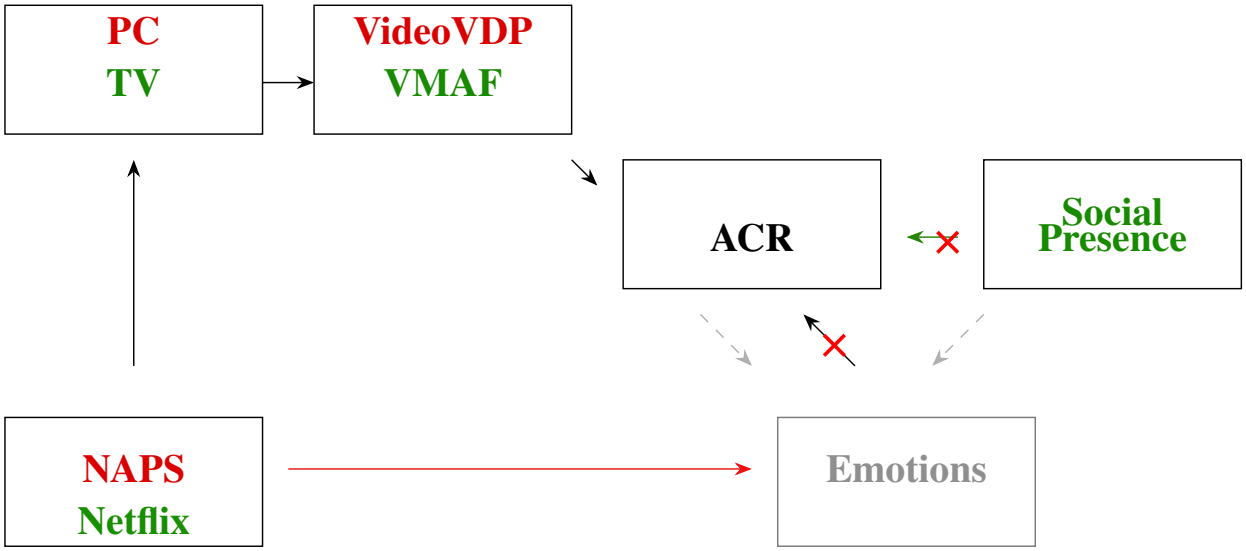

Figure 6.1: This causal graph shows the design of both experiments from chapter 2 and 5 within one theoretical framework. Red text represents variables from the experiment using emotionally evoking content. Green text represents variables from the experiment with social presence manipulation. Arrows represent the causal influence. The red arrow depicts manipulation in the study from chapter 4, and the green arrow depicts manipulation in the study described in chapter 5. Color gray to represent unobserved variables and influences. The red cross marks the lack of influence found in studies.

As those results are contradictory to previously proposed theories, hypotheses drawn from the memory recall questionnaire, and the initial intuition of the entire TUFIQoE project team, some strong claims can be proposed. Below, I present 4 claims that can be inferred from the results.

ACR is highly effective in measuring the perceived quality of multimedia signals.

When participants use the ACR scale to score the quality of the image or video, they can distinguish the quality of the multimedia signal from other factors. The quality distribution provided in QoE studies makes this task easy and the objective of the experiment evident. Thus, highly controlled experiments might be overkill.

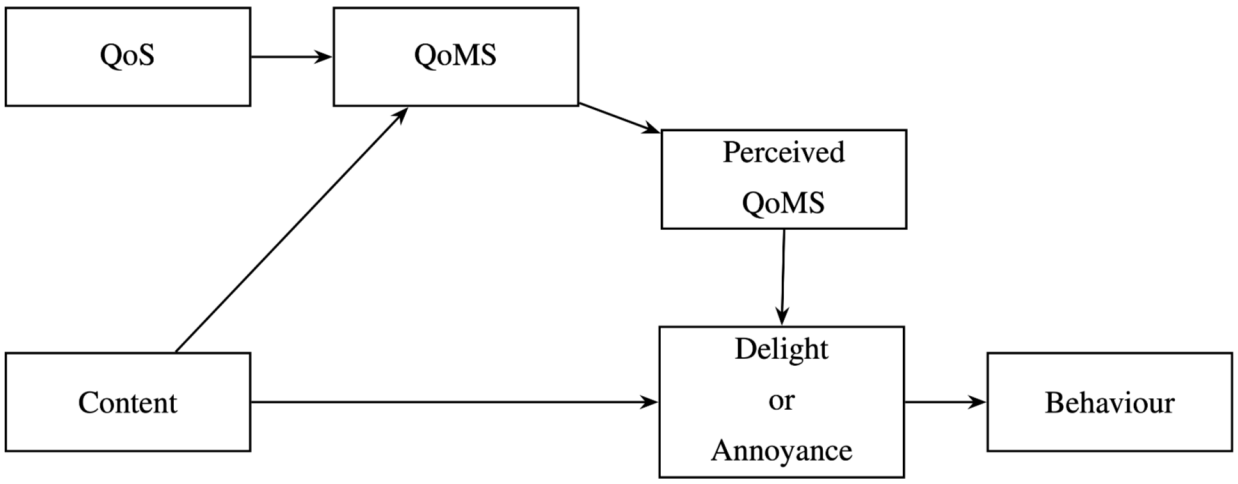

Figure 6.2: Path model, updated after empirical studies. As studies showed no effect on emotional state on the ACR scale path from perceived QoMS to Delight or Annoyance is one-directional.

In addition, both metrics that were used for our studies performed better than expected. Those metrics were tested in a less controlled environment than where they were developed and evaluated before. In addition, new variables that change the emotional state of the participants were implemented. Despite these methodological changes, the included metrics still predict participants' scores, which is equally efficient as typical evaluation studies.

ACR is limited to measures of technical Influential Factors.

Although the above-mentioned robustness of the ACR scale might be advantageous for typical QoE studies, it comes with a price. Based on the lack of influence of social presence and extreme content, ACR might not be a good measure for QoE Influential Factors studies, especially when compared to questionnaire results.

Need for greater clarity in what studies measure

The above effect shows that ACR is not equal to QoE according to its definition. Quality of Experience is defined as the level of delight or annoyance, which refers to the emotional state of the user. Both of manipulations aimed to change the emotional state, yet it has no influence on the ACR scale. Thus, ACR is an assessment of perceived multimedia quality only, which is just one component of QoE. This distinction is crucial, as most QoE studies actually measure only perceived multimedia quality, not the level of satisfaction. For future development of the domain, it is important to distinguish the perceptual level of QoE from the emotional and behavioral levels.

External validity of future QoE studies

The limitation of the ACR scale test is related to the limitation of quality metrics based on scores provided by this scale. To enhance the external validity of QoE studies and predict more than just perceived quality, new experimental protocols are necessary. If new QoE studies incorporate emotional and behavioral information,

metrics predicting actual delight or annoyance, and users' behavior, may be developed. This would be a step forward to improve the external validity of QoE metrics.

6.3 Recommendations

In light of these findings, several methodological directions for future QoE research can be proposed.

Robust ACR usage

Firstly, employ the Absolute Category Rating (ACR) scale in subjective studies that focus on perception or technical aspects of video quality. Its robustness against unrelated factors makes it a reliable tool in such contexts.

Alternative methods for broader factors

Secondly, if the scope of the study is broader than technical factors only, such as emotional or behavioral elements, seek alternatives to ACR due to its limited applicability in these areas. The questionnaire from chapter 2 can be one of the candidates. This questionnaire was also adjusted to be used after the experiment, and all its variations can be found at `https://github.com/TUFIQoE/questionaire`.

Represent variables on path diagrams

For a clear description of the scope of the study, the use of causal models can be beneficial as presented in chapter 3. Such models provide a clear description of the variables and manipulations in the study. They may also lead to better applicability and comparability of QoE studies [103].

Employment of modern metrics

Thirdly, if the main interest of study goes beyond technical factors, adopt advanced metrics such as VMAF or FovVideoVDP as predictive tools instead of asking about quality. These metrics are effective in estimating how participants would rate video quality in classic QoE experiments, serving as an alternative to direct quality inquiries.

Behavioral and emotional assessments

Fourthly, in future QoE studies, prioritize the integration of behavioral and emotional measurements. This should lead to the development of metrics capable of predicting a broader spectrum of user responses, including emotional reactions and behaviors, thus transcending the limits of mere perceptual assessment.

Content control

Finaly, for the development of metrics aiming to model delight or annoyance or behaviour, it is important to consider the role of content as a cofounding factor.

6.4 Application

While the proposed model and its definition were not yet adopted directly in other QoE studies outside of the TUFIQoE project, it has already influenced VQEG discussions and actions. In upcoming „White Paper on Quality of Experience-Aware Management for Collaboration Between Network and Application Providers" [184], new definitions and layers of QoE will be proposed. This VQEG white paper proposes a framework to improve end-user Quality of Experience (QoE) by facilitating cooperation between Content and Application Providers (CAPs) and Communication Service Providers (CSPs). It introduces a layered model that clearly distinguishes between network-level KPIs, application-level KQIs, and user-centric QoE, establishing a common language for research and industry. The core contribution is a shared state table that enables structured information exchange of QoE-relevant metrics, allowing both CAPs and CSPs to optimize services collaboratively. Practical use cases in video streaming and cloud gaming demonstrate the framework's benefits, while privacy considerations and pathways toward standardization ensure its applicability in real-world deployments.

The white paper is currently under review, having been announced at the ITU-T SG12 meeting, and is expected to be published before the end of 2025. It has been prepared by more than 20 experts representing over 15 organizations from both academia and industry, including YouTube, Meta, Telefónica, AT&T, Ericsson, Nokia, and TikTok. The author of this thesis contributed as one of the co-authors. During the preparation of this document, several of the distinctions and definitions proposed in this thesis were discussed and refined to align with the objectives of the white paper.

6.5 Limitations and Future Studies

In this dissertation, two Influential Factors were verified experimentally. If future studies show the influence of different factors, they should be incorporated into future QoE models, such as the one described in chapter 3.

What is more, the above-described conclusions are limited to video-on-demand scenarios only. Current frontiers of QoE studies include research on Extended Reality and Artificial Intelligence. Those areas are drastically different from the video domain, and it is important to highlight those differences. Firstly, both in XR and AI, the possibilities of quality manipulation are limited compared to video. Thus, providing sufficient differences in stimulus or similar operationalization levels using metrics, as in the video, is not possible. Moreover, users have less experience with those technologies. Thus, they can be more prone to biases because rating quality in those scenarios is a more challenging task. Finally, both XR and AI are more multidimensional problems compared to video. Thus, it is important to narrow the scope of the claims from this dissertation only to video-on-demand services.

Nevertheless, I believe that insights about the subjective scales and the rating process itself that were described in this dissertation can be useful in studies of frontier technologies. Moreover, models like the one presented here can be developed and adjusted to model multimodal problems and leverage inference about them. Together, these contributions can help researchers identify the most adequate tools for different layers in multimodal QoE of XR and AI. At the same time, future studies should make a clear effort to

operationalize variables in a transparent and reproducible manner, since only precise definitions allow for meaningful accumulation of knowledge across experiments. The layered perspective on QoE proposed in this work provides a useful framework for guiding such operationalization, ensuring that technical, experiential, and contextual dimensions are addressed coherently. In addition, the findings highlight the importance of measuring external validity and investing in a deeper understanding of how participants actually use rating scales, since the interpretive process of scoring is as critical as the numerical results themselves. By combining methodological rigor with conceptual clarity, QoE research can progress toward frameworks that are both theoretically rich and practically applicable.

Bibliography

[1] Itu-t recommendation p.912: Subjective video quality assessment methods for recognition tasks, March 2016. Series P: Terminals and subjective and objective assessment methods, Audiovisual quality in multimedia services.

[2] Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats, August 2019. Third edition, 2019-08.

[3] Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. ITU-T Recommendation P.913, International Telecommunication Union, June 2021. Series P: Audiovisual quality in multimedia services.

[4] Us streaming satisfaction report 2022. Technical report, Whip Media, Los Angeles, CA, June 2022. Survey of 2,460 US TV Time app users, April 29–May 4, 2022.

[5] Itu-t recommendation p.910: Subjective video quality assessment methods for multimedia applications. Technical Report P.910, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Geneva, Switzerland, October 2023.

[6] Anne Aaron, Anush Moorthy, and David Ronca. Per-title encode optimization. `https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2`, December 2015. Netflix Tech Blog.

[7] Kylee T Ack Baraly, Lydia Muyingo, Christine Beaudoin, Sanaz Karami, Melina Langevin, and Patrick SR Davidson. Database of emotional videos from ottawa (devo). *Collabra: Psychology*, 6(1):10, 2020.

[8] Florence Agboma and Antonio Liotta. Quality of experience management in mobile content delivery systems. *Telecommunication Systems*, 49(1):85–98, 2012.

[9] Ali Ak, Abhishek Gera, Denise Noyes, Hassene Tmar, Ioannis Katsavounidis, and Patrick Le Callet. Comparison of crowdsourcing and laboratory settings for subjective assessment of video quality and acceptability & annoyance. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1159–1164. IEEE, 2024.

[10] Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, and Frédéric Dufaux. Rv-tmo: Large-scale dataset for subjective quality assessment of tone mapped images. *IEEE Transactions on Multimedia*, 25:6013–6025, 2022.

[11] Ali Ak, Andreas Pastor, and Patrick Le Callet. From just noticeable differences to image quality. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pages 23–28, 2022.

[12] Mohammed Alreshoodi and John Woods. Survey on qoe\qos correlation models for multimedia services. *arXiv preprint arXiv:1306.0221*, 2013.

[13] American Psychological Association. Apa dictionary of psychology, n.d. Accessed via APA's official online dictionary of psychology, which contains over 25000 definitions :contentReference[oaicite:1]index=1.

[14] Mehdi Banitalebi-Dehkordi, Abbas Ebrahimi-Moghadam, Morteza Khademi, and Hadi Hadizadeh. No-reference video quality assessment based on visual memory modeling. *IEEE Transactions on Broadcasting*, 66(3):676–689, 2019.

[15] Jasmina Baraković Husić and Sabina Baraković. Multidimensional modelling of quality of experience for video streaming. *Computers in Human Behaviour*, 129, 2022.

[16] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[17] Timothy Beechey. Ecological validity, external validity, and mundane realism in hearing science. *Ear and Hearing*, 43(5), 2022.

[18] Ralf Bender and Stefan Lange. Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4):343–349, 2001.

[19] Niall Bolger and Jean-Philippe Laurenceau. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press, 2013.

[20] Charles Bonnineau, Wassim Hamidouche, Jerome Fournier, Naty Sidaty, Jean-François Travers, and Olivier Deforges. Perceptual quality assessment of hevc and vvc standards for 8k video. *IEEE Transactions on Broadcasting*, 68(1):246–253, 2022.

[21] George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[22] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ..., 1999.

[23] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[24] Steven J. Breckler. Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of personality and social psychology*, 47(6):1191, 1984.

[25] Tobias Brosch, Klaus Scherer, Didier Grandjean, and David Sander. The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, 143(1920):w13786–w13786, 2013.

[26] Egon Brunswik. Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychol. Monogr*, 56(1):i–49, 1944.

[27] ITUR BT. 500-14. bt. 500: Methodologies for the subjective assessment of the quality of television images. *International Telecommunications Union: Geneva, Switzerland*, 2019.

[28] John R Cavanaugh, Richard W Hatch, and John L Sullivan. Models for the subjective effects of loss, noise, and talker echo on telephone connections. *Bell System Technical Journal*, 55(9):1319–1371, 1976.

[29] Central Statistical Office of Poland. Information society in poland in 2022, 2022. Accessed: [11.10.2023].

[30] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 491–500, 2009.

[31] Yanjiao Chen, Kaishun Wu, and Qian Zhang. From qos to qoe: A tutorial on video quality assessment. *IEEE Communications Surveys & Tutorials*, 17(2):1126–1165, 2014.

[32] Natalia Cieplińska. *Quality Assessment of Video Services in the Long Term*. Doctoral dissertation, AGH University of Krakow, Faculty of Computer Science, Electronics and Telecommunications, Kraków, 2024.

[33] Natalia Cieplińska, Lucjan Janowski, Katrien De Moor, and Michał Wierzchoń. Long-term video qoe assessment studies: A systematic review. *IEEE Access*, 10:133883–133897, 2022.

[34] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

[35] OH Coolidge and GC Reier. An appraisal of received telephone speech volume. *The Bell System Technical Journal*, 38(3):877–897, 1959.

[36] David R Cox. Role of models in statistical analysis. *Statistical Science*, 5(2):169–174, 1990.

[37] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[38] Elise S Dan-Glauser and Klaus R Scherer. The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. *Behavior research methods*, 43(2):468–477, 2011.

[39] Katrien De Moor, Hege Krokås Borge, and Poul Heegaard. Young children and the use of video chat: Implications for qoe research. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[40] Katrien De Moor, M Rios Quintero, Dominik Strohmeier, and Alexander Raake. Evaluating qoe by means of traditional and alternative subjective measures: an exploratory'living room lab'study on iptv. *Vienna, Austria*, 2013.

[41] Toon De Pessemier, Luc Martens, and Wout Joseph. Modeling subjective quality evaluations for mobile video watching in a living lab context. In *2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–5. IEEE, 2013.

[42] Adolfo Di Crosta, Pasquale La Malva, Claudio Manna, Anna Marin, Rocco Palumbo, Maria Cristina Verrocchio, Michela Cortini, Nicola Mammarella, and Alberto Di Domenico. The chieti affective action videos database, a resource for the study of emotions in psychology. *Scientific Data*, 7(1):1–6, 2020.

[43] Zhengfang Duanmu, Kede Ma, and Zhou Wang. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach. *IEEE Transactions on Image Processing*, 27(12):6135–6146, 2018.

[44] Sebastian Egger, Peter Reichl, and Katrin Schoenenberg. Quality of experience and interactivity. In *Quality of experience*, pages 149–161. Springer, 2014.

[45] Waqas Ellahi, Toinon Vigier, and Patrick Le Callet. Evaluation of the bubble view metaphor for the crowdsourcing study of visual attention deployment in tone-mapped images. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2021.

[46] Anantrasirichai Nantheera et al. Encoding in the dark grand challenge: an overview. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, 2020.

[47] Avşar Asan et al. Impact of video resolution changes on qoe for adaptive video streaming. In *2017 IEEE ICME*, pages 499–504. IEEE, 2017.

[48] Danieau Fabien et al. Enabling embodiment and interaction in omnidirectional videos. In *2017 IEEE ICME*, pages 697–702. IEEE, 2017.

[49] Fauville Geraldine et al. Zoom exhaustion & fatigue scale. *Available at SSRN 3786329*, 2021.

[50] Ickin Selim et al. Factors influencing quality of experience of commonly used mobile applications. *IEEE Communications Magazine*, 50(4):48–56, 2012.

[51] Kahneman Daniel et al. A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702):1776–1780, 2004.

[52] Kjell Brunnström et al. Qualinet white paper on definitions of quality of experience. 2013.

[53] Reiter Ulrich et al. Factors influencing quality of experience. In *Quality of experience*, pages 55–72. Springer, 2014.

[54] Reuge Nicolas et al. Education response to covid 19 pandemic, a special issue proposed by unicef: Editorial review. *International Journal of Educational Development*, 87:102485, 2021.

[55] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, pages 1–20, 2021.

[56] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, 24(2):36–41, 2010.

[57] Markus Fiedler, Sebastian Möller, Peter Reichl, and Min Xie. QoE Vadis? (Dagstuhl Perspectives Workshop 16472). *Dagstuhl Manifestos*, 7(1):30–51, 2018.

[58] Ebba Fogelberg. We move in order to perceive: A mouse-tracking study of user behaviour during stalling branched videos with a playback bar, 2020.

[59] Bo Fu, Gerald Kunzmann, Daniel Corujo, Michelle Wetterwald, and Rui Costa. Qoe-aware traffic management for mobile video delivery. In *2013 IEEE International Conference on Communications Workshops (ICC)*, pages 652–656. IEEE, 2013.

[60] Xavier Gabaix and DAVID LAIBSON. The seven properties. *The Foundations of Positive and Normative Economics: A Handbook*, 1:292, 2008.

[61] Boni García, Francisco Gortázar, Micael Gallego, and Andrew Hines. Assessment of qoe for video and audio in webrtc applications using full-reference models. *Electronics*, 9(3):462, 2020.

[62] Boni García, Luis López-Fernández, Francisco Gortázar, and Micael Gallego. Practical evaluation of vmaf perceptual video quality for webrtc applications. *Electronics*, 8(8):854, 2019.

[63] David Geerts, Katrien De Moor, István Ketykó, An Jacobs, Jan Van den Bergh, Wout Joseph, Luc Martens, and Lieven De Marez. Linking an integrated framework with appropriate methods for measuring qoe. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 158–163, 2010.

[64] Dale L Goodhue and Eleanor T Loiacono. Randomizing survey question order vs. grouping questions by construct: An empirical test of the impact on apparent reliabilities and links to related constructs. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 3456–3465. IEEE, 2002.

[65] Abhishek Goswami, Ali Ak, Wolf Hauser, Patrick Le Callet, and Frederic Dufaux. Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2021.

[66] TUFIQoE Group. Yournetflixourlab-tufiqoe-2022. `https://github.com/TUFIQoE/YourNetflixOurLab-TUFIQoE-2022`, 2022. Accessed: 2025-06-17.

[67] Rishabh Gupta, Jan-Niklas Antons, Robert Schleicher, Sebastian Möller, Tiago H Falk, et al. Objective characterization of human behavioural characteristics for qoe assessment: A pilot study on the use of electroencephalography features. In *2013 IEEE Globecom Workshops (GC Wkshps)*, pages 1168–1173. IEEE, 2013.

[68] J Herman. Effect of signal distortion on morse telegraph transmission quality. *Bell System Technical Journal*, 8(2):267–285, 1929.

[69] James A Hoffmeyer. Voiceband quality–of–service issues in the post divestiture environment. Technical report, Institute for Telecommunication Sciences, 1985.

[70] Gijs A Holleman, Ignace TC Hooge, Chantal Kemner, and Roy S Hessels. The 'real-world approach'and its problems: A critique of the term ecological validity. *Front. in Psychol.*, 11:721, 2020.

[71] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[72] Miao Hu, Jiawen Chen, Di Wu, Yipeng Zhou, Yi Wang, and Hong-Ning Dai. Tvg-streaming: Learning user behaviors for qoe-optimized 360-degree video streaming. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4107–4120, 2020.

[73] Jasmina Baraković Husić and Sabina Baraković. Multidimensional modelling of quality of experience for video streaming. *Computers in Human Behavior*, 129:107155, 2022.

[74] i3 Forum. Service value and process of measuring qos kpis. White Paper / Technical Report Release 1.0, i3 Forum, May 2010. Proprietary document.

[75] International Telecommunication Union. *Red Book: Volume I. 1st CCITT Plenary Assembly, Geneva, 10–20 December 1956*. International Telecommunication Union, Geneva, 1957. Minutes, Resolutions, and Questions to be studied, including adoption of standardized subjective testing methods such as Mean Opinion Score (MOS).

[76] International Telecommunication Union. ITU-T Recommendation P.10/G.100 (2006) – Amendment 1: Vocabulary for performance and quality of service. New Appendix I – Definition of Quality of Experience (QoE). ITU-T Recommendation P.10/G.100 Amd.1, International Telecommunication Union, Geneva, January 2007. Approved on 25 January 2007 by ITU-T Study Group 12 (2005–2008).

[77] International Telecommunication Union. Definitions of terms related to quality of service. Recommendation E.800, ITU-T, Geneva, September 2008. Approved on 23 September 2008 by ITU-T Study Group 2 (2005–2008).

[78] International Telecommunication Union. ITU-T Recommendation E.804: QoS aspects for popular services in mobile networks. ITU-T Recommendation E.804, International Telecommunication Union, February 2014. Series E: Overall network operation, telephone service, service operation and human factors.

[79] International Telecommunication Union. ITU-T Recommendation P.1203: Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport. ITU-T Recommendation P.1203, International Telecommunication Union, October 2017. Series P: Telephone transmission quality, telephone installations, local line networks.

[80] International Telecommunication Union. Vocabulary for performance, quality of service and quality of experience. ITU-T Recommendation P.10/G.100, International Telecommunication Union (ITU), November 2017. Series P: Telephone transmission quality, telephone installations, local line networks.

[81] International Telecommunication Union. ITU-T Recommendation E.840: Statistical framework for end-to-end network performance benchmark scoring and ranking. ITU-T Recommendation E.840, International Telecommunication Union, June 2018. Series E: Overall network operation, telephone service, service operation and human factors.

[82] International Telecommunication Union. ITU-T Recommendation P.863: Perceptual Objective Listening Quality Prediction. ITU-T Recommendation P.863, International Telecommunication Union, March 2018. Series P: Telephone transmission quality, telephone installations, local line networks; Approval date: 16 March 2018.

[83] ITU-R. 500-14 (10/2019): Methodologies for the subjective assessment of the quality of television images. *ITU: Geneva, Switzerland*, 2020.

[84] ITU-T. Recommendation G.107: The E-model, a computational model for use in transmission planning. ITU-T Recommendation G.107, International Telecommunication Union, Geneva, March 2005. Series G: Transmission Systems and Media, Digital Systems and Networks.

[85] P ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications, 1999.

[86] ITUT. P.800.1 (07/16): Mean opinion score (mos) terminology. *ITU: Geneva, Switzerland*, 2016.

[87] Joint Photographic Experts Group (JPEG). Jpeg aic: Assessment of image coding. `https://jpeg.org/aic/`, 2025. Accessed: 12 August 2025.

[88] Parikshit Juluri, Venkatesh Tamarapalli, and Deep Medhi. Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys Tutorials*, 18(1):401–418, 2016.

[89] Satu Hannele Jumisko, Ville Petteri Ilvonen, and Kaisa Anneli Vaananen-Vainio-Mattila. Effect of tv content in subjective assessment of video quality on mobile devices. In *Multimedia on Mobile Devices*, volume 5684, pages 243–254. SPIE, 2005.

[90] Satu Jumisko-Pyykkö and Miska M Hannuksela. Does context matter in quality evaluation of mobile television? In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 63–72, 2008.

[91] Evangelos Karapanos, Pedro Teixeira, and Ruben Gouveia. Need fulfillment and experiences on social media: A case on facebook and whatsapp. *Computers in Human Behavior*, 55:888–897, 2016.

[92] Evangelos Karapanos, John Zimmerman, Jodi Forlizzi, and Jean-Bernard Martens. User experience over time: an initial framework. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 729–738, 2009.

[93] Ioannis Katsavounidis. Dynamic optimizer — a perceptual video encoding optimization framework. Netflix Technology Blog, March 2018. Senior Research Scientist, Video Algorithms.

[94] Angeliki V Katsenou, Fan Zhang, Kyle Swanson, Mariana Afonso, Joel Sole, and David R Bull. Vmaf-based bitrate ladder estimation for adaptive streaming. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021.

[95] Patrick Van Kenhove and Patrick Desrumaux. The relationship between emotional states and approach or avoidance responses in a retail environment. *The International Review of Retail, Distribution and Consumer Research*, 7(4):351–368, 1997.

[96] Asiya Khan, Lingfen Sun, and Emmanuel Ifeachor. Qoe prediction model and its application in video quality adaptation over umts networks. *IEEE Transactions on Multimedia*, 14(2):431–442, 2011.

[97] Hyun Jong Kim and Seong Gon Choi. Qoe assessment model for multimedia streaming services using qos parameters. *Multimedia tools and applications*, 72(3):2163–2175, 2014.

[98] Hendrik Knoche and Martina Angela Sasse. The big picture on small screens delivering acceptable video quality in mobile tv. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 5(3):1–27, 2009.

[99] Fumiya Kobayashi, Masataka Masuda, and Takanori Hayashi. Qoe assessment method for mobile video services based on user motivation. In *Image Quality and System Performance IX*, volume 8293, pages 301–309. SPIE, 2012.

[100] Kamil Koniuch. Doctoral Consortium 5 - Factors influencing video Quality of Experience: measurements and theoretical model. 6 2022.

[101] Kamil Koniuch. Factors influencing video quality of experience in ecologically valid experiments: Measurements and a theoretical mode. In *Proceedings of the 14th ACM Multimedia Systems Conference*, MMSys '23, page 338–342, New York, NY, USA, 2023. Association for Computing Machinery.

[102] Kamil Koniuch, Sabina Baraković, Jasmina Baraković Husić, Sruti Subramanian, Katrien De Moor, Lucjan Janowski, and Michał Wierzchoń. Top-down and bottom-up approaches to video quality of experience studies; overview and proposal of a new model. *Frontiers in Computer Science*, 6:1305670, 2024.

[103] Kamil Koniuch, Lucjan Janowski, Katrien De Moor, Michał Wierzchoń, and Sruti Subramanian. The role of theoretical models in ecologically valid studies: the example of a video quality of experience

model. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 67–72. IEEE, 2023.

[104] Kamil Koniuch, Lucjan Janowski, Katrien De Moor, Michał Wierzchoń, and Sruti Subramanian. The role of theoretical models in ecologically valid studies: the example of a video quality of experience model. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 67–72, 2023.

[105] Philip Kortum and Marc Sullivan. The effect of content desirability on subjective video quality ratings. *Human factors*, 52(1):105–118, 2010.

[106] Zachary J Kunicki, Meghan L Smith, and Eleanor J Murray. A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, 6(2):25152459231156085, 2023.

[107] Michał Kuniecki, Kinga B Wołoszyn, Aleksandra Domagalik, and Joanna Pilarczyk. Effects of scene properties and emotional valence on brain activations: A fixation-related fmri study. *Frontiers in human neuroscience*, 11:429, 2017.

[108] Fatima Laiche, Asma Ben Letaifa, and Taoufik Aguili. Qoe-aware traffic monitoring based on user behavior in video streaming services. *Concurrency and Computation: Practice and Experience*, page e6678, 2021.

[109] Fatima Laiche, Asma Ben Letaifa, Imene Elloumi, and Taoufik Aguili. When machine learning algorithms meet user engagement parameters to predict video qoe. *Wireless Personal Communications*, 116(3):2723–2741, 2021.

[110] Peter Lang and Margaret M Bradley. The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29:70–73, 2007.

[111] Peter J Lang. International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical report*, 2005.

[112] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*, 1999.

[113] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold POS Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 719–728, 2009.

[114] Patrick Le Callet, Sebastian Möller, Andrew Perkis, Kjell Brunnström, Sergio Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hoßfeld, et al. *Qualinet white paper on definitions of quality of experience*. PhD thesis, Qualinet (www. qualinet. eu), 2013.

[115] Mikołaj Leszczuk and Lucjan Janowski. New recommendation for subjective video quality assessment methods for recognition tasks. *Preprints*, 2021. Not peer-reviewed.

[116] Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet. Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience. *IEEE Transactions on Multimedia*, 21(10):2589–2602, 2019.

[117] Jing Li, Lukáš Krasula, Yoann Baveye, Zhi Li, and Patrick Le Callet. Accann: A new subjective assessment methodology for measuring acceptability and annoyance of quality of experience. *IEEE Transactions on Multimedia*, 21(10):2589–2602, 2019.

[118] Jing Li, Lukáš Krasula, Patrick Le Callet, Zhi Li, and Yoann Baveye. Quantifying the influence of devices on quality of experience for video streaming. In *2018 Picture Coding Symposium (PCS)*, pages 308–312, 2018.

[119] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2):2, 2016.

[120] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[121] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and Jan De Cock. Vmaf: The journey continues. Netflix Tech Blog, October 2018.

[122] Tomasz Lyko, Yehia Elkhatib, Rajiv Ramdhany, and Nicholas Race. Drop or stop: Investigating the impact of playback rate on qoe in adaptive video streaming. In *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 111–117. IEEE, 2024.

[123] Orlewilson Bentes Maia, Hani Camille Yehia, and Luciano de Errico. A concise review of the quality of experience assessment for video streaming. *Computer communications*, 57:1–12, 2015.

[124] Rafal Mantiuk and University of Cambridge. Fovvideovdp: Benchmark of video quality metrics. `https://www.cl.cam.ac.uk/research/rainbow/projects/fovvideovdp/html_reports/benchmark/`, 2023. Accessed: 2025-04-03.

[125] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021.

[126] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior research methods*, 46(2):596–610, 2014.

[127] WH Martin. Rating the transmission performance of telephone circuits. *Bell System Technical Journal*, 10(1):116–131, 1931.

[128] Richard McElreath. Multivariate linear models. In *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, chapter 5, pages 135–172. Chapman & Hall/CRC, 1st edition, 2016.

[129] Richard McElreath. Overfitting, regularization, and information criteria. In *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, chapter 6, pages 153–190. CRC Press, 2 edition, 2020.

[130] Jarosław M Michałowski, Dawid Droździel, Jacek Matuszewski, Wojtek Koziejowski, Katarzyna Jednoróg, and Artur Marchewka. The set of fear inducing pictures (sfip): Development and validation in fearful and nonfearful individuals. *Behavior Research Methods*, 49:1407–1419, 2017.

[131] Sebastian Möller, Marcel Wältermann, and Marie-Neige Garcia. *Features of Quality of Experience*, pages 73–84. Springer International Publishing, Cham, 2014.

[132] Hajer Gahbiche Msakni and Habib Youssef. Impact of user emotion and video content on video quality of experience. In *Proc. 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, pages 97–101, 2016.

[133] Szilveszter Nádas, Lars Ernström, David Lindero, and Jonathan Lynam. On qoe-aware traffic management for real-time, interactive video with time-variant spatial complexity. *arXiv preprint arXiv:2507.11798*, 2025.

[134] Hyunwoo Nam, Kyung-Hwa Kim, and Henning Schulzrinne. Qoe matters more than qos: Why people stop watching cat videos. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[135] Rada Doskonałości Naukowej. Komunikat nr 19/2020 w sprawie składania rozpraw doktorskich. `https://www.rdn.gov.pl/komunikaty/komunikat-nr-192020-w-sprawie-skladania-rozpraw-doktorskich.html`, 2020. Accessed: 2025-09-13.

[136] Omer Nawaz, Markus Fiedler, Katrien De Moor, and Siamak Khatibi. Influence of gender and viewing frequency on quality of experience. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–4. IEEE, 2020.

[137] Omer Nawaz, Markus Fiedler, and Siamak Khatibi. Impact of human and content factors on quality of experience of online video streaming. In *17th International Joint Conference on E-Business and Telecommunications (SECRYPT), Lieusant, Paris on 8-10 July*, pages 59–66. SCITEPRESS, 2020.

[138] Yana Nehmé, Patrick Le Callet, Florent Dupont, Jean-Philippe Farrugia, and Guillaume Lavoué. Exploring crowdsourcing for subjective quality assessment of 3d graphics. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2021.

[139] Per Nilsen. Making sense of implementation theories, models, and frameworks. In *Implementation Science 3.0*, pages 53–79. Springer, 2020.

[140] Even Brattbak Øie, Kamil Koniuch, Natalia Cieplińska, and Katrien De Moor. Factors influencing qoe of video consultations. In *2021 13th QoMEX*, pages 137–140. IEEE, 2021.

[141] Alper Özer, Mehpare Tokay Argan, and Metin Argan. The effect of mobile service quality dimensions on customer satisfaction. *Procedia-Social and Behavioral Sciences*, 99:428–438, 2013.

[142] Borysław Paulewicz. Wprowadzenie do teorii wnioskowania przyczynowego dla psychologów: testowalne i nietestowalne założenia przyczynowe i statystyczne. *Przegląd Psychologiczny*, 66(1):93–124, 2023.

[143] Borysław Paulewicz, Marta Siedlecka, and Marcin Koculak. Confounding in studies on metacognition: A preliminary causal analysis framework. *Frontiers in Psychology*, 11:1933, 2020.

[144] Pablo Pérez. The transmission rating scale and its relation to subjective scores. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 31–36. IEEE, 2023.

[145] Andrew Perkis, Peter Reichl, and Sergio Beker. Business perspectives on quality of experience. In *Quality of Experience*, pages 97–108. Springer, 2014.

[146] Leonardo Peroni and Sergey Gorinsky. Quality of experience in video streaming: Status quo, pitfalls, and guidelines. In *2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 558–567. IEEE, 2024.

[147] Margaret H. Pinson. The precision and repeatability of media quality comparisons: Measurements and new statistical methods. *IEEE Transactions on Broadcasting*, 69(2):378–395, 2023.

[148] Margaret H. Pinson, Lucjan Janowski, and Mark D. Gross. Reasons to replace term ecological validity with terms mundane realism and external validity. In *Proceedings of the 17th International Conference on Quality of Multimedia Experience (QoMEX 2025)*, Madrid, Spain, September 2025. Accepted, to appear.

[149] Qualtrics. Qualtrics, 2005. Version as of *2021*.

[150] Alexander Raake and Sebastian Egger. Quality and quality of experience. In *Quality of Experience: Advanced concepts, applications and methods*, pages 11–33. Springer, 2014.

[151] Alexander Raake and Sebastian Egger. *Quality and Quality of Experience*, pages 11–33. Springer International Publishing, Cham, 2014.

[152] Reza Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–2, 2017.

[153] ITUT Recommendation. E. 800, definitions of terms related to quality of service. *International Telecommunication Union's Telecommunication Standardization Sector (ITU-T) Std*, 2008.

[154] Peter Reichl, Sebastian Egger, Sebastian Möller, Kalevi Kilkki, Markus Fiedler, Tobias Hoßfeld, Christos Tsiaras, and Alemnew Asrese. Towards a comprehensive framework for qoe and user behavior modelling. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2015.

[155] Benjamin F Rex. Liability of telegraph companies for fraud, accident, delay and mistakes in the transmission and delivery of messages. *The American Law Register (1852-1891)*, 32(5):281–294, 1884.

[156] Monika Riegel, Łukasz Żurawski, Małgorzata Wierzba, Abnoss Moslehi, Łukasz Klocek, Marko Horvat, Anna Grabowska, Jarosław Michałowski, Katarzyna Jednoróg, and Artur Marchewka. Characterization of the nencki affective picture system by discrete emotional categories (naps be). *Behavior research methods*, 48:600–612, 2016.

[157] Michael Riordan, Lillian Hoddeson, and Conyers Herring. The invention of the transistor. *Reviews of Modern Physics*, 71(2):S336, 1999.

[158] Werner Robitza, Alexander M Dethof, Steve Göring, Alexander Raake, André Beyer, and Tim Polzehl. Are you still watching? streaming video quality and engagement assessment in the crowd. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2020.

[159] Werner Robitza and Alexander Raake. (re-) actions speak louder than words? a novel test method for tracking user behavior in web video services. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2016.

[160] Werner Robitza, Sabine Schönfellner, and Alexander Raake. A theoretical approach to the formation of quality of experience and user behavior in multimedia services. In *5th ISCA/DEGA Workshop on Perceptual Quality of Systems*, pages 39–43, 2016.

[161] Pablo Antonio Sánchez, Salvador Luna-Ramírez, Matías Toril, Carolina Gijón, and Juan L Bejarano-Luque. A data-driven scheduler performance model for qoe assessment in a lte radio network planning tool. *Computer Networks*, 173:107186, 2020.

[162] Robert Schleicher and Jan-Niklas Antons. Evoking emotions and evaluating emotional impact. In *Quality of Experience*, pages 121–132. Springer, 2014.

[163] Marwin Schmitt, Dick CA Bulterman, and Pablo S Cesar. The contrast effect: Qoe of mixed video-qualities at the same time. *Quality and User Experience*, 3(1):1–17, 2018.

[164] Michael James Scott, Sharath Chandra Guntuku, Weisi Lin, and Gheorghita Ghinea. Do personality and culture influence perceived video quality and enjoyment? *IEEE Transactions on Multimedia*, 18(9):1796–1807, 2016.

[165] Michael Seufert, Sarah Wassermann, and Pedro Casas. Considering user behavior in the quality of experience cycle: towards proactive qoe-aware traffic management. *IEEE Communications Letters*, 23(7):1145–1148, 2019.

[166] Herbert A Simon. Science seeks parsimony, not simplicity: Searching for pattern in phenomena. *Simplicity, inference and modelling: Keeping it sophisticatedly simple*, pages 32–72, 2001.

[167] Lea Skorin-Kapov, Martín Varela, Tobias Hoßfeld, and Kuan-Ta Chen. A survey of emerging concepts and challenges for qoe management of multimedia services. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2s):1–29, 2018.

[168] Joel Sole, Liwei Guo, Andrey Norkin, and Mariana Afonso. Performance comparison of video coding standards: an adaptive streaming perspective, dec 2018.

[169] Jiarun Song, Fuzheng Yang, Yicong Zhou, Shuai Wan, and Hong Ren Wu. Qoe evaluation of multimedia services based on audiovisual quality and user interest. *IEEE Transactions on Multimedia*, 18(3):444–457, 2016.

[170] Wei Song, Dian Tjondronegoro, and Michael Docherty. Exploration and optimization of user experience in viewing videos on a mobile phone. *International Journal of Software Engineering and Knowledge Engineering*, 20(08):1045–1075, 2010.

[171] Wei Song and Dian W Tjondronegoro. Acceptability-based qoe models for mobile video. *IEEE Transactions on Multimedia*, 16(3):738–750, 2014.

[172] Hoong-Cheng Soong and Phooi-Yee Lau. Video quality assessment: A review of full-referenced, reduced-referenced and no-referenced methods. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pages 232–237. IEEE, 2017.

[173] Janina Steinmetz and Stefan Pfattheicher. Beyond social facilitation: A review of the far-reaching effects of social attention. *Social Cognition*, 35(5):585–599, 2017.

[174] Dominik Strohmeier, Satu Jumisko-Pyykkö, Kristina Kunze, and Mehmet Oguz Bici. The extended-opq method for user-centered quality of experience evaluation: a study for mobile 3d video broadcasting over dvb-h. *EURASIP Journal on Image and Video Processing*, 2011:1–24, 2011.

[175] Sruti Subramanian, Katrien De Moor, and Kamil Koniuch. Investigating motivational factors influencing users' consumption of video streaming services: A human factor perspective. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 231–242. Springer, 2023.

[176] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, page 1454, 2017.

[177] Syed Uddin, Michał Grega, Mikolaj Leszczuk, and Waqas ur Rahman. Evaluating has and low-latency streaming algorithms for enhanced qoe. 2025.

[178] UNICEF UNESCO. the world bank: Survey on national education responses to covid-19 school closures, round 2, 2020.

[179] N van den Ende. *Experiencing audio/video quality: an investigation into the relationship between perceived video quality and involvement*. 2015.

[180] Gary R VandenBos. *APA dictionary of psychology.* American Psychological Association, 2007.

[181] Martín Varela, Lea Skorin-Kapov, and Touradj Ebrahimi. Quality of service versus quality of experience. In *Quality of Experience: Advanced Concepts, Applications and Methods*, pages 85–96. Springer, 2014.

[182] Maria Torres Vega, Cristian Perra, Filip De Turck, and Antonio Liotta. A review of predictive quality of experience management in video streaming services. *IEEE Transactions on Broadcasting*, 64(2):432–445, 2018.

[183] Video Quality Experts Group (VQEG). Projects home — vqeg. `https://www.vqeg.org/projects-home/`, 2025. Accessed: September 8, 2025.

[184] VQEG. White paper on quality of experience-aware management for collaboration between network and application providers. Manuscript in review, announced at ITU-T SG12, expected publication 2025.

[185] Katarzyna Wac, Selim Ickin, Jin-Hyuk Hong, Lucjan Janowski, Markus Fiedler, and Anind K Dey. Studying the experience of mobile applications used in different contexts of daily life. In *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack*, pages 7–12, 2011.

[186] Cheyenne Wakeland-Hart and Mariam Aly. Predicting image memorability from evoked feelings, Sep 2023.

[187] Karen D Wang, Zhongzhou Chen, and Carl Wieman. Can crowdsourcing platforms be useful for educational research? In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 416–425, 2024.

[188] Andrew B Watson. Proposal: Measurement of a jnd scale for video quality. *IEEE G-2.1. 6 Subcommittee on Video Compression Measurements*, 2000.

[189] Ina Wechsung and Katrien De Moor. Quality of experience versus user experience. In *Quality of experience: advanced concepts, applications and methods*, pages 35–54. Springer, 2014.

[190] Nikolas Wehner, Nils Mertinat, Michael Seufert, and Tobias Hoßfeld. Studying the impact of the content selection method on the video qoe on mobile devices. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–4. IEEE, 2020.

[191] Jannis Weil, Yassin Alkhalili, Anam Tahir, Thomas Gruczyk, Tobias Meuser, Mu Mu, Heinz Koeppl, and Andreas Mauthe. Modeling quality of experience for compressed point cloud sequences based

on a subjective study. In *2023 15th international conference on quality of Multimedia experience (QoMEX)*, pages 135–140. IEEE, 2023.

[192] M Wessa, P Kanske, P Neumeister, K Bode, J Heissler, S Schönfelder, et al. Emopics: Subjektive und psychophysiologische evaluation neuen bildmaterials für die klinisch-bio-psychologische forschung. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 39(Suppl. 1/11):77, 2010.

[193] Mathias Wien and Joel Jung. Remote expert viewing, laboratory tests or objective metrics: which one (s) to trust? *EURASIP Journal on Image and Video Processing*, 2024(1):16, 2024.

[194] Małgorzata Wierzba, Monika Riegel, Anna Pucz, Zuzanna Leśniewska, Wojciech Łukasz Dragan, Mateusz Gola, Katarzyna Jednoróg, and Artur Marchewka. Erotic subset for the nencki affective picture system (naps ero): cross-sexual comparison study. *Frontiers in psychology*, 6:1336, 2015.

[195] Wanlu Yang, Kai Makita, Takashi Nakao, Noriaki Kanayama, Maro G Machizawa, Takafumi Sasaoka, Ayako Sugata, Ryota Kobayashi, Ryosuke Hiramoto, Shigeto Yamawaki, et al. Affective auditory stimulus database: An expanded version of the international affective digitized sounds (iads-e). *Behavior research methods*, 50:1415–1429, 2018.

[196] Y Ben Youssef, Abdelhamid Mellouk, Mériem Afif, and Sami Tabbane. Video quality assessment based on statistical selection approach for qoe factors dependency. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2016.

[197] Jia Zhang, Yixuan Zhang, Enhuan Dong, Yan Zhang, Shaorui Ren, Zili Meng, Mingwei Xu, Xiaotian Li, Zongzhi Hou, Zhicheng Yang, et al. Bridging the gap between {QoE} and {QoS} in congestion control: A large-scale mobile web service perspective. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, pages 553–569, 2023.

[198] Xu Zhang, Hanchen Li, Paul Schmitt, Marshini Chetty, Nick Feamster, and Junchen Jiang. Vidplat: A tool for fast crowdsourcing of quality-of-experience measurements. *CoRR*, 2023.

[199] Tiesong Zhao, Qian Liu, and Chang Wen Chen. Qoe in video transmission: A user experience-driven strategy. *IEEE Communications Surveys & Tutorials*, 19(1):285–302, 2016.

[200] Qi Zheng, Yibo Fan, Leilei Huang, Tianyu Zhu, Jiaming Liu, Zhijian Hao, Shuo Xing, Chia-Ju Chen, Xiongkuo Min, Alan C Bovik, et al. Video quality assessment: A comprehensive survey. *arXiv preprint arXiv:2412.04508*, 2024.

[201] Yi Zhu, Ingrid Heynderickx, and Judith A Redi. Understanding the role of social context and user factors in video quality of experience. *Computers in Human Behavior*, 49:412–426, 2015.