Please cite this paper as:

K. Koniuch, L. Janowski, K. De Moor, M. Wierzchoń and S. Subramanian, "The Role of Theoretical Models in Ecologically Valid Studies: the example of a video Quality of Experience model,"2023 15th International Conference on Quality of Multimedia Experience (QoMEX), Ghent, Belgium, 2023, pp. 67-72, doi: 10.1109/QoMEX58391.2023.10178561.

```
@INPROCEEDINGS{10178561,
  author={Koniuch, Kamil and Janowski, Lucjan and De Moor,
  Katrien and Wierzchoń, Michał and Subramanian, Sruti},
  booktitle={2023 15th International Conference on Quality
  of Multimedia Experience (QoMEX)},
  title={The Role of Theoretical Models in Ecologically
  Valid Studies: the example of a video Quality of Experience model},
  year={2023},
  volume={},
  number={},
  pages={67-72},
  doi={10.1109/QoMEX58391.2023.10178561}
}
```

The Role of Theoretical Models in Ecologically Valid Studies: the example of a video Quality of Experience model

1st Kamil Koniuch
AGH University of Krakow
Institute of Communication Technologies
Kraków, Poland
0000-0002-8243-7155

2nd Lucjan Janowski
AGH University of Krakow
Institute of Communication Technologies
Kraków, Poland
0000-0002-3151-2944

3rd Katrien De Moor
Department of Information
Security and Communication Technology
Norwegian Univeristy of Science and Technology
Trondheim, Norway
0000-0003-2163-8778

4th Michał Wierzchoń
Institute of Psychology
Jagiellonian University
Kraków, Poland
0000-0002-7347-2696

5th Sruti Subramanian
Department of Information
Security and Communication Technology
Norwegian Univeristy of Science and Technology
Trondheim, Norway
sruti.subramanian@ntnu.no

Streszczenie—**This paper discusses the problem of ecological validity of Quality of Experience experiments in a broader context of internal and external validity. We argue that the issue of trade-off between control in the experiment and generalization of the results requires diverse experimental protocols. To enable this process comparability of experiments has to be guaranteed. To improve comparability and communication between researchers, we propose using theoretical models as tools for clearly communicating the assumptions underlying subjective experiments. These models describe the relationships between variables and enable comparability of research. We demonstrate how the theoretical model of video QoE can be used for the comparison of three independent QoE studies. In conclusion, we stress that ecological validity is a mean to an extended external validity. We suggest that new experimental protocols that increase the number of Influence Factors (IFs) measured have to provide comparability with other experiments. Theoretical models can facilitate this process and stimulate higher comparability between studies in QoE subdomains. A community effort is needed to adjust the theoretical model to new use cases, particularly with new immersive multimedia. This paper inspires further research on ecological validity in QoE and encourages the adoption of theoretical models for more comparable research designs.**

Index Terms—**quality of experience, model, content, behavior, multimedia quality, network, subjective quality, video, user perception, ecological validity**

I. Introduction

Modern telecommunications services require quality assurance of the presented multimedia. For example, quality assurance mechanisms allow video streaming platforms to provide optimal quality levels, with the least possible amount of data transferred. Subjective assessments of quality are an important source of information that is used for the optimization of multimedia services. However, subjective data collection can be challenging due to the inherently subjective nature of human perception and the need to carefully operationalize experiments to ensure reliable and valid results. One of the biggest challenges in studies based on the subjective rating of quality is the ecological validity of experimental protocols.

The concept of ecological validity refers to the naturalness of the experimental setup [12]. In the context of Quality of Experience (QoE) research, it is related to the environment, content, and data collection used in experiments. Typical video QoE experiments can be characterized as non-ecologically valid due to the unnatural lighting in the laboratory, short soundless content, and intrusive measurements displayed during the experiment [3]. On the other hand, researchers run experiments measuring behavioral reactions(e.g. [26]) or crowdsource studies (e.g. [19], [25]) to increase ecological validity. In this context ecological validity can be described as part of a broader term - external validity. External validity describes to which extent one can extrapolate the conclusions from a performed study [28].

Conclusions from studies with high external validity can be generalized to different contexts and to diverse groups of people. In a typical video QoE study, the context of measurement and the context from which researchers draw conclusions is distanced. Researchers try to predict the delight or annoyance of everyday users based on perceptual measurements conducted in laboratories. Thus, the generalizability of results gathered with this method is not obvious. Although, it does not mean that those perceptual-focused experiments are not

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

crucial for understanding the satisfaction of users. Moreover, it does not imply that current methods have to be fully replaced with more ecologically valid experimental protocols. However, in this article, we argue that "typical"QoE experiments can and should to a larger extent be **complemented** with more ecologically valid experiments and user studies.

Nevertheless, to be able to achieve such complementary research designs, better comparability of research is necessary. Therefore, we propose using theoretical models, to a much larger extent than is the case today, as tools for clearly communicating the assumptions underlying subjective experiments. Theoretical models based on path diagrams describe the relationships between variables measured in experiments. Thus, they clearly picture the research assumptions and the measured context. Most importantly, diagrams enable comparability of research. In this paper, we use the theoretical model of video QoE [13] to describe 3 video QoE studies performed by other researchers. These studies used different methodologies and were not designed for comparability. We use them as case examples to show how the application of the theoretical model can elevate communication between researchers.

Our goal was to present to the community how theoretical models can be implemented in designing a series of comparable experiments. As we argue below, this methodology might be a key to increasing ecological validity with the preservation of control over variables in studies. Moreover, the described approach can be used for the implementation of Directed Acyclic Graphs (DAGs) [15] in the data analysis process. This method is especially useful for the analysis of experiments with various variables included.

II. Background

Manipulation of one variable to observe consecutive changes in another variable is the essence of experimentation [5]. One of the key requirements for a successful experiment is to ensure and be able to verify that the change in the observed variable is *caused* by the manipulated variable. The extent to which the experimental design can ensure this causal effect is directed as the *internal validity* of an experiment. Unfortunately, the internal validity of the experiment is not obvious to conserve, especially when measuring humans. In the context of QoE research, the observed change in quality ratings might not only be caused by the change in the objective stimuli quality. Other Influence Factors (IFs) can also evoke changes in this dependent variable [24]. Therefore, strictly controlled experimental protocols are recommended to ensure the high internal validity of QoE experiments. The aim of those protocols is to limit the scope of IFs [13] that may play a role. As a result, the *external validity* of the experiments based on such protocols can be characterized as low.

The external validity of the experiment describes the generalizability of the results gathered in the experiment [17]. In other words, it answers the question of how far we can draw generalizable conclusions from the experiment. In table I we present aspects of external validity with examples of questions relevant to QoE studies.

Tabela I
Aspects of External Validity (EV) with Examples Relative to Quality of Experience Research

Aspects	Examples for QoE research
Situations	Can the results from the experiment be generalized to other situations like everyday usage of the considered service or to other multimedia?
People	What population can we conclude about based on the measured sample? Is the result unique for groups of different ages, sex, and cultural background?
Stimuli	Will the measured effect be the same with everyday content? Can we extrapolate the conclusions to different data sets?
Time	Is it possible to predict long-time consumer behavior based on the experiment results?

From the described perspective, ecological validity is an important part of external validity. While external validity describes all of the features important for the generalization of the study, ecological validity focuses on the distance between the measured context and the context about which researchers conclude. E.g., according to the community-supported QoE definition [2], researchers who study QoE try to predict the level of delight or annoyance of everyday service users. Yet, in video QoE studies, the most common measure of QoE is the Absolute Category Scale (ACR), which is focused on the perception of video quality. Moreover, the experimental context is very far from how users interact with video services in a natural context. During a typical video QoE experiment, soundless 10s long video clips are presented multiple times to the testers [3]. Thus, those studies can be described as not ecologically valid. However, that does not mean that they are useless or wrong.

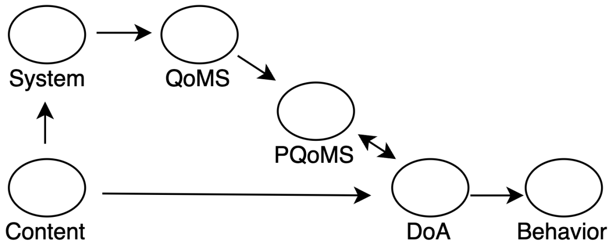
Classical QoE studies such as the ones described above can provide a lot of information about the perception of quality, which can be an important part of everyday users' experience. Nevertheless, in a natural setup, more factors than the objective video quality can lead to users' delight or annoyance. More ecologically valid studies (e.g [1], [29]) therefore include a wider range of influence factors compared to classical experiments.

On the other hand, studies that measure a lot of influence factors can provide lower internal validity. In this type of setup, it is harder to establish a causal relationship between variables. In effect, experimental design is always a trade-off between internal and external validity. To face this challenge, more complementary research approaches are needed. Ecologically valid studies can be a great supplement to classical, perception-focused experiments. Although, despite the development of ecologically valid experimental protocols, conclusions from those studies are rarely implemented in follow-up studies. In order to address this issue, a tool for comparing different studies could be helpful.

Both classical experiments and more ecologically valid QoE studies hold diverse sets of assumptions. We argue that for the comparability of QoE studies, a precise description of assumptions about the relationships between the considered variables is necessary. Moreover, a clear characterization of the research setup and context about which the study concludes is needed. As we will present in the following sections, theoretical models can be very useful for the graphical presentation of study assumptions. Together with a clear description of the experimental procedure theoretical models can elevate the replicability and comparability of QoE studies. In this way, it is possible to design a set of comparable experiments which estimate different influence factors of QoE. The aim of this approach is to increase the external validity of QoE studies with preserving internal validity.

For the presentation of properties of theoretical models, we choose to use the video QoE model described in [13] because of its high generalizability. In other words, the simple structure of this model enables adjustments necessary to describe diverse experiments. Moreover, the causal structure of this model makes it possible to adjust it to adequate statistical methods.

III. Method



Rysunek 2. Path diagram describing video Quality of Experience model [13].

We use diagrams to describe causal relationships between variables in the selected case example studies. In this method, units (depicted as ovals) illustrate variables and arrows represent the assumption that variable A *might* influence variable B. On the other hand, the lack of an arrow represents a much **stronger** assumption, namely a lack of direct influence between variables [18]. The structural representation clearly illustrates assumptions about relationships between measured variables. Moreover, this clear description of variables can be used for presenting the measured context. Thus, it is easier to understand the distance between the measured context and the context about which the study concludes. Moreover, based on described models it is possible to use Directed Acyclic Graphs (DAGs) [15] for data analysis. Thus, diagrams can be a great tool for providing comparability of both designs and conclusions of experiments.

However, to use graphs in a comparable manner a common theoretical structure is needed. For that purpose, we use the model described in [13]. In table II we present the descriptions of model units presented in Fig. 2. This model is generalized and presents complex processes in a simplified manner. Depending on the scope of the research its' complexity can be increased by adding adequate variables. For example, Fig. 1 shows the complexity behind the "system"unit with causal relations between successive processes. With this visualization, it is easier to understand how these processes are represented in the model [13].

The scope of the model is limited to the video QoE only. Thus, we limit the scope of discussed experiments only to the video subdomain of QoE studies and the above-mentioned theoretical model provided us with a minimal structure that we could develop to describe three video QoE studies.

Our aim during the selection of studies was to present 3 different approaches to video QoE studies which are by design incomparable. We look for one "typical"study, one which measures user behavior, and one that introduces a new, uncommon influence factor as the independent variable. For the representation of the classical QoE experiment, we chose a well-known HDTV project [9], [10]. We used the experiment presented in [26] as an example of a behavioral study and [29] as a representation of a study that aims for the measurement of new IFs. We discuss them in the following section in detail.

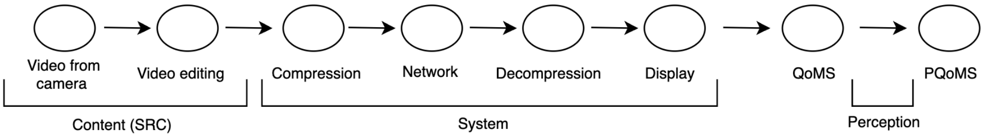
IV. Results

We use the theoretical model described in [13], for the description of the assumptions of the following studies [9], [10], [26], [29]. The goal is to picture the differences and similarities between discussed experiments. Green units represent the manipulated variables. With blue units, we denoted measured variables. Gray units represent the variables that were omitted in the experiment. In parentheses, we describe the operationalization of units in these three experiments.

A. HDTV Project

1) Study context: Researchers gathered data in two projects of Video Quality Expert Group (VQEG) [9], [10]. The main goal of those projects was to provide datasets that can be used by the community to evaluate objective quality metrics. The data sets analyzed were collected in accordance with the classical QoE experiment protocols [21].

2) Application of the theoretical model: In Fig. 3 we describe the above-mentioned processes with the operationalization of units denoted in parentheses. As the content is only operationalized in terms of frame rate, there is not enough information to predict the emotional response of testers (DoA). Moreover, the random assignment of content and the



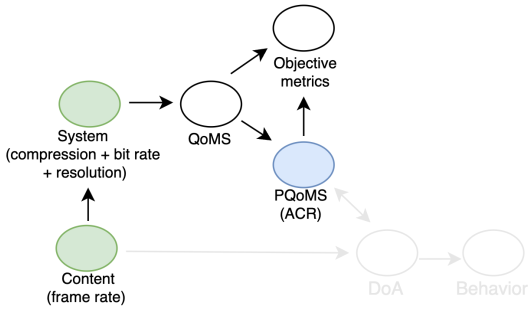
Rysunek 1. Processes behind "System"unit

Tabela II
UNITS OF VIDEO QOE THEORETICAL MODEL [13].

Unit	Definition	Potential Measurements
Content	Video content characteristics are multidimensional and, depending on the research question, different variables might be taken into the account to analyze its influence.	E.g. compression complexity [16], average engagement of users of the YouTube platform, content genre.
System	In [24] System Influence Factors (SIFs) are defined as “properties and characteristics that determine the technically produced quality of an application or service”. Depending on the scope of the study, network, compression methods, or hardware can be analyzed as crucial parts of the system.	The network can be expressed in terms of performance indicators such as throughput, latency, jitter, and packet loss.
Quality of Multimedia Signal (QoMS)	In this model Quality of Multimedia Signal represent the objective properties of visual stimuli. In the context of video streaming, it is video displayed on the user’s device. In previously proposed models [2], [20], [27], QoMS is described as the physical representation of the signal.	It can be assessed with objective metrics like signal-to-noise ratio, or VMAF [7], [8], [14], [22].
Perceived QoMS (PQoMS)	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. In general QoE models [2], [20] these descriptions are the outcome of the quality formation process.	5-point Absolute Category Rating scale [11]
Delight or annoyance (DoA)	We assume that the state of delight or annoyance (DoA) of the user is the outcome of both quality and content properties. According to the general definition, this is in fact the measure of QoE.	Differential Emotions Scale [6].
Behavior	Depending on the scope of the study behavior might be the short-term reaction for quality-related events [26], habits evaluation [25] or even consumer attitude [23] predictor. As long-term behavior toward network providers might be influenced by a set of extra, important variables (e.g. pricing) we focus on short-term behavior.	Interaction with service(e.g. change of the video)

repetition of the same content at different degradation levels were implemented in the study design. Thus, the influence of participants’ emotional reactions to content was minimized. Therefore, units "DoA"related to the emotional reaction and its further connection to behavior were left gray.

The goal of the study is depicted by the relation between "QoMS", "PQoMS", and the objective metrics units. Researchers provided data sets with technical aspects of the system and the correlates in the subjective rating of quality (on the ACR scale). This type of data is crucial for building metrics for quality predictions. Some metrics (e.g. Peak Signal-to-Noise Ratio) focus mostly on the QoMS and are based on objective properties of Multimedia Signal. On the other hand, new metrics like P.1204 [22] use both information about Multimedia Signal and scores of Perceived Quality of Multimedia Signal to predict the quality.

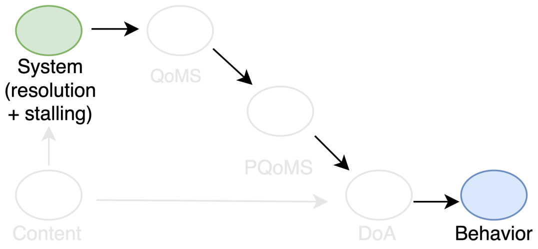


Rysunek 3. HDTV project assumptions described with path diagram model

B. *(Re-) actions speak louder than words? A novel test method for tracking user behavior in Web video services*

1) Study context: In this study [26], researchers developed their own video portal mimicking the YouTube service but without social features. 32 video clips from existing web video portals were implemented in the player. Each clip was shortened to 1.5 to 3 minutes. 15 participants used this service in the laboratory with a room-like set-up on a 13” MacBookProo device. Participants could interact with the player by pausing, changing quality, seeking forward and backward, switching to full screen or window mode, reloading the page, and selecting another video. Researchers were manipulating the video quality and implemented stalling events to evoke participants’ behaviors. The quality ratings were not gathered to avoid distracting the experience.

2) Application of the theoretical model: In Fig. 4 we describe the relationships between the variables in the discussed study. The influence of the system factors on participants’ behavior was the only one measured in the study. In this setup, the behaviors of participants could not change the system thus the only possible arrows go from system to behavior. Content had some unknown variability and its influence was not examined in the study so we decided not to draw the arrow from the content to behavior. Moreover, no objective metrics were used to estimate the QoMS. The influence of other variables from the model [13] was omitted in the study design, thus they are left gray.

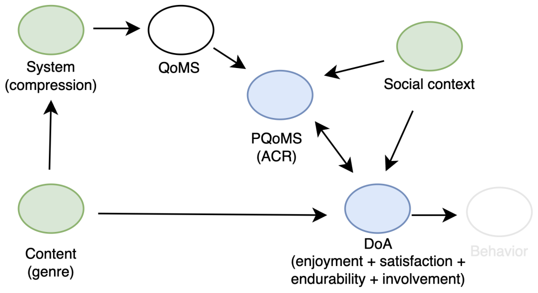


Rysunek 4. Path diagram describing video Quality of Experience model.

C. Understanding the role of social context and user factors in video Quality of Experience

1) Study context: In this experiment, researchers used real-life viewing situations with control of both quality and social context. Participants were randomly assigned to two groups A or B. Group A participated in the experiment alone, and group B consisted of groups of three friends. In total 60 participants participated in the experiment. The study was preceded by a survey of 80 PhD students who scored different genres of content as more appropriate for watching alone or with friends. Video genres refer to categories or classifications of video content, such as comedy, drama, action, romance, horror, or documentary. Based on responses, 3 genres out of 15 were assigned as stimuli for the experiment to ensure content variability. Comedy was preferred to watch with friends, educational videos alone, and for the sports category, almost half of the respondents stated that it does not matter. Each category was assigned two different video clips obtained from the YouTube platform, each at least 5 minutes long. Videos were encoded with H.264/AVC at two different bitrates: high(2000 kbps) and low (600 kbps). The videos were displayed to the participants in random order. After the presentation of each video, participants filled out a questionnaire measuring enjoyment, endurability, satisfaction, involvement, and Perceived Quality of Multimedia Signal.

2) Application of the theoretical model: In Figure 5 we present one of the possible graphical interpretations of discussed study variables. Researchers introduced the social context as a new influence factor in the design of the study. The content and its compression were by design independent of the influence of the social context. Thus, the social context could influence only the perceptive Quality of Multimedia Signal and the Delight or Annoyance. We chose to classify enjoyment, endurability, satisfaction, and involvement as dimensions of overall Delight or Annoyance. Those dimensions are similar in meaning and were highly correlated. Moreover, they were measured with the questionnaire, so a causal relationship between those dimensions is impossible to establish. The influence of content can be described with both possible arrows. This is because researchers choose diverse content that could have potentially different evoke changes in the "DoA"unit. The relationship to the behavior of the participants was not measured, so this unit is left gray.



Rysunek 5. Path diagram describing video Quality of Experience model.

V. Discussion

All of the discussed studies were designed to investigate video Quality of Experience and used adequate ITU-T recommendations for that purpose. Despite this fact, it is not easy to combine the conclusions provided by the discussed studies. It is not only a matter of differences measured context-diverse stimuli and experimental setups. Comparison is also difficult because the different studies draw conclusions about different contexts. Study IV-A allows for the prediction of subjective rating of the quality. Thus, it is rather focused on the perceptual ability to distinguish between different levels of quality degradation, not on the emotional response to this degradation. Study IV-B predicts the behavior of video service users and as a result, the context of conclusions goes far beyond noticing quality degradation. Additionally, study IV-C concludes about the context in which users watch videos together with friends. While these underlying assumptions are embedded into the design, they remain rather implicit. In effect, designing comparable follow-up studies is challenging. Using diagrams to describe relationships between variables might be useful to explicitly describe the scope of an experiments' conclusions and their true limitations.

Moreover, differences in described studies show the multidimensional character of the Quality of Experience. Depending on the scope of the study, Perception of Quality of Multimedia Signal, Delight or Annoyance, or behavior might be used as a measure of QoE. Moreover, diverse Influence Factors (IFs) can be used for experimental manipulation. This diversity makes looking for *one* experimental protocol, which will have a satisfying level of ecological validity, unrealistic.

This is also due to the fact that introducing a higher number of IFs in a single experimental manipulation makes concluding cause and effect troublesome. Changes in the dependent variable can be *caused* by various variables and their interactions. For example, quality ratings might be influenced by both quality distortions and some properties of the source content such as aesthetics. If the content included in the study design has a variability of aesthetics or can evoke diverse emotional reactions, it can be described as a confounding factor. In this scenario, content influences both an independent variable (e.g. compression) **and** dependent variable (quality rating).

The problem of an increased number of variables further has severe consequences for the data analysis. Adding predictors to statistical models, without considering their casual relationships can lead to incorrect predictions [15]. Using directed acyclic graphs (DAGs) to describe a causal relationship between variables can be used to establish how to properly implement variables into the statistical model [4]. Nevertheless, DAGs require to include in the model all of the common causes of each variable. In other words, if pair of variables have some common cause, it has to be represented in a graph. From the perspective of QoE research, fulfilling this property of DAGs in a study that includes lots of IFs seems impossible.

Taking all of the above into consideration, we propose using a series of comparable experiments with diverse levels of ecological validity to increase the external validity of QoE studies. This approach will allow the influence factors to be measured in an additive manner. However, it is important to note that designing comparable experiments requires not only the similarity of stimuli and measurements. It also mandates the comparability of experiment assumptions. Using the same theoretical model to design a series of experiments can make this task easier. In fact, the theoretical model described in this paper is currently used in our "TUFIQOE"project both for conceptualization and data analysis. Our goal is to validate the theoretical assumptions of the model in a series of experiments. Moreover, describing experiment assumptions in form of a diagram facilitates communication between researchers. The measured context and the scope of the conclusion can be depicted clearly and in its entirety. In a such diverse domain as QoE, it might help to establish similarities and differences between studies.

VI. Conclusions

Ecologically valid experiments are means to extend the external validity of QoE studies. To achieve this goal new experimental protocols should not only increase the measured number of Influence Factors (IFs) but also provide comparability with other experiments. This methodology could lead to measuring IFs in an additive manner in a series of experiments. We propose to use theoretical models to facilitate this process. Adding graphical representation to study assumptions can elevate the communication between researchers and stimulate higher comparability between studies. Moreover, theoretical models like the one described in [13] can be used for providing comparable experimental design in QoE subdomains. As we showed in this paper, the multidimensional character of QoE can be a challenge in the relatively simple case of two-dimensional video. The problem of the trade-off between the internal and external validity of the experiment is even bigger with new immersive multimedia. For that reason, a community effort is needed to provide adjustment of the theoretical model [13] to new use-cases.

Literatura

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

[2] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. Qualinet white paper on definitions of quality of experience. 2013.

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

[4] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552, 2020.

[5] Thomas D Cook, Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA, 2002.

[6] 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.

[7] 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.

[8] 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.

[9] Video Quality Experts Group et al. Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i. 2008.

[10] Video Quality Experts Group et al. Report on the validation of video quality models for high definition video content. 2010.

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

[12] John F Kihlstrom. Ecological validity and "ecological validity". *Perspectives on Psychological Science*, 16(2):466–471, 2021.

[13] Kamil Koniuch, Sabina Baraković, Jasmina Baraković Husić, 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. *arXiv preprint arXiv:2301.11648*, 2023.

[14] 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.

[15] Richard McElreath. The many variables & the spurious waffles. In *Statistical Rethinking*, pages 123–160. Chapman and Hall/CRC, 2020.

[16] Vignesh V Menon, Christian Feldmann, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer. Vca: Video complexity analyzer. In *Proceedings of the 13th ACM Multimedia Systems Conference*, pages 259–264, 2022.

[17] Mark Mitchell and Janina Jolley. *Research design explained.* Holt, Rinehart & Winston Inc, 1988.

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

[19] Pablo Pérez, Ester González-Sosa, Redouane Kachach, Francisco Pereira, and Álvaro Villegas. Ecological validity through gamification: an experiment with a mixed reality escape room. In *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 179–183. IEEE, 2021.

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

[21] ITUT Rec. P. 910: Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union, Geneva*, 2, 2008.

[22] ITUTP Recommendation. 1204,"video quality assessment of streaming services over reliable transport for resolutions up to 4k,", 2019.

[23] 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.

[24] Ulrich Reiter, Kjell Brunnström, Katrien De Moor, Mohamed-Chaker Larabi, Manuela Pereira, Antonio Pinheiro, Junyong You, and Andrej Zgank. Factors influencing quality of experience. *Quality of experience: Advanced concepts, applications and methods*, pages 55–72, 2014.

[25] 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.

[26] 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.

[27] 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.

[28] Colin Robson and Kieran McCartan. *Real world research.* Chichester, 2016.

[29] 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.