Generalized Score Distribution: A Two-Parameter Discrete Distribution Accurately Describing Responses from Quality of Experience Subjective Experiments

Jakub Nawała, Lucjan Janowski, Bogdan Ćmiel, Krzysztof Rusek, Pablo Pérez

***Abstract*—Subjective responses from Multimedia Quality Assessment (MQA) experiments are conventionally analyzed with methods not suitable for the data type these responses represent. Furthermore, obtaining subjective responses is resource intensive. Thus, a method that allows the reuse of existing responses would be beneficial. Applying improper data analysis methods leads to difficulty in interpreting results. This increases the probability of drawing erroneous conclusions. Building upon existing subjective responses is resource friendly and helps develop machine learning (ML) based visual quality predictors. In this work, we show that using a discrete model for analyzing responses from MQA subjective experiments is feasible. We indicate that our proposed Generalized Score Distribution (GSD) properly describes response distributions observed in typical MQA experiments. We also highlight interpretability of GSD parameters and indicate that the GSD outperforms the approach based on sample empirical distribution when it comes to bootstrapping. Furthermore, we provide evidence that the GSD outcompetes the state-of-the-art model both in terms of goodness-of-fit and bootstrapping capabilities. To accomplish the aforementioned objectives, we analyze more than one million subjective responses from over 30 subjective experiments. Finally, we make the code implementing the GSD model and related analyses available through our GitHub repository: https://github.com/Qub3k/subjective-exp-consistency-check.**

***Index Terms*—Discrete distribution, generalised score distribution, GSD, subjective experiments, quality of experience.**

I. Introduction

THERE are phenomena that require gathering opinions from a panel of people. One significant example here is the notion of Quality of Experience (QoE). Contrary to the Quality of Service (QoS), the QoE also depends on how a user of a system perceives its performance (with the word *perceives* assuming the greatest significance here). Although technical factors do influence the QoE, ultimately, it is a subjective opinion of a user that represents the most direct indication of the QoE. (Refer to Sec. 2.2.2 of [1] for a formal definition of the QoE.)

Multimedia Quality Assessment (MQA) is a sub-field of the QoE related research activities. It focuses on understanding how people perceive quality of multimedia content as well as the effects of its processing and performance of multimedia services. It is a common and recommended [2][3] practice to organize experiments in which a panel of observers provides its opinion on the quality of multimedia materials presented. We refer to such experiments as *subjective experiments* and to the opinions provided by the panel of observers as *subjective responses*. Importantly, we narrow our discussion down to subjective experiments in which participants judge the technical reproduction quality of stimuli presented only. In other words, we do not take into account subjective experiments, where observers voice their opinions regarding the content of stimuli (e.g., plot of a story or artistic properties of an image).

Lack of access to ground truth information is an inherent feature of subjective experiments. Put differently, we observe subjective responses, but have no way of directly measuring the quality of a given stimulus. One solution to this problem is gathering a large number of responses per stimulus. By doing so, we are able to ensure that any summary statistic we use to estimate stimulus quality adequately reflects population level opinions. An increasing number of researchers are following this intuition and switching from small scale controlled experiments to large scale crowdsourcing experiments [4]. Unfortunately, switching to crowdsourcing experiments usually corresponds to less precise measurements. On the other hand, organizing large scale controlled subjective experiments is money- and time-intensive. For these reasons, we want to draw as many learnings as possible from limited information controlled subjective experiments provide.

To fully use the information that controlled subjective experiments provide, we cannot merely rely on summary statistics only (among which the Mean Opinion Score or MOS is the most popular [2][3]). Instead, we need to construct models that try to capture the underlying, unobservable structure of subjective responses and understand how this structure maps to

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599. Furthermore, the work was supported by the PL-Grid Infrastructure. *(Corresponding author: Jakub Nawała.)*

J. Nawała, L. Janowski, and K. Rusek are with the Institute of Telecommunictions, AGH University of Science and Technology, 30059 Kraków, Poland. (e-mail: jakub.nawala@agh.edu.pl)

B. Ćmiel is with the Department of Mathematical Analysis, Computational Mathematics and Probability Methods, AGH University of Science and Technology, 30059 Kraków, Poland.

P. Pérez is with Applications & Platforms Software Systems, Nokia Bell Labs, Madrid, Spain.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a detailed description of selected parts of the methodology used in this paper. This material is 372 KB in size.

quality. To construct such models, we use various assumptions based on domain knowledge and experiences gathered from previous subjective experiments.

There are better and worse models. Likewise, there are tools to assess how well a model performs. We claim that using models reflecting data type that subjective responses represent is a better approach than assuming that continuous models can be applied to discrete data. For one thing, models reflecting underlying data type generate interpretable results. Increased interpretability makes it easier to understand the result, thereby protecting against ill posed conclusions.

A. Problem Statement and Contributions

Subjective responses from Multimedia Quality Assessment (MQA) experiments are conventionally analyzed with methods not suitable for data type that these responses represent. In particular, continuous models are used even though subjective responses are discrete in most cases [5][6]. Furthermore, obtaining subjective responses is money- and time-intensive. Thus, a method that allows the reuse of existing responses would be beneficial.

Applying improper data analysis methods may lead to results that are difficult to interpret. This, in turn, may result in erroneous conclusions. Liddell and Kruschke provide a convincing overview of mistakes that arise when data is analyzed using an improper model [7]. One of our goals is to protect researchers analyzing responses from MQA experiments against the mistakes Liddell and Kruschke mention.

In terms of building upon existing subjective responses, the approach is especially important if it is used to generate large samples from small real-life samples. This procedure is also referred to as *bootstrapping*. Properly applied bootstrapping allows for the production of sample sizes sufficient for developing machine learning (ML) visual quality predictors. Naturally, reusing existing subjective responses is also resource friendly.

We show that it is feasible to use a discrete model for the data analysis of responses from MQA subjective experiments is feasible. We also present benefits stemming from this approach. Specifically, we indicate that our proposed Generalized Score Distribution (GSD) properly describes response distributions observed in typical MQA experiments. We also highlight interpretability of GSD parameters. This GSD feature makes it possible to easily describe and intuitively understand non-trivial dependencies between various response distributions. Finally, we point out that the GSD outperforms the traditional approach based on a sample empirical distribution when it comes to bootstrapping.

Our work is novel in two respects. First, to the best of our knowledge, the GSD is the first two-parameter discrete distribution proposed in the field of MQA that properly models per stimulus response distribution. Second, we are the first ones to demonstrate that our subjective response modelling approach (i.e., the GSD) outperforms the standard approach based on empirical distribution in regard to bootstrapping.

Being more suitable for bootstrapping than sample empirical distribution, the GSD can generate a large data set of responses by taking advantage of only a small data set of real-life responses. In turn, large data sets generated this way may allow for the creation of next generation ML based perceptual visual quality predictors. This is because ML based solutions are data hungry by nature, and typical MQA experiments are capable of gathering only few dozens of responses per stimulus. Moreover, knowing a correct data model (which we show the GSD is for typical MQA experiments) allows for the proposal of a parametric hypothesis testing framework. Using such a framework results in higher power (when compared to conventionally used here non-parametric methods), thus allowing the reduction of costs related to organizing subjective experiments. This is because more powerful statistical methods allow the detection of smaller effect sizes, while keeping the sample size constant. Finally, interpretability of GSD parameters makes it easier to summarize subjective responses and perform non-misleading intuitive inferences based on this summary.

With this work, we put forward the following contributions:

- We evidence that analyzing subjective responses from MQA experiments with a discrete model (specifically, the GSD) is feasible and brings easy to interpret results.
- We indicate that the GSD properly describes responses from typical MQA subjective experiments.
- We show that the GSD outperforms empirical distribution when it comes to bootstrapping for responses from MQA subjective experiments.
- Finally, we demonstrate that the GSD outperforms the state-of-the-art model when it comes to bootstrapping and goodness-of-fit testing.

The main main objective of this paper is to convince the MQA research community that using the GSD model to analyze subjective responses is better than following current recommendations and the practices put forward in the literature. Specifically, we want to demonstrate that the GSD outperforms non-disrete models used in the literature and that the GSD also performs better than the standard approach based on a sample empirical distribution when it comes to bootstrapping. To make it easier for others to use our work, we invite everyone to visit our GitHub repository (https://github.com/Qub3k/subjective-exp-consistency-check). There, we provide a code that allows the analysis of subjective responses with the use of the GSD model and to reproduce a significant part of results presented in this paper.

B. Related Works

There is a trend in the MQA community to favour response distribution analysis over relying on summary statistics (e.g., the MOS) only. An important recent contribution in this topic is the work by Seufert [8]. There, he highlights fundamental advantages of considering response distributions over summary statistic-based evaluations. Hoßfeld *et al.* take this idea further and show how to approximate response distributions given a QoS-to-MOS mapping function [9]. Our work follows the trend of response distribution analysis. At the same time, we indicate that interpretable GSD model parameters can serve

as summary statistics by adequately describing underlying response distribution.

Modelling individual responses generation process is another important thread of MQA research focusing on response distribution analysis. The idea was first proposed by Janowski and Pinson and termed *subject model*[1] [10]. Li and Bampis took on the approach and proposed an extended subject model [5]. In their formulation of the model, they considered subject bias, subject inconsistency and stimulus ambiguity. Reference [6] proposes an updated, simpler version of the same model. Authors of [6] convincingly show that their model addresses the shortcomings of subjective data analysis methods presented in several MQA-related ITU recommendations. Our work extends this arc of research. We model individual responses generation process with the Generalized Score Distribution (GSD) model. The model was introduced in [11], where we showed how it could be applied to check subjective responses consistency. Recently, we also made available a paper formally describing the GSD family of distributions [12]. There, we highlighted and provided mathematical proofs of a few important properties of the GSD family. There, we also revealed the details underlying the GSD parameter estimation procedure.

We are not the first ones to notice that subjective responses modelling approach should reflect data it operates on. Specifically, both [13] and [14] propose models that take into account ordinal nature of subjective responses coming from MQA experiments.

Compared to our other works on the GSD [11], [12], this paper puts forward a series of new contributions. Unlike [12], it specifically targets practitioners from the MQA community. It thus focuses on GSD properties relevant for this community. This paper also compares the GSD to a state-of-the-art model from the MQA community and checks GSD's performance on data going beyond typical MQA experiments. Both those analyses are novel when compared to [12]. Unlike [11], this paper focuses on the general applicability of the GSD to MQA data analysis (instead of showing only one applicability area of the model). Overall, this paper concentrates on the broad consequences of using the GSD to model MQA data, rather than limiting itself to presenting the GSD as a tool that resolves one particular problem [11] or focuses on a formal description of the model [12].

II. Methodology

In this section, we describe the methodology we use to substantiate the claims made in the introduction. Section II-A elucidates our idea of treating subjective responses from MQA experiments as realizations of a discrete random variable. Section II-B shows how we test the goodness-of-fit of the models that we take into account. It also presents how we interpret resulting p-values. Section II-C highlights the data sets that we use to test the GSD on real data. Finally, Sec. II-D details the procedure that we use to test GSD's performance when it comes to subjective responses bootstrapping.

[1]The word *subject* refers to subjective experiment participant.

A. Subjective Response as a Random Variable

We propose to think about responses from MQA subjective experiments as realizations of a discrete random variable U. Since we focus on responses expressed on the 5-level Absolute Category Rating (ACR) scale (cf. Sec. 6.1 of [15]), U can take values from the $\{1, 2, 3, 4, 5\}$ set. To make the distribution of U practically useful, we need to parametrise it. Our experiences show that distributions with one parameter do not properly fit real data. Thus, we focus on two-parameter distributions. The following shows a general formulation of such distributions

$$U \sim F(\lambda, \theta), \tag{1}$$

where $F()$ denotes a cumulative distribution function, λ is a parameter describing central tendency of the distribution and θ expresses distribution spread. Importantly, we assume that $F()$ reflects the response distribution of each stimulus in a subjective experiment. Per-stimulus values of λ and θ define the exact shape of $F()$.

Now, there are at least two approaches to proposing the exact formulation of $F()$. The first one (which is more popular in the MQA literature) is to assume that subjective responses follow a continuous normal distribution. The second one (which we take in this paper) is to assume that responses follow a discrete distribution and, more precisely, the Generalised Score Distribution (GSD).[2]

The approach assuming that subjective responses follow continuous normal distribution is best described by introducing an intermediate continuous random variable $O \sim \mathcal{N}(\mu, \sigma^2)$, where μ describes the mean and σ^2 is the variance of the normal distribution. Since U is discrete and O is continuous, we need to introduce a mapping between the two. In other words, O must be discretized and censored as follows:

$$P(U = s) = \int_{s-0.5}^{s+0.5} \frac{1}{2\pi\sigma} e^{-\frac{(o-\mu)^2}{2\sigma^2}} do \tag{2}$$

for $s = \{2, 3, 4\}$ and

$$P(U = 1) = \int_{-\infty}^{1.5} \frac{1}{2\pi\sigma} e^{-\frac{(o-\mu)^2}{2\sigma^2}} do, \tag{3}$$

$$P(U = 5) = \int_{4.5}^{\infty} \frac{1}{2\pi\sigma} e^{-\frac{(o-\mu)^2}{2\sigma^2}} do. \tag{4}$$

Such a construct (i.e., a thresholded cumulative normal distribution) is quite popular in latent variable analysis [16]. Thus, we follow the appropriate nomenclature and refer to this model as Ordered Probit.

1) Generalized Score Distribution: Our approach to modelling subjective responses does not require any mapping between an intermediate random variable and U. This is because the GSD already is a discrete distribution. Thus, we can directly write $U \sim GSD(\psi, \rho)$, where ψ expresses the so called true quality and ρ expresses responses spread. The true quality parameter ψ can be intuitively understood as a mean response for a given stimulus, if we were to ask for the opinion of the complete population of observers. Contrary to

[2]Although GSD's name refers to scores, we use the word "responses" to refer to opinions formulated by observers taking part in a MQA experiment.

Ordered Probit's μ, ψ reflects the 5-level ACR scale and is bounded to the $[1, 5]$ range.[3] The other GSD's parameter, ρ, is a linear function of $V(U)$ (i.e., variance of U). Furthermore, ρ is bounded to the $[0, 1]$ interval and expresses what portion of possible variance is present in realizations of U. Please note here that any discrete distribution with a limited domain (e.g., $U \sim F(\lambda, \theta)$) has its mean value $E(U)$ and variance $V(U)$ bounded (cf. Fig. 2). One more important property of ρ is that it represents responses confidence. Put differently, it is inversely proportional to the variance observed in responses (the higher the observed variance, the lower the value of ρ). Yet another way to put it is to say that the greater the value of ρ, the closer to ψ observed responses are. Importantly, the GSD is able to model the complete range of possible variances for a given M-point scale (with $M \in \mathbb{N} : M > 2$). For more details regarding the GSD, we refer the reader to [12].

To further concretize GSD's description, let us take a look at its internal structure. We start by showing a more detailed form of the $U \sim GSD(\psi, \rho)$ expression:

$$U \sim \psi + \epsilon, \tag{5}$$

where ϵ expresses uncertainty regarding the mean response represented by ψ. ψ is one constant number estimated for a stimulus of interest. Notice that $\psi = E(U)$. ϵ, on the other hand, follows a distribution parameterized with a single parameter ρ. Furthermore, ϵ's distribution satisfies the following two criteria: (i) its mean equals zero and (ii) its variance is a linear function of ρ. In Appendix A (see the supplemental material), we show the exact formulation of ϵ's distribution. Here, we only mention that this distribution is a mixture of the following distributions: binomial, beta-binomial, and one- or two-point distribution (whether one- or two-point distribution is used depends on the value of ψ). Importantly, we reparameterize the distributions in the mixture to make them satisfy the two criteria that ϵ's distribution must follow. As a result, the reparameterised distributions in the mixture depend only on a single parameter ρ.

Fig. 1 illustrates various realisations of the GSD for different values of ψ and ρ. Please notice how flexible the GSD is. For example, in Fig. 1c, for the case of $\rho = 0.38$, the GSD takes the form of a distribution with two modes (one mode at response 1 and another at response 5). Apart from this extreme example, GSD's shape follows common sense intuition regarding the response distributions observed in typical MQA subjective experiments. (For an in-depth discussion regarding response distribution shapes acceptable by the GSD, we refer the reader to [11].)

B. G-test and P–P Plot

In order to validate if a distribution (or a model) fits specific data, we need to perform a two-step procedure. The first step is to estimate distribution parameters for a sample of interest. The second step is to test a null hypothesis stating that the sample truly comes from the assumed distribution (GSD or Ordered Probit in our case), given the parameters estimated in the first step. We use a standard likelihood ratio approach to test the goodness-of-fit (GoF) of the two models (the GSD and Ordered Probit). More precisely, we use the so called G-test of GoF (cf. Sec. 14.3.4 of [17]). We do not use the asymptotic distribution for calculating the p-value because sample sizes we consider are predominantly small. On the contrary, we estimate the p-value utilising a bootstrapped version of the G-test. Please refer to Appendix B in the supplemental material to learn exactly how we use the G-test of goodness-of-fit. (For broader theoretical considerations on the topic, please take a look at [18].)

Since each MQA subjective experiment that we analyze contains multiple stimuli, we need to perform the G-test multiple times (as many as there are stimuli in the experiment). The result of each G-test is a p-value. This means we get a vector of p-values for each experiment that we take into account. To be able to efficiently draw conclusions regarding a vector of p-values, we use p-value P–P plots (where P–P stands for probability–probability) [19]. For a detailed discussion regarding p-value P–P plots for the GSD, we refer the reader to [11].

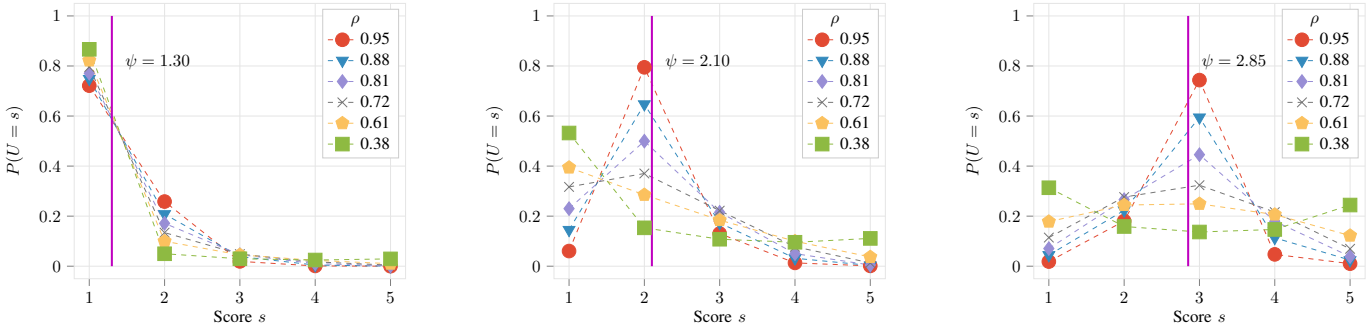
C. Data Sets

To test the GSD in practice, we make use of more than one million individual subjective responses (to be precise, 1 183 696). We take into account the responses coming from 33 subjective experiments that assess quality (or other traits) of more than nine thousand stimuli (exactly 9 290). Table I presents an overview of data sets we use. Importantly, we classify data sets into three types: (i) typical, (ii) broadly understood, and (iii) non-MQA. The types reflect how much a given data set follows best practices and recommendations regarding organizing MQA experiments. *Typical* experiments tightly follow best practices and recommendations. *Broadly understood* experiments follow these best practices and recommendations generally, but deviate from them in some aspects. Finally, *non-MQA* experiments are not MQA experiments at all. We include these to check GSD's performance on data outside of GSD's intended application scope. Please note that one data set may correspond to multiple experiments (cf. the "No. of Exp." row of Table I). For example, the MM2 data set consists of 10 separate experiments. Thus, although we use data from 11 data sets, they amount to 33 experiments.

We do not provide detailed descriptions of the data sets here. Instead, we link to references describing each data set in Table I. The only exception to this rule is the NFLX data set. Since its description has not yet been published, we describe the data set briefly.

Experiments included in the NFLX data set investigated the influence of per-scene quality changes on the opinions of human observers. Two hundred observers assessed quality of 320 stimuli.[4] Ten seconds long video clips (without audio) were used as stimuli. The clips had a resolution of 1920x1080 pixels. Quality degradations were applied solely through video compression. However, since per-scene compression was used,

[3]To make the discussion easy to comprehend, we limit ourselves to the version of the GSD reflecting a 5-level scale. However, the GSD can describe any discrete process with domain of size M, where M is a natural number greater than 2.



(a) (b) (c)

Fig. 1. Realizations (in the form of probability mass functions) of the GSD for a 5-point scale and various values of parameters ψ and ρ. Notice how the growing value of ρ corresponds to more responses accumulating close to the value of ψ.

TABLE I
AN OVERVIEW OF DATA SETS WE USE TO TEST THE GSD ON REAL DATA

Study	ITU [20]	HDTV [21]	MM2 [22]	14-505 [23]	ITS4S [24]	NFLX	ITERO [25]	ITS4S2 [26]	Naderi [27]	MovieLens [28]	Personality [29]
Year	1995	2010	2012	2014	2018	2018	2019	2019	2020	2003	2018
No. of Exp.	1	6	10	1	2	4	3	1	3	1	1
Total No. of Stimuli	176	1 008	600	114	1 025	720	330	1 429	170	3 708	10
Total No. of Responses	4 224	24 192	12 780	7 076	26 926	36 000	25 080	22 864	22 511	1 000 209	1 834
Type	typical	typical	typical	typical	typical	typical	bu	bu	bu	non-MQA	non-MQA
Stimulus type	speech	video	av	video	video	video	video	image	speech	mr	mr

Exp. stands for experiments; av stands for audiovisual; mr stands for movie recommendation; bu stands for broadly understood.

quality switches occurred during playback as well. Contents spanned a wide range of categories and were taken from Netflix's catalogue. This made the experiments more ecologically valid, but also meant that the clips could not be publicly shared. The clips were displayed on either a TV or a tablet (both with the native resolution of 1920x1080 pixels). Moreover, some participants were asked to provide their opinions during the video playback. They were encouraged to use a software slider displayed at the bottom of the screen. In total, four experiments were performed: (i) with the TV and the software slider, (ii) with the TV and without the slider, (iii) with the tablet and the slider, and (iv) with the tablet and without the slider. Participants were recruited through a temporary working agency. Care was taken not to over-represent the 18 to 25 age group. All four experiments were carried out in a controlled environment and were generally following the provisions of Rec. ITU-T P.913 [2]. The experiments were performed in accordance with the Absolute Category Rating with Hidden Reference (ACR-HR) method (cf. Sec. 7.2.2 of [2]). Thus, participants provided their responses using the 5-level ACR scale.

D. Bootstrapping

To compare GSD's generalizability to that of the empirical distribution (which is typically used for resampling), we introduce the following procedure. We start by generating MC (e.g., $MC = 10\,000$) bootstrap samples from the empirical probability mass function (EPMF) of the large sample. Importantly, we generate bootstrap samples with significantly fewer observations than those in the large sample (e.g., $n = 24$ observations in each bootstrap sample for $N = 200$ observations in the large sample). Next, we fit the GSD to each r-th bootstrap sample. This yields estimates of each response category probability $(\hat{q}_1^r, \hat{q}_2^r, \hat{q}_3^r, \hat{q}_4^r, \hat{q}_5^r)$. We use those estimates to calculate the likelihood function for the large sample $\mathcal{L}_{GSD}^r$. We repeat the procedure for each bootstrap sample, but use the EPMF of the bootstrap sample this time to find the response category probability estimates $(\hat{v}_1^r, \hat{v}_2^r, \hat{v}_3^r, \hat{v}_4^r, \hat{v}_5^r)$. Having the likelihood for both the GSD $(\mathcal{L}_{GSD}^r)$ and empirical distribution $(\mathcal{L}_e^r)$, we introduce a statistic W_r based on the quotient of the two values. In other words, we introduce a statistic based on the likelihood ratio: $W_r = \ln(\mathcal{L}_{GSD}^r / \mathcal{L}_e^r)$. Value of the quotient signifies which approach better describes the large sample. (Note that there are as many quotients as there are bootstrap samples.) Now, we use the quotients to estimate the probability p_{GSD} that the GSD model-based estimates of response category probabilities in the large sample yield a higher likelihood function value $(\mathcal{L}_{GSD})$ than the likelihood function value we get if we use the EPMF-based estimates $(\mathcal{L}_e)$. We also do the same for the empirical distribution and estimate the probability that the EPMF-based estimates yield a higher likelihood function value than that yielded by the GSD model-based estimates and denote this probability

[4]In Table I we write about 720 stimuli in this data set, since we treat each of the four experiments as separate. Because each of the four experiments investigated 180 stimuli, we end up with 720 stimuli in total.

by p_{e}. We then calculate the 95% confidence interval for $p_{\mathrm{GSD}} - p_{\mathrm{e}} = P(W_r > 0) - P(W_r < 0)$ and denote its lower (or left) bound as L and upper (or right) bound as R. If $L > 0$, then the GSD performs better than the empirical distribution. If $R < 0$, then the empirical distribution performs better. If $[L, R]$ contains zero, there is no significant difference between the GSD and empirical distribution. We provide the precise description of the aforementioned procedure in Appendix C (refer to the supplemental material).

Since we use the subsample to make inferences about the large sample, there is a risk of overfitting. Put differently, by fitting any model too precisely to the subsample, we are confronted with the risk of finding model parameter estimates that are suboptimal from the point of view of the large sample. This is because the subsample represents only limited information about the large sample. Intuitively, we should not entirely trust the data that we observe in the subsample. To address this issue, we apply parameter estimation modification that prevents probability estimators we use from yielding response category probabilities equal to 0 (for any response category). In other words, we expect that, at the population level, there is no response category that would be assigned no observations (even if the estimation result for the subsample suggests something else). This results in modified estimation procedures for both the GSD and empirical distribution. The detailed estimation correction procedures that we use are described in Appendix C-A (refer to the supplemental material).

III. Results

Here, we present the results reflecting our contributions mentioned in the introduction. Sec. III-A puts forward the evidence supporting the claim that the GSD has easy to interpret parameters. Sec. III-B shows that the GSD properly describes response distributions from typical MQA experiments. It also indicates that GSD does not perform well for atypical MQA and non-MQA experiments. Sec. III-C reveals that the GSD outperforms empirical distribution when it comes to subjective responses bootstrapping. Finally, Sec. III-D evidences that the GSD outperforms the state-of-the-art model both in terms of goodness-of-fit testing and bootstrapping.

A. Interpretable Parameters

Fig. 2 presents how Ordered Probit model parameters map to the $E(U)$ and $V(U)$ space. In other words, the figure shows how the parameters of the Ordered Probit model that we use to describe observed data (cf. Fig. 2a and Fig. 2e) map to summary statistics computed directly on these observed data (Fig. 2b and Fig. 2f). Intuitively, Fig. 2a and Fig. 2e illustrate how the Ordered Probit model "sees" observed data. Fig. 2b and Fig. 2f show us how observed data actually look like in terms of two basic summary statistics (i.e., mean $E(U)$ and variance $V(U)$). Put differently, any point along any line in Fig. 2a or Fig. 2e corresponds to a fixed pair of Ordered Probit parameters. The Ordered Probit model with these parameters is then used to generate discrete responses (being realizations of the random variable U). Summary statistics (i.e., $E(U)$ and $V(U)$) computed on these generated responses yield a single point in Fig. 2b or Fig. 2f, respectively. (Note that these generated responses can be thought of as representing individual responses that we observe in real subjective experiments.) Importantly, plots in Fig. 2 should be analyzed in pairs, row-wise. In other words, the leftmost (red) line in Fig. 2a corresponds to the same data series as the leftmost (red) line in Fig. 2b. The same is true for Fig. 2e and Fig. 2f, and so on. When analyzing Fig. 2, please also keep in mind that $E(O) = \mu$ and $V(O) = \sigma^2$ (cf. Sec. II-A for more context).

We want model parameters to accurately reflect phenomena occurring in observed data. For example, we naturally associate the μ parameter with the central tendency of observed data and the σ parameter with their variance. Thus, if we keep μ constant and increase the value of σ, we expect this should correspond to $E(U)$ staying constant and $V(U)$ to increase. However, this is not the case. Instead, keeping μ constant and increasing σ corresponds to changes in both $E(U)$ and $V(U)$. This can be observed by following same-colored lines[5] in Fig. 2a and Fig. 2b. Specifically, let us take the leftmost (red) line from Fig. 2a. It corresponds to Ordered Probit's μ fixed at a value slightly larger than zero. Moving vertically upwards along this line, μ stays constant and σ increases. If we were to stop at various points along this line and generate discrete responses (being realizations of the random variable U) from the Ordered Probit model with μ and σ parameters fixed, we expect each such sample to have a constant and same expected value $E(U)$, but a changing variance $V(U)$. The corresponding leftmost (red) line in Fig. 2b shows the $E(U)$ and $V(U)$ we actually observe when generating responses from the Ordered Probit model. As shown in the figure, the samples generated do not have a constant expected value. On the contrary, it changes in rather unexpected ways, as we move along increasing the values of σ. (The only exception to this rule is when $\mu = 3$.) This property of Ordered Probit model parameters makes them counter-intuitive. Unfortunately, this is not the only limitation of Ordered Probit's parameterisation. Another one relates to how changes in μ correspond to changes in $E(U)$. Looking at Fig. 2e and Fig. 2f, we see that the same range of μ values maps to different ranges of $E(U)$ values as the σ parameter changes. For example, let us compare the topmost pink curve ($\sigma = 8$) with the second topmost green one ($2 < \sigma < 4$). In Fig. 2e they both span the same range of μ values. However, in Fig. 2f, the pink curve corresponds to a much narrower range of $E(U)$, in comparison to the green curve. This leads us to another limitation of Ordered Probit's parameterisation. Although subjective responses that we take into account here span the range from 1 to 5, the μ parameter takes values exceeding the 1–5 range. Practically speaking, although it is tempting to treat μ as an MOS-related measure, μ can and will exceed the 1–5 range (which the MOS never does). Thus, μ should not be intuitively interpreted as MOS counterpart for the Ordered Probit model. To ensure completeness, we mention that the Ordered Probit model is able to describe the complete ghost-like area shown in Fig. 2b and Fig. 2f. However, this is only possible if we allow its

[5]The ordering of lines in Fig. 2a and Fig. 2e is the same as the ordering of lines in Fig. 2b and Fig. 2f. Thus, the figure can be interpreted in black-and-white print as well.

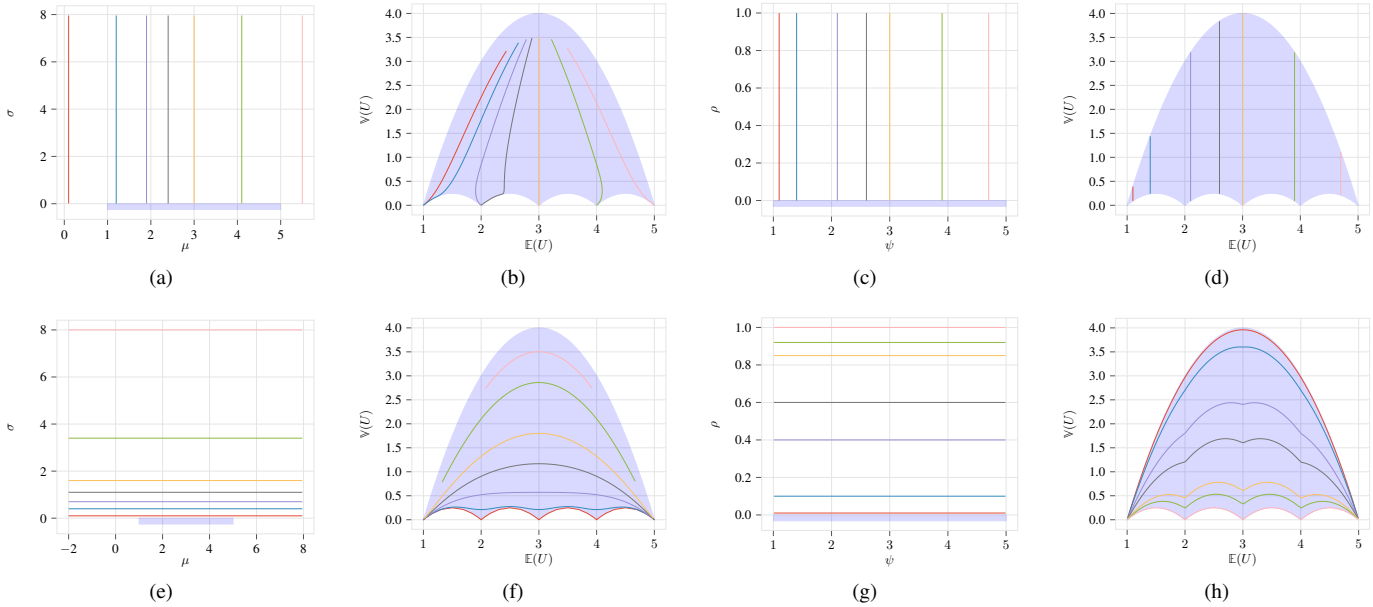


Fig. 2. Mapping of Ordered Probit parameters to the $E(U)$ and $V(U)$ space (plots (a), (b), (e) and (f)). Mapping of GSD parameters to the $E(U)$ and $V(U)$ space (plots (c), (d), (g) and (h)). The violet area marks all possible $(E(U), V(U))$ pairs for a discrete process with values $\{1, 2, 3, 4, 5\}$. The violet bar below plots (a), (c), (e) and (g) shows the 1–5 range (reflecting the range of values of random variable U).

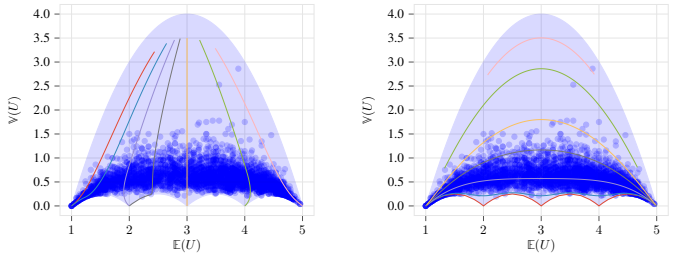


Fig. 3. Mean response ($\mathbb{E}(U)$) and response variance ($\mathbb{V}(U)$) pairs of all stimuli from the typical MQA experiments shown along with (a) the lines with constant μ parameter of the Ordered Probit model for varying σ parameter and (b) the lines with constant σ parameter for varying μ. To understand the meaning of the curves in the plot, please refer to Fig. 2.

parameters to change without bounds. In other words, when $(\mu, \sigma) \in (-\infty, +\infty) \times (0, +\infty)$.

The right hand side of Fig. 2 is GSD's counterpart of Ordered Probit plots on the left. Fig. 2c and Fig. 2g present GSD parameters space. Fig. 2d and Fig. 2h present the $E(U)$ and $V(U)$ space. (Note that there is an inverse relationship between ρ and $V(U)$.) As can be readily seen, the GSD is not fraught with problems inherent to the Ordered Probit model. In particular, keeping the ψ parameter constant and changing ρ parameter's value, we keep the $E(U)$ constant and vary $V(U)$ only. This means that GSD's parameterisation allows for treating ψ as observed data's central tendency and ρ as a measure of their variability. Following same-colored lines in Fig. 2c and Fig. 2d evidences how keeping ψ constant corresponds to constant $E(U)$. Notably, the same range of ψ values for different values of ρ, always corresponds to the same range of $E(U)$. We can take a curve of any color from Fig. 2g and Fig. 2h, and see that it always spans the entire range of $E(U)$. Although the bumpy shape of multiple curves in Fig. 2h may initially seem counter-intuitive, it reflects an important property of ρ. The ρ parameter expresses what ratio of available variance for a given mean is present in observed data. Thus, to keep this ratio constant across different mean values, the curve has to follow the bottom part of the $E(U)$ and $V(U)$ space. Thanks to its properties, $\rho = 0.5$ means that we deal with data being at the midpoint between minimum and maximum possible variance. Finally, GSD parameters cover the entire space of $E(U)$ and $V(U)$, while staying in the well defined bounds. Specifically, $(\psi, \rho) \in [1, 5] \times [0, 1]$. Practically speaking, ψ can be regarded as GSD's counterpart of the MOS.

Here, it is noteworthy that both models share one limitation. Even when data variability related parameter (σ or ρ) stays constant and the central tendency related parameter (μ or ψ) changes, $V(U)$ changes across different values of $E(U)$. Ideally, $V(U)$ should follow the variability related parameter and stay constant across chaning $E(U)$. However, since we are dealing here with a discrete, limited domain process (only values $\{1, 2, 3, 4, 5\}$ can be observed), the mean is naturally coupled with variance. In other words, changes to the mean inherently influence variance.

In order to show that problems with model parameters interpretability do apply to real data, we overlay on top of Fig. 2b and Fig. 2f $(E(U), V(U))$ pairs of real response distributions. Specifically, we take all stimuli from typical MQA experiments (cf. Tab. I). Then, we compute per stimulus mean response $(E(U))$ and per stimulus response variance $(V(U))$. Finally, we place each $(E(U), V(U))$ pair on top of the ghost-like shapes shown in Fig. 2b and Fig. 2f. Fig. 3 presents the end result.

Looking at Fig. 3, we can make a few observations. First, it

is clear that the $(E(U), V(U))$ pairs cover the entire range (1 to 5) of possible mean responses. Second, the pairs correspond mostly to variances in the interval 0 to 1.5 (although there are points corresponding to variance almost as high as 3). However, the most important observation is that the data cloud in Fig. 3a covers the area where the vertical curves bend. For example, looking at the second left-most (blue) curve, it becomes evident that there are data points that fall along this curve. Please recall that this curve corresponds to a constant Ordered Probit's μ with a value of roughly 1.2. Although μ stays *constant*, points falling along the blue curve correspond to *varying* values of mean response. This indicates that problems with Ordered Probit's parameterisation do apply to real data as well. Please also note that since the data cloud in Fig. 3 covers the entire range of mean responses, the Ordered Probit model is forced to use μ exceeding the 1 to 5 range (cf. Fig. 2a). Again, this confirms that applying Ordered Probit's parameterisation in practice is problematic. Finally, we observe that the problems with Ordered Probit's parameterisation described in this paragraph do *not* apply to GSD's parameterisation. In other words, GSD's ψ never exceeds the range of 1 to 5, and the data points corresponding to different values of $E(U)$ also correspond to different values of ψ.

B. Good Description of Typical MQA Experiments

Fig. 4a shows the results of applying a bootstrapped G-test of goodness of fit to responses from typical MQA experiments, as modelled by the GSD or by the Ordered Probit model. If any of the two models truly reflects response distributions observed in real data, a related p-value histogram should resemble the uniform distribution (or any other nonincreasing distribution) in the region of low p-values (roughly between 0 and 0.2) [11]. It can be clearly seen that the histogram for the Ordered Probit model does not resemble the uniform distribution. The most important indication of this fact is the height of the leftmost bar, which is significantly greater than that of the second leftmost bar. GSD's histogram does resemble the uniform distribution for the p-values region of interest. However, to decisively assess GSD's performance, we need to resort to p-value P–P plot (cf. Fig. 5). Since all GSD-related data points fall below the black diagonal line, we can safely infer that the results do not contradict the null hypothesis of the GSD truly reflecting response distributions observed in real data. In other words, the GSD adequately reflects response distributions observed in typical MQA experiments. However, the same is not true for the Ordered Probit model. This is indicated by all Ordered Probit related data points falling above the black diagonal line. Put differently, the Ordered Probit model does not properly reflect response distributions observed in typical MQA experiments.

If we now also consider MQA experiments that do not strictly follow international recommendations (let us call them *broadly understood MQA experiments*), we see that the performance of the both models deteriorates (cf. Fig. 4b). This is best indicated by the height of the leftmost bar. On both histograms, its height is significantly greater than the reference height corresponding to approximately 279 stimuli or 5% of all stimuli investigated. We do not show a related P–P plot since it simply reaffirms that both models do not reflect response distributions observed in real data.

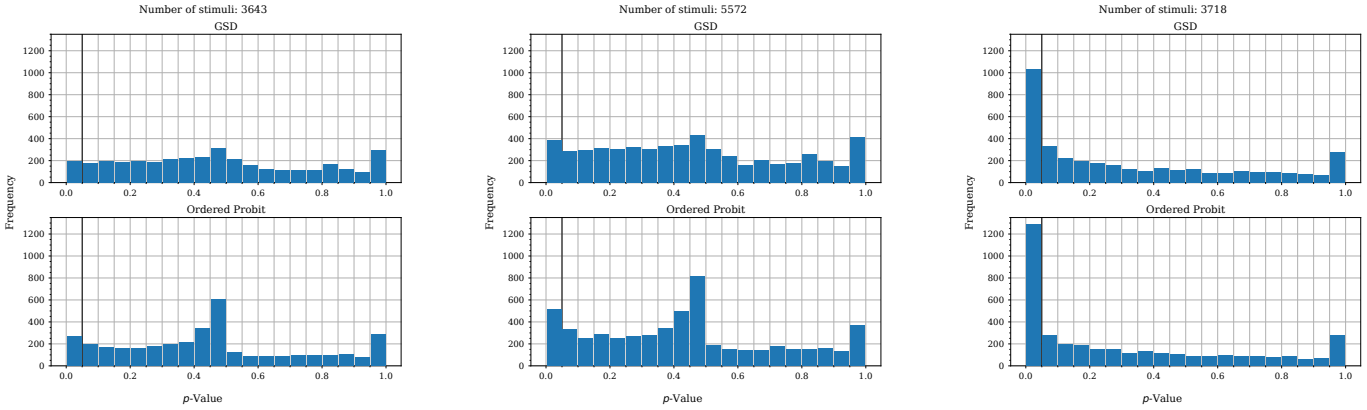
We also investigated how the GSD and Ordered Probit models would perform on a data set unrelated to MQA experiments. To this end, we chose two data sets popular in the movie recommendation systems research community: (i) MovieLens 1M [28] and (ii) Personality 2018 [29]. Although the two data sets are outside of MQA, they use the 5-level Likert scale to collect subjective responses. Our hypothesis was that since the GSD performed well on MQA data using the 5-level Liker scale, then it would probably also perform well on these data sets. However, looking at Fig. 4c, we can clearly see that both the GSD and Ordered Probit models do not reflect response distributions observed in the data. In other words, neither the GSD nor Ordered Probit model properly describe response distributions observed in data concerning movie recommendation systems.

C. Better Than Empirical Distribution

It is interesting to determine whether the GSD brings any advantage if it comes to generalizability. We define generalizability as the ability of a model to capture large sample phenomena when observing only a subsample of the large sample. In particular, we would like to ascertain whether the GSD better captures large sample's distribution shape in comparison to the empirical distribution of the subsample. Put differently, we would like to check whether the GSD is better suited for bootstrapping than the empirical distribution. If this proves to be the case, then the GSD could be used for resampling. One important consequence of this would be an opportunity to build better machine learning (ML) models aimed at predicting subjectively perceived multimedia quality (which is a difficult, important and still open challenge). It is often the case in the field of Multimedia Quality Assessment (MQA) that only up to 30 responses per stimulus are available. If one wants to create an ML model, this may prove insufficient and therefore, resampling must be applied to generate more responses per stimulus. Should the GSD prove to be better for bootstrapping than the empirical distribution (which is typically applied in this context), the GSD could be used to generate more reliable samples during resampling.

To test GSD's generalizability capabilities in practice, we use data from four MQA studies: (i) MM2 [22], (ii) VQEG HDTV Phase I [21], (iii) NFLX (cf. Section II-C to learn more about this study) and (iv) ITERO [25]. From these studies, we extract responses for selected stimuli. More precisely, we select stimuli with at least 144 responses. This way, we get 234 stimuli. The number of responses per stimulus spans from 144 to 228. There are only four unique numbers of responses per stimulus. Table II shows the four numbers of responses and the count of stimuli with a given number of responses. Furthermore, it shows the study from which a given set of stimuli was taken.

One may wonder why we use only selected experiments and not all experiments given in Tab. I. This is attributed to the



(a) (b) (c)

Fig. 4. p-Value histograms for the GSD (upper) and Ordered Probit (lower) models. p-Values come from the G-test of goodness-of-fit applied to stimuli from (a) typical Multimedia Quality Assessment (MQA) experiments, (b) typical and broadly understood MQA experiments and (c) non-MQA experiments. The thick vertical line marks the 0.05 significance level.

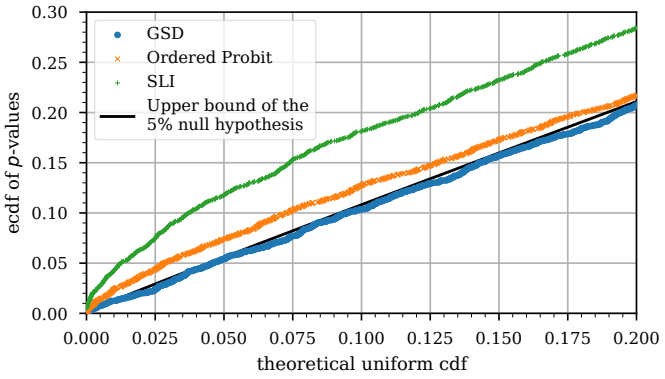


Fig. 5. p-Value P–P plot for typical MQA experiments. p-Values come from the G-test of goodness-of-fit applied to the GSD, Ordered Probit and Simplified Li2020 (SLI) models, fitted to responses from typical MQA experiments. CDF stands for cumulative distribution function and ECDF for empirical cumulative distribution function.

TABLE II
DISTRIBUTION OF RESPONSES AMONG THE FOUR STUDIES USED IN THE BOOTSTRAP ANALYSIS. HDTV CORRESPONDS TO VQEG HDTV PHASE I STUDY.

No. of Responses	No. of Stimuli	Study
144	24	HDTV
200	40	NFLX
213	60	MM2
228	110	ITERO

fact that for the analysis presented in this section, we need experiments with a high number of responses per stimulus. Typical MQA experiments gather roughly between 12 to 50 responses per stimulus. Since we focus on the bootstrapping capabilities of a model in this analysis, we need significantly more observations per stimulus than the standard 12 to 50 range. We decide to use a slightly arbitrary threshold of 100 responses per stimulus. In other words, we take into account only those experiments that provide at least 100 responses per stimulus. Most of the experiments listed in Tab. I do not satisfy this requirement at all or satisfy it only under some special assumptions. Overall, in this section, we use the experiments coming from the four studies listed above (i.e., MM2, VQEG HDTV Phase I, NFLX, and ITERO). The following paragraphs describe a few special assumptions we had to make in order to include these experiments in the analysis.

The NFLX study contains responses given to stimuli displayed on one of the two display devices—tablet and TV. In principle, responses for different display devices shall be analyzed separately. However, since the responses for the two devices are highly correlated and since the same visual content was presented to participants during the sessions with each device, we decide to include the combined responses from the two display devices in this analysis.[6]

If it comes to the HDTV Phase I study, we only focus on responses provided to the so called *common set* of stimuli. The stimuli from the common set were presented to participants in all the six experiments that were part of the HDTV Phase I study. Although the six experiments were conducted by different research teams and used different display devices, the experimenters declared that actions were taken to make the six experiments similar to each other. Specifically, all video stimuli were displayed with the same resolution and in a room conforming to guidelines of Rec. ITU-R BT.500-11. Following experimenters declaration, we combine responses for the common set stimuli. That is, we treat the six experiments, with 24 participants each, as one large experiment with 144 participants. This way, we end up with 24 stimuli (that many are in the common set), each having 144 responses.

The MM2 study is a set of ten experiments. Responses in the experiments were collected by six laboratory teams from four countries. Different subject pools and environments were used in each experiment. The common denominator of all the experiments was the same set of 60 audiovisual stimuli and a very similar test procedure. According to the authors

[6]The exact correlation between mean responses for the two display devices is 0.988. The scatter plot of mean responses is shown in Fig. 8 in App. D (see the supplemental material).

of [22], the experiments were highly repeatable. Thus, we combine responses from the ten experiments. This yields 213 responses (that many participants in total took part in the ten experiments) for each of the 60 audiovisual stimuli.

The ITERO study collected responses from 27 subjects, who rated the same set of 110 stimuli. The study was carried out by three research teams. The experiment design was atypical of how MQA experiments are usually conducted. Subjects were instructed to repeat the experiment ten times. In total, 110 stimuli were assigned 228 responses each (however, not all subjects repeated the experiment ten times). Although these subjects were allowed to repeat the experiment at their leisure and the majority of them did not use the same display device, we combine the responses from the ten repetitions. In other words, we treat the responses as though they come from one large subjective experiment with 110 stimuli and 228 subjects (wherein each subject rates the same set of 110 stimuli).

We utilize three small sample sizes, i.e., $n = \{12, 24, 50\}$. This allows us to observe how the GSD performs (when compared to the empirical distribution) for different fractions of the large sample information available. Intuitively, we expect the empirical distribution's performance to improve as the small sample size increases. If the GSD proves to perform differently than the empirical distribution, we would observe how the increasing small sample size influences the difference between the two approaches. Here, we emphasize that the increasing small sample size always favors the empirical distribution. On the other hand, the performance of the GSD depends on how well it fits to the distribution of responses observed in the large sample. If the fit is good, the increasing small sample size also favors the GSD. On the other hand, if the fit is poor, the increasing small sample size does not necessarily improve GSD's performance.

Fig. 6a presents the results of the analysis. They take the form of three histograms. These histograms visualize probability differences $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}}$ for the three investigated small sample sizes (i.e., 12, 24 and 50). Now, greater probability mass to the right of 0 indicates that the GSD performs better than the empirical distribution. Greater probability mass to the left of 0 corresponds to the opposite situation, i.e., empirical distribution outperforms the GSD. To simplify this analysis, in the plot, we show red hatched bars that indicate for how many stimuli the GSD outperforms the empirical distribution (the red hatched bar on the right) and for how many the empirical distribution performs better than the GSD (the red hatched bar on the left). Blue-colored parts of the bars represent statistically insignificant probability differences. When assessing which approach performs better, these blue parts of the bars are discarded.

Clearly, the GSD outperforms empirical distribution for all three small sample sizes. The effect is most clearly visible for the small sample size of 12. As expected, as the size of a small sample grows, empirical distribution's performance improves as well. Nevertheless, even for as many as 50 observations per small sample (which rarely happens in typical MQA subjective experiments), the GSD still significantly outperforms the empirical distribution.

According to the results, the GSD is a better choice than the empirical distribution (which is typically used in this context) when it comes to the resampling of subjective responses from MQA studies. This opens up an opportunity to train better ML models for the MQA applications, without having to organize large subjective experiments (i.e., experiments with a large number of participants). This result is yet another indication of GSD's superiority over methods typically used for MQA data analysis.

D. Comparison With the State-of-the-Art Model

To the best of our knowledge, the GSD is the first two-parameter discrete distribution proposed in the field of MQA that models per stimulus subjective responses. Thus, there are *no* state of the art solutions that could be directly compared with the GSD. Existing solutions either rely on discretizing continuous probability distribution [6], [30], require more data than per stimulus subjective responses only [6], [13], or use a multinomial distribution to model subjective responses [30], [13].

The problem with solutions discretizing a continuous probability distribution is that their parameters are difficult to interpret. Such models suffer from a similar set of problems as the Ordered Probit model does (cf. Sec. III-A). If it comes to solutions that require access to more data than responses for a stimulus of interest only, these cannot be directly applied to our goodness-of-fit and bootstrap capabilities testing frameworks. In fact, such solutions are fundamentally different from the GSD in the sense that they try to model the entire subjective experiment, rather than a response distribution of a single stimulus. Finally, solutions that utilize a multinomial distribution can be practically equated to using an empirical distribution of a sample of interest. For example, if we were to extract from [30] and [13] just the part of the model describing the per stimulus distribution, we would get a four-parameter multinomial distribution. Since the data set on which we operate in this paper consists of subjective responses expressed on the *five*-level ACR scale, the empirical distribution of per stimulus responses always takes the form of a *four*-parameter multinomial distribution. In other words, models based on a four-parameter multinomial distribution do not bring about any reduction in the number of parameters. Thus, in principle, they do not confer any advantage over using the empirical distribution of a sample.

Another challenge that we confront when trying to compare the GSD with other similar solutions is that we would have to modify the existing solutions first. Put differently, to the best of our knowledge, there are *no* existing models that could be used as a drop-in replacement for the GSD. We argue that comparing the GSD with a modified version of existing solutions would be unfair and provide very limited information regarding GSD's performance.

To avoid modifying existing solutions, we could follow the methodology proposed by Gao *et al.* in [31]. There, they take several continuous probability distributions (among others, Gaussian, half normal and Weibull distributions) and discretize them before comparing them with their solution. We claim that following this methodology would *not* be a good idea as well.

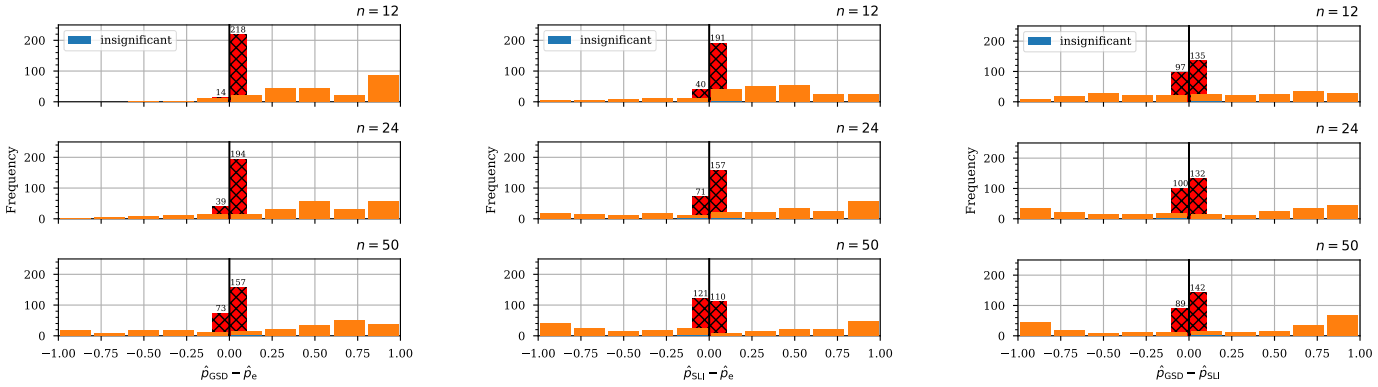


Fig. 6. Histograms depicting the distribution of probability differences $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}}$ in (a), $\hat{p}_{\mathrm{SLI}} - \hat{p}_{\mathrm{e}}$ in (b), and $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{SLI}}$ in (c). Three small sample sizes are considered: 12, 24 and 50. Blue-coloured parts of the bars represent statistically insignificant probability differences. (There are so few insignificant results that they are barely visible.) Red bars with the hatching indicate the sum of probability differences to the right and to the left of zero (excluding insignificant results).

First, we do not know exactly how to discretize the continuous distributions selected for comparison. Specifically, it is unclear how to divide the probability mass of continuous distributions to map it to a discrete domain that is of our interest.[7] Second, even if we discretized a continuous distribution, we would end up with a distribution with difficult to interpret parameterisation (cf. the discussion about the Ordered Probit model in Sec. III-A). Finally, the authors of [31] show that even the best performing discretized continuous distribution is able to properly model only 44.4% of the response distributions that they consider. As discussed in Sec. III-B, the GSD is able to properly model *all* response distributions originating from typical MQA experiments (assuming the 5% significance level for the goodness-of-fit hypothesis testing).

Notwithstanding the discussion above, we do compare the GSD with a modified version of one model from the literature. More specifically, we compare the GSD with a modified version of the solution presented in [6]. To make the description easier to comprehend, we refer to the model from [6] as Li2020 model. Although we need to modify the Li2020 model (before it could be compared with the GSD), we decide to use it anyway. This is primarily due to the Li2020 model's popularity in the MQA community. Suffice to say that, to the best of our knowledge, the Li2020 model is currently the only candidate for ITU standardization if it comes to modelling of subjective responses from MQA experiments.

The GSD operates only on responses given to a single stimulus. In other words, the GSD needs to know only about these subject responses that were assigned to a single stimulus of interest. The Li2020 model requires information regarding all responses of all subjects that scored the stimulus of interest. Put differently, even though we are interested in responses assigned to a single stimulus, we need to know about all responses assigned by a given subject to all other stimuli in the experiment to estimate the model parameters. Neither the bootstrapped G-test nor the bootstrapping effectiveness test we use satisfy Li2020 model's requirements. Both tests rely on the assumption that responses assigned to the single stimulus of interest are sufficient for the model.

Not to abandon the comparison between the GSD and Li2020 models completely, we simplify the Li2020 model. Specifically, we make it function in a manner similar to that of the GSD. Put differently, we make the Li2020 model only require responses assigned to a single stimulus of interest. This results in a model defined by a Gaussian probability density function (PDF) with its mean (μ) equal to sample mean (i.e., the MOS) and variance (σ^2) equal to the sample variance (s^2). (Note that "sample" here means a set of responses assigned to a single stimulus of interest.) In the following text, we refer to the modified Li2020 model as the *Simplified Li2020* model or SLI for short.

After estimating Simplified Li2020 model's parameters, we end up with a continuous normal distribution $\mathcal{N}(\mathrm{MOS}, s^2)$. Since real subjective responses take the form of discrete numbers ($\{1, 2, 3, 4, 5\}$ in our case), we need to map from the continuous domain of the normal distribution to the 5-level scale of interest. To this end, we proceed in the same manner as we do when fitting the Ordered Probit model to the data. Specifically, we apply equations (2), (3), and (4).

Although the Ordered Probit and Simplified Li2020 models look very similar, they are not identical. The key difference lies in model parameters estimation. The Simplified Li2020 model assumes observed subjective responses are realizations of a continuous random variable following the normal distribution. Importantly, realizations of such a random variable can take any value (from plus to minus infinity). Hence, observing values from the $\{1, 2, 3, 4, 5\}$ set exclusively is, probabilistically speaking, very rare. The Simplified Li2020 model ignores this fact and fits the normal distribution to these data using sample mean and variance.[8] Contrary to the Simplified Li2020 model, the Ordered Probit model does not assume that observed subjective responses are realizations of a continuous random variable. More precisely, the continuous normal distribution

[7]Unfortunately, the authors of [31] provide very few details regarding their methodology to be able to reproduce their method of discretizing the continuous distributions that they consider.

[8]This approach exemplifies what Liddell and Kruschke warn against in [7].

present in the Ordered Probit model is treated as a latent trait of the data. This latent continuous distribution is always mapped to a discrete scale of interest first (cf. (2), (3) and (4)), before fitting the model. Finally, although we describe the Simplified Li2020 model here, the same discussion applies to the original Li2020 model as well. Put differently, the original full model also assumes that observed subjective responses are realizations of a continuous normal random variable (even though these responses only take values from the $\{1, 2, 3, 4, 5\}$ set).

1) G-test of Goodness-of-Fit: Let us first check how the SLI model performs when it comes to describing response distributions observed in typical MQA experiments. In this regard, we will use the same G-test-based procedure, as the one we applied to the GSD in Sec. III-B. Fig. 5 shows the comparison of GSD, SLI and Ordered Probit in terms of G-test results. Since only GSD data points fall below the black diagonal line, it is the only model that properly reflects response distributions observed in typical MQA experiments. Furthermore, the SLI model performs worse than both the GSD and the Ordered Probit models. Performance inferior to the Ordered Probit model may be ascribed to SLI's lack of mapping to the 5-level scale, when estimating its parameters. In short, the SLI and Ordered Probit models do not properly describe response distributions observed in typical MQA experiments, whereas the GSD model does.

2) Bootstrapping: We now test whether the SLI model can perform better than the GSD if it comes to bootstrapping. To ascertain this, we apply the same procedure to the SLI model that we had applied to the GSD model in Sec. III-C. The result is given in Fig. 6b. As shown in the figure, the SLI model performs better than the empirical probability mass function (EPMF) for small samples of size 12 and 24. However, it performs worse than the EPMF for small samples of size 50. Figure 6c presents the result of directly comparing the GSD with the SLI model. Put succinctly, the GSD outperforms the SLI model for all small sample sizes.

IV. Discussion

Section III-A shows that GSD's parameterisation is more interpretable and intuitive when compared to that of Ordered Probit. Importantly, Ordered Probit's parameterisation is not erroneous. Still, using it may lead to mistaken conclusions, if used carelessly. Our results indicate that GSD's parameterisation should be preferred over that of Ordered Probit. This insight is relevant for the MQA research community since many practitioners decide to first try using continuous models (Ordered Probit being one of them) when they start working with subjective responses modelling. Arguably, their preference to choose continuous models stems from easier availability of methods operating on such models. We can also argue that continuous models elicit more attention during standard statistics and probability classes and thus, are a nature choice when it comes to data modelling. We want to protect MQA practitioners against potential mistakes arising from the use of continuous models to analyze discrete data. Our results indicate that the GSD is a viable and better alternative to continuous models when it comes to subjective responses analysis (with responses expressed on a discrete scale).

Section III-B reveals that the GSD adequately describes responses from typical MQA experiments. This property of the GSD indicates that the GSD can serve as a basis for building parametric methods for subjective responses analysis. Notably, the power of parametric methods is greater than that of their non-parametric counterparts. For example, a parametric hypothesis testing framework can detect a smaller effect size for a given sample size, when compared to a nonparametric framework. This increased power may prove essential when analyzing responses from controlled subjective experiments. Since such experiments usually take place in a laboratory environment and require the direct involvement of a researcher, they can become resource intensive (both money- and time-wise). It is desirable (or sometimes necessary) to reduce the sample size of such experiments.[9] A parametric GSD-based data analysis framework would help address this problem. Due to its parametric nature, it would be able to detect smaller differences between various test conditions for a given sample size, in comparison to other nonparametric methods.

Importantly, neither the GSD nor the Ordered Probit model properly describe response distributions observed in broadly understood or non-MQA subjective experiments (cf. Fig. 4b and Fig. 4c). This implies that the models are not globally applicable to modelling subjective responses expressed on the 5-level Likert scale. Potentially, more complicated models (i.e., models with more than two parameters) are necessary to model phenomena present in responses from broadly understood or non-MQA experiments.

Although our results indicate that the GSD does not properly model responses coming from broadly understood and non-MQA experiments, we strongly believe that the model can be applied to data sets other than those originating from the MQA community. Specifically, we expect the GSD to function well in all those situations where we expect people to agree and where their responses are expressed on a discrete scale. Put differently, in MQA research, we assume that although we gather *subjective* responses, these are the *objective* characteristics of the stimulus that decide what is the consensus opinion of human observers. Whenever this line of thinking applies to a given data set, there is a strong likelihood that the GSD will properly model the related response distributions.

Sec. III-C makes it clear that the GSD outperforms the traditional approach (based on empirical distribution) when it comes to subjective responses bootstrapping. This result means that whenever there is a need to generate more results from a small real-life sample, the GSD should be preferred over empirical distribution to perform resampling. Such resampling may prove necessary when building an ML-based perceptual quality predictor. Building such a predictor requires a significant amount of data. Sufficiently large sample sizes may be difficult to generate through a controlled experiment. For this reason, a small real-life sample can be collected through a controlled experiment. Then, the GSD-based bootstrapping can

[9]In MQA experiments sample size usually corresponds to the number of people invited to assess quality of a set of stimuli.

be used to enlarge the small real-life sample to a larger sample (of a size sufficient for building an ML-based perceptual quality predictor). Significantly, having such a mechanism at hand also addresses the issue of controlled experiments being money- and time-intensive. As discussed previously, a small and not so expensive experiment may be organized to generate a small real-life sample of responses. This sample can then be enlarged using the GSD-based bootstrapping to achieve a sample size that would otherwise require organizing a larger and more expensive controlled experiment. At this point, we would like to remind the reader that our results indicate that the mechanism described above applies to responses from typical MQA experiments exclusively.

Looking at Sec. II-D, the reader may wonder whether there are methods of checking GSD's bootstrapping capabilities other than comparing the GSD with the empirical distribution. Naturally, the answer is yes. For example, one could use either the Akaike Information Criterion (AIC) [32] or the Bayesian Information Criterion (BIC) [33]. Both AIC or BIC could be computed for the GSD and the SLI model. Then, the resultant AIC or BIC values could be compared and the model with a lower value selected as the winner. However, since both AIC and BIC are based on the number of model parameters and the maximized value of the likelihood function of the model, using those two measures would lead to the same conclusions as the ones presented in Sec. III-D. This is because both the GSD and SLI model have the same number of parameters and because the GSD always attains a higher likelihood value (at least when the likelihood is computed as shown in point 2-c) of the procedure given in App. C in the supplemental material).

Sec. III-D evidences that the GSD outperforms the state-of-the-art model, namely the Simplified Li2020 (SLI) model from [6]. GSD's superiority is clear in terms of the goodness-of-fit testing for data from typical MQA experiments. Out of the three models tested (GSD, SLI and Ordered Probit), only the GSD properly describes response distributions observed in the data. If it comes to bootstrapping, the SLI, similarly to the GSD, outperforms the empirical distribution for small sample size of 12 and 24. However, GSD's improvement over the empirical distribution is greater for the two cases. Furthermore, only the GSD outperforms the empirical distribution for a small sample size of 50. Finally, when we directly compare the GSD with the SLI model, the former performs better for all small sample sizes.

In Sec. III-D, we mention that the Li2020 model [6] is currently the most popular method of subjective data analysis in the MQA community. We also say that the model has been presented to ITU as a candidate for standardization. The reader may be surprised to see that the Li2020 model performed rather poorly in the analyses that we presented in Sec. III. A few things require consideration here. First, we do *not* use the Li2020 model as is. Put differently, we simplify its structure to make it function in accordance with the GSD model. This means that the conclusions we reach do not necessarily apply to the full formulation of the Li2020 model presented in [6]. Second, the Li2020 model takes into account (and corrects for) subject bias present in subjective responses. Undoubtedly, this is a positive feature of that model. For example, [34] shows that Li2020 model's ability to take into account subject bias makes it function very well as a subjective experiment precision estimator. Importantly, the simplified version of the Li2020 model, that is, the SLI, does *not* take into account subject bias. Undoubtedly, this impacts model performance. Finally, although we argue in this paper that GSD's ability to model discrete data directly is something good, it comes at a cost. Specifically, if subjective responses are, for some reason, non discrete (e.g., 3.2 instead of 3), the GSD *cannot* be used to model them. The Li2020 model does not suffer from this limitation. In other words, the Li2020 model can be used to analyze both discrete and non-discrete subjective responses.

V. Conclusion

Our work substantiates the following four claims:

1) The GSD has interpretable parameters that clearly and intuitively describe response distribution shape (for responses gathered in MQA subjective experiments).
2) The GSD properly models response distributions observed in typical MQA subjective experiments.
3) The GSD is better suited for bootstrapping of responses from MQA subjective experiments in comparison to the traditional approach based on empirical distribution.
4) The GSD outperforms the state-of-the-art model in terms of goodness-of-fit testing (on data from typical MQA experiments) and bootstrapping.

The results indicate that the GSD-based bootstrapping of subjective responses from MQA experiments can be used to build new ML-based perceptual quality predictors, without having to organize large-scale controlled experiments. This makes it possible to build ML-based predictors cheaper than would otherwise be possible.

We hope that our discussion regarding interpretable GSD parameters and risks inherent to using continuous models to analyze discrete subjective responses, will convince the MQA research community to reconsider current best practices and recommendations.

There are at least three directions that our future work may take. First, we would like to build a ML-based perceptual quality predictor. In this regard, we plan to use the GSD-based bootstrapping. Second, we would like to propose a GSD-based parametric hypothesis testing framework for the analysis of subjective responses from MQA experiments. Third, we aim to test GSD's performance on other openly available data sets with subjective responses [35], [36], [37]. Importantly, we need access to *individual* subjective responses for the latter to be possible. Many researchers make available only aggregated data (e.g., MOS scores [36], [37] or the number of responses per response category [35]). Thus, we would like to openly ask the researchers in the MQA community to publicly share their data in the per response format. In other words, we would ask them to structure their data so that there are as many data rows as there are individual responses gathered in the course of a subjective experiment. Each row should state the experiment in which the participant issued the response and to which stimulus. In general, the data format should conform to the rules of the so-called *tidy data* [38]. Only the data sets that

are structured this way can help extend the work presented in this paper.

Finally, we invite everyone to use the GSD to analyze subjective responses from their experiments and to make use of the tools presented in this paper. Our GitHub repository (https://github.com/Qub3k/subjective-exp-consistency-check) contains software tools that make it easier to start using the GSD. We hope the model and related tools will allow other MQA researchers and practitioners to analyze their data more efficiently and effectively.

ACKNOWLEDGMENT

We would like to express our gratitude to Netflix, Inc. for sponsoring initial stages of this research and for the funding they provided to organize and conduct the NFLX experiment (cf. Sec. II-C). We would also like to thank Anush Krishna Moorthy for coordinating the NFLX experiment and providing valuable feedback on a draft version of this work. We acknowledge the helpful and constructive comments of the members of the Video Quality Experts Group's (VQEG) Statistical Analysis Methods (SAM) group as well. Last but not least, we also warmly acknowledge comments and suggestions from Zhi Li of Netflix.

REFERENCES

[1] S. Möller and A. Raake, Eds., *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer International Publishing, 2014. [Online]. Available: http://link.springer.com/10.1007/978-3-319-02681-7

[2] ITU-T, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," Rec. ITU-T P.913, 6 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.913/en

[3] ITU-R, "Methodologies for the subjective assessment of the quality of television images," Rec. ITU-R BT.500-14, 2020. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500/en

[4] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-vq: 'patching up' the video quality problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2021, pp. 14 019–14 029.

[5] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data Compression Conference (DCC)*, 2017, pp. 52–61.

[6] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Electronic Imaging*, vol. 2020, pp. 131–1–131–14, 1 2020. [Online]. Available: https://www.ingentaconnect.com/content/10.2352/ISSN.2470-1173.2020.11.HVEI-131

[7] T. M. Liddell and J. K. Kruschke, "Analyzing ordinal data with metric models: What could possibly go wrong?" *Journal of Experimental Social Psychology*, vol. 79, pp. 328–348, 2018. [Online]. Available: https://doi.org/10.1016/j.jesp.2018.08.009

[8] M. Seufert, "Statistical methods and models based on quality of experience distributions," *Quality and User Experience*, vol. 6, pp. 1–27, 2021. [Online]. Available: https://doi.org/10.1007/s41233-020-00044-z

[9] T. Hoßfeld, P. E. Heegaard, M. Varela, L. Skorin-Kapov, and M. Fiedler, "From qos distributions to qoe distributions: a system's perspective," in *2020 6th IEEE Conference on Network Softwarization (NetSoft)*. IEEE, 6 2020, pp. 51–56. [Online]. Available: https://ieeexplore.ieee.org/document/9165426/

[10] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec 2015.

[11] J. Nawała, L. Janowski, B. Ćmiel, and K. Rusek, "Describing subjective experiment consistency by p-value p-p plot," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 10 2020, pp. 852–861. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413749

[12] B. Ćmiel, J. Nawała, L. Janowski, and K. Rusek, "Generalised score distribution: Underdispersed continuation of the beta-binomial distribution," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2204.10565

[13] J. Li, S. Ling, J. Wang, and P. L. Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 10 2020, pp. 3339–3347. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413619

[14] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests-beyond subjects' mos," *IEEE Transactions on Multimedia*, vol. 23, pp. 2505–2519, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9154548/

[15] ITU-T, "Subjective video quality assessment methods for multimedia applications," Rec. ITU-T P.910, 2008. [Online]. Available: http://handle.itu.int/11.1002/1000/9317

[16] W. E. Becker and P. E. Kennedy, "A graphical exposition of the ordered probit," *Econometric Theory*, vol. 8, pp. 127–131, 1992.

[17] A. Agresti, *Categorical Data Analysis*, 2nd ed. Wiley, 2002.

[18] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[19] T. Schweder and E. Spjøtvoll, "Plots of P-Values to Evaluate Many Tests Simultaneously," *Biometrika*, vol. 69, no. 3, pp. 493–502, 12 1982. [Online]. Available: https://doi.org/10.1093/biomet/69.3.493

[20] ITU-T Study Group 12, "ITU-T Coded-Speech Database," 1998. [Online]. Available: http://handle.itu.int/11.1002/1000/4415

[21] M. Pinson, F. Speranza, M. Barkowski, V. Baroncini, R. Bitto, S. Borer, Y. Dhondt, R. Green, L. Janowski, T. Kawano *et al.*, "Report on the validation of video quality models for high definition video content," *Video Quality Experts Group*, 2010. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[22] M. H. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, oct 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6286980/

[23] M. H. Pinson and L. Janowski, "Agh/ntia: A video quality subjective test with repeated sequences," NTIA Technical Memo TM-14-505, 6 2014. [Online]. Available: https://www.its.bldrdoc.gov/publications/details.aspx?pub=2758

[24] M. H. Pinson, "Its4s: A video quality dataset with four-second unrepeated scenes," NTIA Technical Memorandum 18-532, 2 2018. [Online]. Available: https://www.its.bldrdoc.gov/publications/details.aspx?pub=3194

[25] P. Perez, L. Janowski, N. Garcia, and M. Pinson, "Subjective assessment experiments that recruit few observers with repetitions (fowr)," pp. 1–12, 2021. [Online]. Available: http://arxiv.org/abs/2104.02618

[26] M. H. Pinson, "Its4s2: An image quality dataset with unrepeated images from consumer cameras," NTIA Technical Memorandum 19-537, Apr. 2019. [Online]. Available: https://www.its.bldrdoc.gov/publications/details.aspx?pub=3219

[27] B. Naderi, T. Hosfeld, M. Hirth, F. Metzger, S. Moller, and R. Z. Jimenez, "Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 5 2020, pp. 1–6. [Online]. Available: http://arxiv.org/abs/2003.11300https://ieeexplore.ieee.org/document/9123115/

[28] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, 12 2015. [Online]. Available: https://doi.org/10.1145/2827872

[29] T. T. Nguyen, F. M. Harper, L. Terveen, and J. A. Konstan, "User personality and user satisfaction with recommender systems," *Information Systems Frontiers*, vol. 20, pp. 1173–1189, 2018.

[30] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Modeling image quality score distribution using alpha stable model," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1574–1578.

[31] ——, "Parameterized image quality score distribution prediction," *arXiv preprint arXiv:2203.00926*, 2022.

[32] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[33] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, p. 461–464, 1978. [Online]. Available: https://doi.org/10.1214/aos/1176344136

[34] J. Nawała, T. Hoßfeld, L. Janowski, and M. Seufert, "Systematic analysis of experiment precision measures and methods for experiments comparison," *arXiv*, 2022, submitted for review to IEEE Transactions on Multimedia. [Online]. Available: http://arxiv.org/abs/2204.07131
[35] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.
[36] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3674–3683.
[37] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3572–3582.
[38] H. Wickham, "Tidy data," *Journal of Statistical Software*, vol. 59, pp. 1–23, 2014. [Online]. Available: https://www.jstatsoft.org/v059/i10
[39] M. Pagano and K. Gauvreau, *Principles of Biostatistics*. Chapman and Hall/CRC, 2018.

Jakub Nawała is a PhD candidate in Information and Communication Technology at the Institute of Telecommunications at AGH University of Science and Technology (AGH UST) in Krakow, Poland. His research interests include subjectively perceived multimedia quality, quality of experience, subjective testing, and subjective data analysis. Jakub Nawała received the B.Eng. and M.Eng. degree in electronics and telecommunications from AGH UST, graduating with honors both times, in 2017 and 2018, respectively. He is an author of 10 scholarly articles and has participated in 11 national and international research projects. He was the recipient of the award for outstanding achievements granted by the Polish Ministry of Science and Higher Education in 2016 and 2017.

Lucjan Janowski received the Ph.D. degree in telecommunications from the AGH University of Science and Technology, Krakow, Poland, in 2006. In 2007, he was a Postdoctoral Researcher with the Laboratory for Analysis and Architecture of Systems, Centre National de la Recherche Scientifique, Paris, France. From 2010 to 2011, he was a Postdoctoral Researcher with the University of Geneva, Geneva, Switzerland. From 2014 to 2015, he was a Postdoctoral Researcher with the Telecommunications Research Center Vienna, Vienna, Austria. He is currently an Assistant Professor with the Institute of Telecommunications, AGH University of Science and Technology. His research interests include statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation.

Bogdan Ćmiel received the PhD degree in Mathematics at the Department of Applied Mathematics, AGH University of Science and Technology in Krakow, Poland, in 2013. In 2014 he was a postdoctoral researcher at the Institute of Mathematics, Polish Academy of Sciences. He is currently an Assistant Professor at the AGH University of Science and Technology. His research interests include theoretical statistics (mainly nonparametric statistics), statistical modeling and applied statistics in medicine and engineering.

Krzysztof Rusek received the Ph.D. degree in queuing theory from the AGH University of Science and Technology in 2016. He has worked as a system administrator and machine learning engineer with the research group focused on the processing and protection of multimedia content. He is currently an Assistant Professor at the AGH University of Science and Technology and also a Data Scientist at the Barcelona Neural Networking Center, UPC. He is currently working on the applications of Graph Neural Networks and probabilistic modeling for performance evaluation of communication systems and data mining in astronomy. His main research interests include performance evaluation of telecommunication systems, machine learning, and data mining.

Pablo Pérez is Lead Scientist at Nokia Extended Reality Lab (Madrid, Spain). He is a Telecommunication Enginner (BSc+MSs, 2004) and a PhD in Telecommunications (2013) from Universidad Politécnica de Madrid, Spain, and Nokia Distinguished Member of Technical Staff (2022). He has worked as R&D engineer of digital video products and services in Telefónica, Alcatel-Lucent and Nokia; as well as a researcher in future video technologies in Nokia Bell Labs. He is currently leading the scientific activities of Nokia XR Lab, addressing the end-to-end technological chain of the use of Extended Reality for human communication: networking, system architecture, processing algorithms, quality of experience and human-computer interaction.