The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model

Lucjan Janowski and Margaret Pinson

***Abstract*—How accurately are people able to use the absolute category rating (ACR) 5-level scale? Put another way, how repeatable are an individual subject's scores? Several subjective experiments have asked subjects to rate the same sequences a couple of times. Analyses indicate that none of the subjects exactly repeated their prior scores for these sequences. We would like to better understand this imperfection. This paper uses ACR subjective video quality tests to explore the precision of subjective ratings. To make formal measurements possible, we propose a theoretical subject model that is the main contribution of this paper. The proposed subject model indicates three major factors that influence accuracy: subject bias, subject inaccuracy, and stimulus scoring difficulty. These appear to be separate random effects and their existence is a reason why none of the subjects were able to perfectly repeat scores. There are three key consequences. First, subject scoring behavior includes a random component that spans approximately half of the rating scale. Second, the sensitivity and accuracy of most subjective analyses can be improved if the subject scores are normalized by removing subject bias. Third, to some extent, multiple subjects can be replaced with a single subject who rates each sequence multiple times.**

***Index Terms*—Design of experiments, mean opinion score, quality of experience (QoE), subject model, subjective ratings, video quality assessment.**

I. Introduction

SUBJECTIVE experiments are key tools that link technical solutions with human perception. A typical subjective experiment presents a particular aspect of a service to a group of subjects, to validate the service in specific way. The collected answers are used to reach conclusions and make product development decisions.

A. Motivation

The goal of many subjective experiments is to revel the Quality of Experience, typically scored on a scale ranging from "Bad" to "Excellent." A subjective experiment, as any measuring process, contains errors. The most common way to deal with those errors is focusing on mean opinion scores (MOS), which are more stable. While MOSs portray elegant trends, individual subjective ratings are messy indeed. Each subject's ratings have a confidence interval (CI), an error term if you will. Subjects use the scale differently, and these differences can seem troubling. Our inclination is to seek order in this chaos. We believe that understanding the source and magnitude of a single subject error helps to plan and analyze more precisely future subjective experiments. Especially it allows to use simulation as a tool analyzing an experiment with low cost before the real subjects are invited to the laboratory. An example of simulation analysis of a subjective experiment can be found in [1]. A validated subject model makes such simulation accurate and more general.

Some researchers proactively seek improved subjective test methods. One example is to change the scale by using Double Stimulus Continuous Quality Scale (DSCQS), Double Stimulus Impairment Scale (DSIS), ACR 11-grade scale (ACR11), PC (Pair Comparison), or another scale, in the hope that the precision of a subject's ratings will improve. Alas, comparisons between discrete and continuous scales indicate no improvement [2], [3], suggesting that subjective test methodology is not the main reason why the obtained results are so strongly scattered. The interesting question is, why?

Other researchers reactively reduce chaos by screening subjects. The goal is to eliminate irrelevant subjects, such as people who misunderstood the task. That is difficult, so we rely upon post-screening algorithms with somewhat arbitrary thresholds (e.g., Clause 11.3 of ITU-T Rec. P.913). The consequence is to discard subjects with noisy data—regardless of whether or not these are valid subjects. The scientific method instead demands that we seek a deeper understanding of subject rating behaviors and account for these human failings in our technique.

B. Contribution

The main innovation within this paper is the modeling of subject rating behavior. We propose an equation that describes subject rating behaviors; we name this equation the "theoretical subject model." Thanks to this equation, different characteristics of subjects and subjective experiments can be derived.

The main goal of the paper is to prove that the propose model is valid, covers different aspects of subjects behavior, and can be used to improve subjective experiments. It is important to understand that the proposed model does not model the quality of a video or other stimulus. The goal is to model the process of a subject giving an answer.

The remainder of this paper is structured as follows. We begin by examining the rating behaviors detected in prior subjective tests. Next, we present a subject scoring model and use discrete

Manuscript received February 11, 2015; revised June 03, 2015 and July 26, 2015; accepted September 25, 2015. Date of publication October 02, 2015; date of current version November 13, 2015. The work of L. Janowski was supported by Contract 11.11.230.018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonggang Wen.

L. Janowski is with the AGH University of Science and Technology, Kraków 30-059, Poland (e-mail: janowski@kt.agh.edu.pl).

M. Pinson is with the Institute for Telecommunication Sciences, Boulder, CO 80305 USA (e-mail: margaret@its.bldrdoc.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2484963

TABLE I
NULL IMPAIRMENT RATINGS FOR ITS1 DATASET

DMOS	4.66	4.89	4.89	4.91	4.93	4.96
Stdev	0.52	0.48	0.31	0.28	0.25	0.20

mathematics to analyze subject rating behaviors within discrete subjective testing data. Then we derive formulae that measure specific terms in our subject model. We look for direct evidence supporting our subject model by applying these formulae to subjective datasets, and then look for indirect proof by examining behaviors predicted by our subject model. We close by examining the practical implications of our subject scoring model.

II. RELATED WORK

Publicized subjective quality tests occasionally include limited statistics on the distribution of rating changes. Most of this information comes from analyses of the original stimuli. This is called a hidden reference or the null impairment.

Let us begin with subjective video quality experiment ITS1, conducted in 1992 [4]. The ITS1 experiment used the double stimulus impairment scale (DSIS). With DSIS, the subject watches the original video, watches the impaired video, and then rates the severity of the impairment on a discrete 5-level scale:

1) imperceptible;
2) perceptible but not annoying;
3) slightly annoying;
4) annoying; and
5) very annoying.

The ITS1 experiment was conducted on broadcast quality, uncompressed analog format tapes (Betacam-SP). None of the original videos were digitally compressed, so the only likely difference between two copies of the original video was a small amount of analog noise due to copy generation or repeated tape use.

ITS1 included the null impairment for six original video sequences (SRC). Thus, the original video was played twice and the difference rated. These null impairments ratings are shown in Table I. From this, we see that subjects do not always rate the null impairment as imperceptible. The SRC with the lowest differential mean opinion score (DMOS) depicts the beginning of a car race and it is the most difficult to code.

The T1A1 subjective experiment [5] was conducted in 1994-1995 to analyze objective video quality models. T1A1 was a subcommittee of the American National Standards Association (ANSI) accredited Alliance for Telecommunications Industry Solutions (ATIS). Subjective data was gathered at three laboratories using DSIS, four viewing sessions and Betacam-SP tapes. Like ITS1, none of the original SRCs were digitally compressed.

The T1A1 test included the null impairment for all 25 SRCs. Of these approximately 900 ratings, in 19 cases (2%) a subject rated the null impairment three or less (i.e., slightly annoying, annoying or very annoying). For DSIS there is a "true" answer (imperceptible), so these are two or more levels lower than the expected rating. These 19 cases were all associated with three SRCs. Of these, two are difficult to code. The quality of the third SRC is poor, as per the absolute category rating (ACR) scale. The average null impairment rating over all SRCs and subjects was 4.87 [6].

TABLE II
NULL IMPAIRMENT RATINGS FOR MUSHRA

Paper	[8]	[9]	[10] Narrowband	[10] Wideband
DMOS	$\approx$96	$\approx$97	99.47	99.44

TABLE III
ABSOLUTE DIFFERENCE BETWEEN T1A1 REPEATED RATING PAIRS

Difference	0	$\pm$1	$\pm$2	$\pm$3 or 4
Count	124	190	103	36
Percent	27%	42%	23%	8%

This behavior has been interpreted as a flaw in the DSIS method, such that it imposes a negative bias on the DMOS rating received by original. Some researchers have theorized that the bias results from the lack of a "quality improvement" rating level, which forces quality improvements to be rated as perceptible but not annoying (4). Note however that the T1A1 experiment did not include any intended quality improvement.

The DSIS null impairments scores may instead demonstrate the existence of subject imprecision, by which we mean imperfections in perception, memory and judgment. The subject may notice a flaw in the second viewing of the original that was not noticed upon the first viewing. As an example, one of the three T1A1 scenes that received low ratings is ANSI T1.801.01 standard test sequence cirkit, which zooms in on a circuit diagram depicting thin black lines on a white sheet of paper. These black lines occasionally appear to shimmer when viewed on a CRT.

Other double stimulus ratings of original stimuli can be found in audio quality subjective tests conducted with MUSHRA [7]. A computer interface presents the subject with multiple versions of the same audio clip, including the explicit reference (labeled), the hidden reference (unlabeled) and multiple impairments. The subject replays and compares stimuli before rating each on a continuous scale spanning [0..100]. The subject is instructed to rate the explicit reference as excellent (100). Although the hidden reference is always rated, papers do not always report the ratings for the hidden reference. Table II reports a few MUSHRA hidden reference MOS values that were clearly labeled as such. These values show the same trend that was observed with DSIS: the null impairment is occasionally rated slightly lower than expected.

The T1A1 test provides us with a limited number of direct observations of multiple ratings by a single subject for identical stimuli. Within each of the four sessions, the subject rated one stimulus twice for the purpose of examining viewer reliability. Subjects were divided into three pools, and each pool used a different set of stimuli for these viewer reliability checks. Table III shows the distribution of rating differences between these two scores for the same stimuli, aggregated over all 114 subjects and all 12 stimuli [6]. Three ratings were missing, and these statistics include subjects who were later removed.

Note that score changes of $\pm$2 or more are much more common in Table III than we saw for the prior null impairment

statistics. This may reflect a difference in the difficulty of the scoring task. The null impairment statistics reflect the simple task of comparing a stimulus to itself. For both DSIS and MUSHRA, the expected rating value is pegged at the top end of the quality scale. By contrast, the stimuli used to calculate Table III lay in the middle of the quality range (from 1.96 to 4.04, average 3.17) and contained digital coding artifacts. Moreover, 54% of the subjects had no prior experience with this technology [6].

The T1A1 subjective test relied upon repeated sequence scores and vision tests to screen subjects. An analysis of standard error indicated that reducing the number of subjects from 114 to 90 was more harmful than including the noisy data from the 24 screened subjects. Standard error was smaller for the 114 subjects, which is advantageous for all analyses that use averaged data (e.g., MOS). [6]. This suggests that subject scoring imprecision acts as a random variable within subject ratings.

An examination of 95% confidence intervals for MOSs shows greater or lesser agreement among subjects depending upon the stimuli (e.g., Fig. 2 of [11]) that cannot be explained by differences in MOSs. Kovács *et al.* [12] performed a subjective test that mimics a "tumbling E" eye chart, to determine the monitor's effective spatial resolution. The results shows large error bars for the worst quality level—larger than we would expect if the data were only driven by random chance. This indicates that even if we eliminate differences of opinion, subjects do not agree (e.g., due to differences in rating accuracy, visual system, and stimuli).

Cermak and Fay [6] analyzed the behavior of subjects in the T1A1 experiment and concluded that subjects center their scores around different fulcrum points. Cermak and Fay tried to explain this behavior using the subject ratings and questionnaire data gathered for the 114 subjects (i.e., gender, experience with video teleconferencing, age, visual acuity, color vision). Their conclusion was that demographic differences and the subject's choice of fulcrum point were not meaningful in predicting ratings of video quality. Their theory is supported by Pinson *et al.* [13], which serves as the starting point for the subject model presented in this paper.

Ostaszewska and Żebrowska-Łucyk [14] and van Dijk *et al.* [15] model subject ratings around the assumption that neither the subject's choice of fulcrum point nor the range of the subject's ratings are meaningful. In van Dijk *et al.* [15], the data is normalized without proposing a detailed subject model. Some of the conclusions that we reach in this paper support the normalization method proposed in [15]. We also include a deeper understanding of when it can be used. An interesting alternative was proposed by Ostaszewska and Żebrowska-Łucyk [14], where a subject model is proposed. That model focuses on errors generated by differences between subjects. We believe that such differences cannot always be removed. Individual opinion is a core factor in quality of experience (QoE) research.

Hossfeld, Schatz and Egger [16] examined the relationship between MOS and standard deviation of opinion scores (SOS). Their analysis of the discrete 5-level scale indicates a square relationship between MOS and SOS as follows:

$$SOS(x)^2 = a(-x^2 + 6x - 5) \quad (1)$$

Fig. 1. AGH/NTIA was run in a sound isolation chamber.

where x is MOS and a characterizes the dataset. Based on publicly available datasets, a was estimated to be 0.04 for image coding artifacts; 0.13 to 0.21 for video streaming; 0.27 to 0.34 for cloud gaming; and 0.27 to 0.59 for web surfing delay patterns. Variable a appears to measure the difficulty that subjects had scoring the stimuli within a particular dataset. The model proposed in our paper also contains a term describing stimulus difficulty, and as such is in line with the research presented in [16].

III. Datasets

We will use 18 different subjective experiments to analyze our subject model. Dataset AGH/NTIA provides us with training data, while the other 17 experiments are used to verify our results.

A. Subjective Video Quality Experiment AGH/NTIA

We designed the AGH/NTIA experiment to investigate the following three issues:

- the behavior of subjects;
- the impact of source video reuse on subjective data; and
- the suitability of subject screening methods.

All three investigations depend upon the availability of repeated scores for the same stimulus (e.g., subjects can be screened by repeated scoring of the same stimulus). However, the three design goals resulted in a complicated experiment design. A simplified experiment summary is presented here, and the full description can be found in [17].

Experiment AGH/NTIA is an ACR test conducted according to ITU-T Rec. P.910. This video-only experiment used 1080p 30fps content displayed on a laptop using a beta version of the web-enabled subjective test (WEST) software [18]. The WEST software runs subjective experiments from a local drive or over a network. It is able to randomize the provided sequences or display them in particular order, as was done for AGH/NTIA. The AGH/NTIA videos were played on a laptop with 17" screen (see Fig. 1). The WEST software recorded the ACR scores and ended each session with an automated questionnaire. Subjects rated video in three sessions.

The experiment included five hypothetical reference circuits (HRC). The HRCs were manually chosen to present five quality levels:

- original;
- good plus;

- good minus;
- poor plus; and
- poor minus.

The experiment included 94 SRCs. A full matrix of SRC by HRC was not used. Of the 470 possible processed video sequences (PVS), only 110 appeared in the experiment. Four of the SRCs were impaired with all five HRCs. The other 90 SRCs were only impaired with one HRC, such that there was an even balance of 18 SRC per HRC.

The AGH/NTIA experiment was designed around three viewing sessions. Each PVS was viewed and rated either once, three times, or six times. By rating three or six times, we mean that exactly the same 8 second long PVS was shown to a subject three or six times. Never was the same SRC shown one after the other, therefore the same PVSs shown in the same session were always separated by other PVSs. Subjects did not know that some PVSs were shown more than once. PVSs that were rated six times appeared twice in each of the three sessions; PVSs rated three times appeared once in each session; and PVSs rated once appeared either in session two or session three. Note that this paper uses three subsets of the AGH/NTIA experiment for different purposes: PVSs rated three times, PVSs rated six times, and PVSs rated either three or six times.

Twenty eight subjects participated in the experiment. Subjects were randomly assigned to one of two orderings ("red" and "blue"). Due to the nature of the experiment, session orderings were held constant (e.g., all people in ordering red had an identical viewing experience with regards to stimulus ordering). Twenty seven of the subjects were obtained from a temporary hiring agency. One subject was a visiting researcher, who was instructed to maintain maximum scoring consistency (i.e., try to repeat the prior rating for each stimulus).

Two subjects were intentionally given incorrect instructions, in an attempt to simulate two subjects who misunderstood the test instructions. These incorrect instructions asked the subject to jointly rate both the quality of the sequence and their opinion of the content. Only one of these two subjects appeared to be an outlier (i.e., ratings were obviously different than MOS computed with the other subjects' ratings). Those subjects should be removed from a typical experiment. We intended to add them to increase the variability of the subject's behavior in order to verify the proposed model. Nevertheless, the primary reason to run this experiment was to derive the model for a typical subject.

All subjects were retained for this analysis, because our focus was subject repeatability. For this analysis, we were interested only in the repeatability of subjects' ratings, measured through repeatedly rating the same PVS. We did not care whether or not a subject was using the subjective scale correctly, by some measure of correctness.

B. Other Datasets

We used 17 other subjective datasets to analyze our subject model. Our evaluation is focused on the 5-point scale. The possibility of using this model for different scales is left as future research. Therefore, all datasets presented within this paper used the discrete 5-point ACR scale from "Excellent" to "Bad." These 17 ACR datasets were divided into three sets.

First is a collection of six high definition television (HDTV) experiments conducted by the Video Quality Experts Group (VQEG) to validate HDTV objective quality metrics. These datasets are named vqegHD1, vqegHD2, vqegHD3, vqegHD4, vqegHD5, and vqegHD6. The individual subject ratings are included in the VQEG report [19]. Each subjective experiment was designed according to identical specifications, to contain a full matrix of 9 SRCs by 16 HRCs, plus a common set of 24 PVSs, for a total of 168 PVSs. The subjective data were collected using the ACR method.

Second, dataset vqegMM2 is an audiovisual subjective dataset that contains 60 PVSs. Subjective data was collected at six different labs in ten different environments, for a total of 213 subjects. The subjective data were collected using the ACR method. For a summary of the experiment and access to the subjective scores, see [20].

Third, dataset NTIA/Verizon [21] compares the performance of MPEG-2 and AVC/H.264 on HDTV, both coding only and in the presence of transmission errors. This experiment contains a partial matrix design, drawn from 12 SRCs and 9 HRCs, for a total of 144 PVSs. The subjective data was collected using the ACR method.

IV. Subject Scoring Model

The above cited work and our observations indicate that subjects do not provide stable and repeatable scores. In order to rigorously evaluate and be able to model such answers, we propose a subject scoring model. This model helps with not only predicting but also describing the uncertainty coming from the subjects.

A subjective quality test is a measurement of users' opinions. As with any measurement, subject opinion can be described by a model. We propose the model given by

$$o_{ijr} = \psi_j + \Delta_i + \epsilon_{ijr} \tag{2}$$

where:

- o_{ijr} is the observed rating for subject i, PVS j, and repetition r;
- ψ_j is the true quality of PVS j;
- Δ_i is the overall shift between the ith subject's scores and the true value (i.e., opinion bias); and
- ϵ_{ijr} is the error (i.e., scoring imprecision)

We assume that:

- there is a true value ψ_j, despite our inability to measure this value in absolute terms;
- the random variable Δ_i has zero mean when observed across all subjects;
- ϵ_{ijr} is influenced by both a subject's imprecision and the PVS scoring difficulty (e.g., PVSs with consistent quality throughout are easier to score consistently than PVSs with spatial or temporal quality changes); and
- random variable ϵ_{ijr} has zero mean both generally (over all subjects and PVSs and repetitions) and conditionally (for a particular subject or PVS)

Thus, if the ith subject is asked to rate the quality of PVS j many times, the obtained value should converge to $\psi_j + \Delta_i$. At this point we are not discussing how to estimate ψ_j. This true

value could be based on many ratings from many users, who could rate the sequence more than once. Equation (2) is the same as (1) from [22], except that the subscript r has been added.[1]

The model given by (2) assumes that there is a true value ψ_j and a subject bias Δ_i. In the case of quality measurement, this is not obvious. QoE research shows that a subject's opinion is influenced by many factors [23]. Moreover, humans in general are not precise in their answers and many theoretically irrelevant factors influence them [24]. Nevertheless, the most probable answer [25] or mean can be used as ψ_j, even if o_{ijr} is influenced by many factors. The important contribution of (2) is to divide scoring error into two different terms, ϵ_{ijr} and Δ_i, which can be drawn from the greater complexity of a subject's behavior.

Therefore, we will focus on major trends (measured systematically) and omit higher order variables (observed anecdotally). We know that (2) may be too simplistic. Nevertheless, the model can be useful only if it is simple enough that the model parameters can be estimated and detailed enough to provide useful information. With in this paper, we prove that our model fulfills the above criteria.

Our analysis of the proposed model will be conducted in three steps. First, the notation is described. Second, the rating differences are analyzed based on the discrete nature of the answers. Third, we assume that ψ_j and Δ_i are real numbers, and use continuous variable analysis to further investigate the proposed model.

A. Notation

The subject answer notation is a simplification of BT.500 notation wherein u_{ijkr} means a score for subject i, condition j, source k, and repetition r. In this paper, o_{ijr} marks a subject score, where j and k are combined to single letter j meaning PVS (i.e., specific condition for specific source).

We present two separate analyses: one assuming ψ_j, Δ_i, and ϵ_{ijr} are discrete values and the other assuming continuous values. The estimation of discrete values is denoted by m with an index denoting the estimated model parameter, for example m_{Δ_i}. Similarly estimation of the continuous values is denoted by μ with an index, for example μ_{Δ_i}.

An example of our estimate of ψ_j is the MOS (Mean Opinion Score) obtained for PVS j. This is denoted by μ_{ψ_j} and given by

$$\mu_{\psi_j} = \frac{1}{I_j R_j} \sum_{i=1}^{I_j} \sum_{r=1}^{R_j} o_{ijr} \tag{3}$$

where:

- μ_{ψ_j} denotes estimation from the data, assuming that the model with continuous ψ_j is considered;
- I_j is the total number of subjects for PVS j; and
- R_j is the total number of repetitions for PVS j.

Besides MOS, standard deviation is also used and denoted by σ and the variance is denoted by σ^2. For example, the standard deviation of all answers obtained for PVS j is denoted by s_j and given by

$$s_j = \underset{i,r}{\sigma}(o_{ijr}) = \sqrt{\frac{1}{I_j R_j - 1} \sum_{i=1}^{I_j} \sum_{r=1}^{R_j} \left(o_{ijr} - \mu_{\psi_j}\right)^2} \tag{4}$$

where:

- s denotes estimated standard deviation; and
- $\underset{i,r}{\sigma}$ denotes standard deviation computed over all I_j subjects and R_j repetitions.

In some places we aggregate specific values. Thus we introduce $\bigcup$ as a set aggregation. For example, a set of differences between all repeated scores and the first score given by the same subject is given by

$$F_i = \bigcup_j \bigcup_{r=2}^{R} \{o_{ijr} - o_{ij1}\} \tag{5}$$

where j under $\bigcup$ means that all possible j values are included. F_i (capital letter to indicate set) represents a set of values, and a single value can appear multiple times. Therefore, σ_{F_i} is the standard deviation of this set. To remove some elements of a set we use the set subtraction sign $\backslash$. For example, $C = A \backslash B$ means that set C contains elements of set A decreased by number of the same elements in set B: thus if set A contains five elements equal to one and set B contains three elements equal to one, then the obtained set C contains two elements equal to one.

V. Discrete Model Analysis

The first step of the analysis is based on discrete values. Such an assumption is correct from the formal mathematical point of view. The ACR scale is ordinal so o_{ij} are ordinal and in order to find the true quality we should use mode not mean [25]. Discrete analysis assumes that the distance between "Excellent" and "Good" is not necessary the same as the distance between "Good" and "Average." The mean operation assumes that those distances are the same and therefore it is incorrect to use mean in the case of ordinal data. In addition, for this analysis we also assume that ψ_j and Δ_i are discrete. In order to drive the results presented in this section, we need PVSs scored by the same subject more than once. Therefore, these analyses are based only on the AGH/NTIA experiment, using both three and six repeats.

We think that in the case of the discrete analysis, it is reasonable to choose ψ_j as the most probable answer. In this case Δ_i and ϵ_{ijr} distract from the answer that we observe the most often. For discrete analysis we use m_{ψ_j} to denote ψ_j

$$m_{\psi_j} = \underset{i,r}{\text{mode}}(o_{ijr}). \tag{6}$$

Similarly to the previous equation, in the case of discrete analysis, Δ_i is computed as the most probable difference between a subject score and m_{ψ_j}. For discrete analysis we use m_{Δ_i} to denote Δ_i

$$m_{\Delta_i} = \underset{j,r}{\text{mode}}(o_{ijr} - m_{\psi_j}). \tag{7}$$

[1] In [22] the model is named "theoretical user model." We changed the name to "theoretical subject model" to emphasize that we are modeling subjects, i.e., users asked to rate quality in an experiment not users judging a service in the real-life situation.

Using the notation defined in (6) and (7), different subject error distributions can be displayed. We start from an error distribution based only on the repeated scores and then we show an error distribution according to the subject answer model [see (2)].

The subject scoring of a PVS can be influenced by many factors. Nevertheless, if a subject scores the same PVS multiple times, we can find the most probable answer (let us denote it by m_{ij}). Thus m_{ij} is the intended answer of subject i for PVS j and is given by

$$m_{ij} = \underset{r}{\text{mode}}(o_{ijr}) \tag{8}$$

where $\underset{r}{\text{mode}}$ calculates a discrete mode. If two values are equally likely we select one at random. If three values are equally likely we select the central one.

Knowing m_{ij}, a difference between a subject answer and m_{ij} can be treated as an error. A set of all errors given by a single subject i is denoted by D_i and given by

$$D_i = \bigcup_{j,r} \{o_{ijr} - m_{ij}\} \setminus \bigcup_{j=1}^{J} \{0\}. \tag{9}$$

In (9), the term $\bigcup_{j=1}^{J}\{0\}$ is needed to remove all zeros that result from removing m_{ij}. Removing m_{ij} generates one zero for each PVS. This extra zero is a property of the algorithm that computes D_i (not of the subject accuracy). Therefore, the extra zero has to be removed.

The distribution of D_i for each subject shows how far they are scoring from their own most probable answer. Note that values of D_i different from zero do not come from a scoring model inaccuracy, since no scoring model is used to calculate D_i. The only assumption is that multiple scores given by the same subject for the same PVS should be the same.

Fig. 2(a) shows the observed distribution of errors D_i from (9) as a normalized histogram. In Fig. 2(a) the image intensity indicates the fraction of ratings in the bin, light yellow indicates an empty histogram bin, and the y-axis identifies the histogram bin ($o_{ijr} - m$). Subjects are sorted by the probability of the correct repetition (i.e., $o_{ijr} - m_{ij} = 0$), which ranges from 50% to 82%.

From Fig. 2(a), we see that subjects are not capable of perfectly repeating subjective scores, even when using the cognitively simple five-level ACR scale. All subjects occasionally changed their score by one category. Even the most repeatable subjects had 18% of scores different than the most probable score given by the same subject for the same PVS. An interesting observation is that the two most repeatable subjects, in terms of the greatest probability of repeating m_{ij}, had D_i values of two, which is half of the scale.

Computing D_i is difficult since a relevant m_{ij} value calls for many repetitions and the estimated value m_{ij} can be imprecise. On the other hand, in a typical experiment only one score per PVS per subject is collected. We next address the case where some repeats exist, but not enough to calculate m_{ij}. In this case an interesting question is: how does a score change, assuming

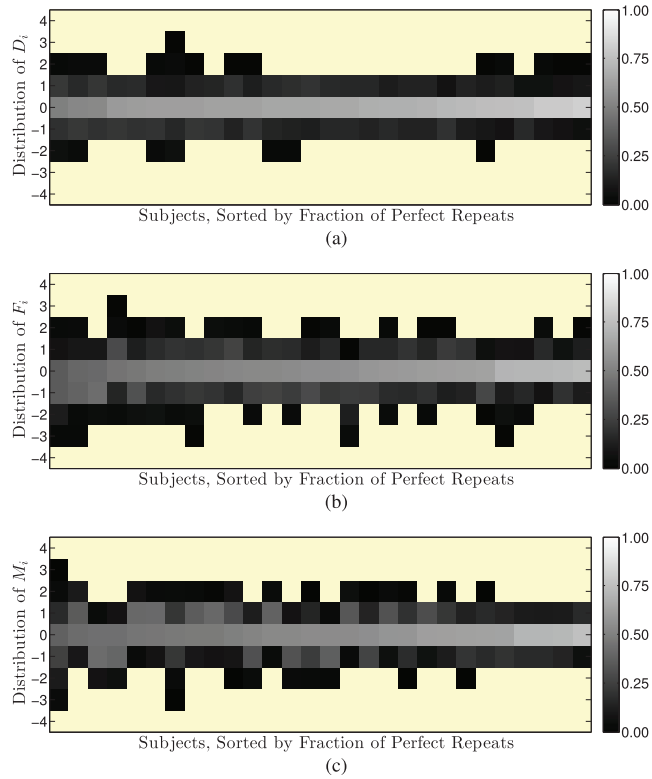

Fig. 2. Error distribution for each subject sorted by probability of correct answer. (a) Compared to mode. (b) Compared to the first score. (c) Compared to model.

TABLE IV
DIFFERENCE BETWEEN AGH/NTIA REPEATED RATINGS

Difference	0	±1	±2	±3	±4
Count	2066	1361	188	24	1
Percent	57%	37%	5%	1%	0%

that the first score given by a subject i for PVS j is the correct one. The set of such answers is denoted by F_i and is given by

$$F_i = \bigcup_{j} \bigcup_{r=2}^{R} \{o_{ijr} - o_{ij1}\}. \tag{10}$$

The F_i distribution for different subjects is shown in Fig. 2(b), and the probability of correct repetition ranges from 36% to 74%. Comparing Figs. 2(a) and 2(b) we see lower values of correct answers and stronger scattering. This is an obvious consequence of choosing the first instead of the most common answer. Table IV shows the overall distribution of F_i for all subjects.

Analyzing Fig. 2(b), we can be sure that if a PVS is repeated it has a high probability of being scored differently. Moreover the difference by ±2 does not indicate that a subject is not relevant since even those with the highest repeatability made such an error. We also cannot see any clear pattern in the change of the obtained distribution. One could expect that the order of repeatability will also determine the order of the maximum F_i values. To the contrary, the fifth most repeatable subject has at least one value of $F_i = -3$ and the third least repeatable subject

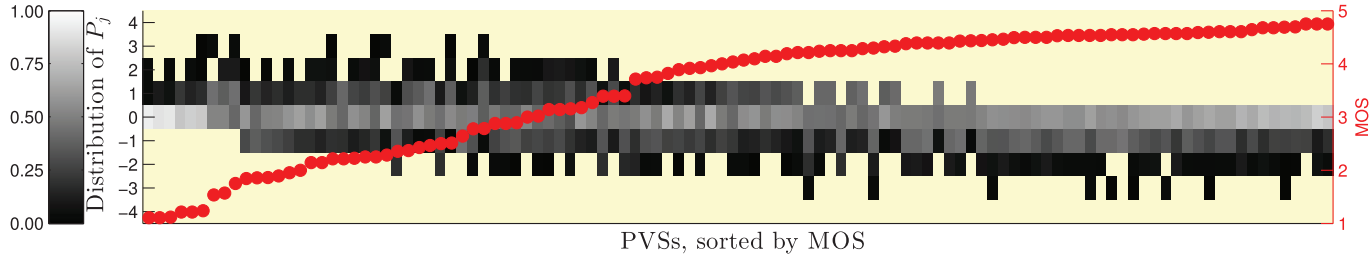

Fig. 3. Observed distribution of rating changes (r_j) for each PVS, sorted by MOS (red dots).

has only very small number of $F_i = -2$ and no other values $|F_i| > 1$.

The distribution of F_i needs repeated scores. In a typical experiment the scores are not repeated and therefore an error distribution cannot be obtained. Nevertheless, thanks to the subject model such a distribution can be estimated.

The distribution shown in Fig. 2(c) is based on the proposed model given by (2). In this case a set of errors is denoted by M_i and given by

$$M_i = \bigcup_{j,r} \{o_{ijr} - m_{\psi_j} - m_{\Delta_i}\} \tag{11}$$

where m_{ψ_j} and m_{Δ_i} are given by (6) and (7), respectively. In Fig. 2(c), the probability of correct repetition ranges from 38% to 76%.

In Fig. 2(c) the M_i distribution is shown. Comparing M_i and F_i distributions we can see that they are similar from the point of view of the range. The obtained distributions are statistically different, which is to be expected as M_i includes between subject–sequence interactions while F_i does not. Regardless, this shows that the proposed model covers subject variability well. Therefore, (11) is recommended to be used to estimate error distribution in different experiments.

Scattering of the results obtained for each subject shows that individual subjects cannot repeat answers precisely. We also know that individual subjects do not agree on answers. One of the questions is how much PVSs influence this lack of repeatability and agreement. In order to answer this question of variations both within and between subjects, we computed distribution of differences between the most probable answer for a particular PVS (i.e., m_{ψ_j}) and the actual answer. Such set is denoted by P_j and given by

$$P_j = \bigcup_{i,r} \{o_{ijr} - m_{\psi_j}\}. \tag{12}$$

Fig. 3 shows the observed distribution of P_j from (12) as a normalized histogram, with PVSs sorted by MOS. The image intensity indicates the fraction of ratings in the bin, light yellow indicates an empty histogram bin, and the left side y-axis identifies the histogram bin ($o_{ijr} - m_{\psi_j}$). MOS is displayed on the right side y-axis and plotted as red dots.

From this figure, we see a spread of ratings for each PVS that typically spans three of the five levels. In no case do all responses fall on a single rating, and in some cases the ratings span the entire scale. We see more PVSs with the higher MOS since original sequences are analyzed as well.

The greatest repeatability (within and between subjects) is 93% in case of very bad quality. The best quality PVSs cannot obtain such vote uniformity. This shows that there is greater agreement (within and between subjects) for bad quality than for good quality. The worst repeatability is 31% for sequence `bluespruce2a_poorplus`; further investigation of this sequence shows that it cuts between content with different qualities and thus is difficult to score consistently. We will come back to this problem later in the case of continuous analysis.

Note that P_j includes the influence of both Δ_i and ϵ_{ijr} on the observed scores' unrepeatability. From Fig. 3 it cannot be deduced if the observed spreading of scores is caused by differences among subjects (Δ_i) or by errors they make (ϵ_{ijr}). Such analysis for discrete ψ_j and Δ_i requires a large number of repetitions and subjects. Therefore, in the next section we present an analysis based on the assumption that ψ_j and Δ_i are continuous values.

VI. Continuous Model Analysis

We now choose to assume that ψ_j and Δ_i are continuous, so their estimations must be redefined compared to (6) and (7). The continuous subject model is different from the discrete subject model. A continuous model will reveal different characteristic of the subject scoring process. For the continuous analysis we assume that $\Delta_i + \epsilon_{ijr}$ is a deviation from a true value ψ_j. We assume that such deviations have a mean value of zero [see (3)]. Therefore, ψ_j in the case of the continuous analysis is the mean value denoted by μ_{ψ_j} and is given by (3).

Let us define the observed bias of each subject with respect to μ_{ψ_j}. It is denoted by μ_{Δ_i} and given by

$$\mu_{\Delta_i} = \frac{1}{\sum_{j=1}^{J_i} R_j} \sum_{j=1}^{J_i} \sum_{r=1}^{R_j} \left(o_{ijr} - \mu_{\psi_j}\right). \tag{13}$$

In other words, μ_{Δ_i} estimates Δ_i by taking the average difference between the ith subject's ratings and all subjects' ratings. Equation (13) follows from substituting μ_{ψ_j} for o_{ijr} in our subject model (2), as shown in Janowski and Pinson [22].

We will now present a series of analyses that indicate the correctness of our proposed model, using continuous analysis. Taken together, these analyses will prove the utility of our subject model from (2).

A. Analysis of Δ_i

We begin by demonstrating that each subject has a bias and that the bias is stable.

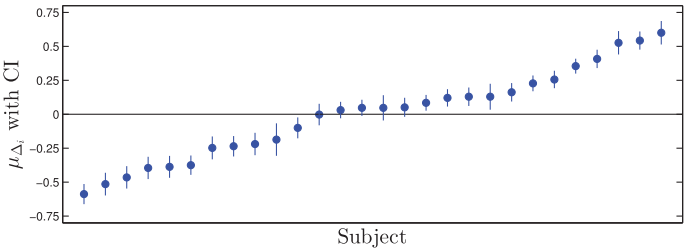

Fig. 4. Observed distribution of subject bias (μ_{Δ_i}) with 95% confidence intervals, for dataset AGH/NTIA.

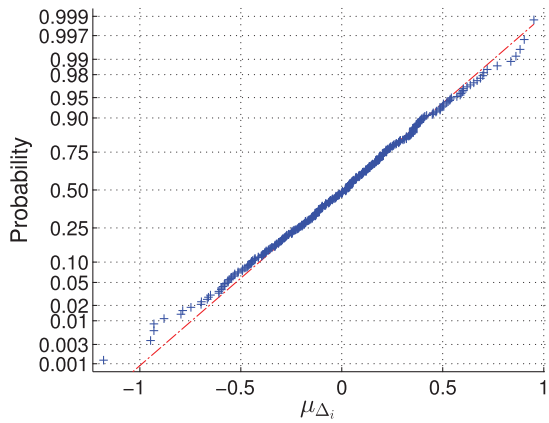

Fig. 5. Normal probability plot of μ_{Δ_i} indicates an approximately normal distribution.

Fig. 4 shows μ_{Δ_i} for dataset AGH/NTIA, with subjects sorted by μ_{Δ_i}. The small CIs[2] indicate that μ_{Δ_i} is stable.

Fig. 5 shows the normal probability plot of μ_{Δ_i} for all nine datasets mentioned in Section III. Variable μ_{Δ_i} has mean zero and standard deviation 0.34. From this, we can see that μ_{Δ_i} has a normal distribution and 85% of the calculated values of μ_{Δ} fall into the range ± 0.5 (which is one-quarter of the full scale).

We have shown that μ_{Δ_i} has a small confidence interval. This demonstrates that μ_{Δ_i} is stable, and we can reasonably treat this as a variable that depends upon i. It appears to be reasonable to ignore the underlying complexities that we know occur (e.g., compression near the end of the ACR scale).

B. Analysis of ϵ_{ijr}

We now examine the error term, ϵ_{ijr}. We understand error to mean the standard deviation of the error. In Section V, we saw evidence that the error term is a function of both subject and PVS. We now seek a model that incorporates both factors. Let α_i be an error parameter for the ith subject and let β_j be an error parameter for the jth PVS. The two simplest possible models for ϵ_{ijr} are as follows.

1) *Multiplicative model*: ϵ_{ijr} is a random variable with mean zero and standard deviation $\alpha_i\beta_j$, i.e.

$$\epsilon_{ijr} = \alpha_i \beta_j X \tag{14}$$

where $X \sim N(0,1)$.

[2]CI is CI for mean with assumption that the sample follows normal distribution.

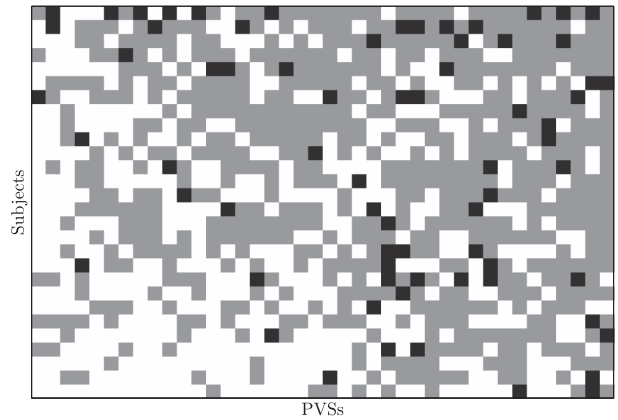

Fig. 6. Scoring difficulty measured by number of rating levels used over three repetitions: white = 1 level, gray = 2 levels, and black = 3 levels. Subjects and PVSs are sorted by the average number of levels used.

2) *Additive model*: ϵ_{ijr} is a sum of two random variables with standard deviations α_i and β_j, i.e.

$$\epsilon_{ijr} = \alpha_i X + \beta_j Y \tag{15}$$

where $X, Y \sim N(0,1)$, and X and Y are independent.

The philosophy behind each model is different. In the case of the multiplicative model we say that if a PVS is very easy (small β_j) the observed variance, even for an inaccurate subject (high α_i), is small. Also the reverse situation is true. So if a subject is very stable (small α_i), even for a very difficult PVS (high β_i) the obtained variance is relatively small. On the other hand, for the additive model we say that PVS and subject error level are independent. So for a difficult PVS even a perfect subject will make significant errors because there is a PVS error level which cannot be decreased.

With the data we have it is not possible to decide whether (15) or (14) is more correct. Nevertheless, we think that the additive model is more appropriate. Evidence is shown in Fig. 6, which is a visual representation of the rating accuracy for each subject and each PVS, limited to PVSs that were rated exactly three times (i.e., once per session). Fig. 6 indicates that an additive error model is more likely than a multiplicative model. We see that the most difficult PVSs are not much more repeatably scored by the most repeatable subjects than by the least repeatable subjects.

Equation (15) may be too simplistic. It is possible that the subject model should contain both terms, (15) and (14), along with other variables not yet considered. However, we want model parameters that can be estimated and the theoretical equations are much easier if we choose (15). Therefore, we will move forward with the additive model given by

$$o_{ijr} = \psi_j + \Delta_i + \alpha_i X + \beta_j Y \tag{16}$$

where $X, Y \sim N(0,1)$, and X and Y are independent.

The first step of the model analysis is α_i and β_j estimation. Since the standard deviation of ϵ_{ijr} in the case of additive model is $\sqrt{\alpha_i^2 + \beta_j^2}$, estimation of α_i and β_j is based on the variance of ϵ_{ijr}, not standard deviation.

1) Estimation in the Case of Multiple Answers Per PVS: We start with PVSs that were scored numerous times by each subject. In this case, a standard deviation of the jth PVS scored by the ith subject can be computed. Let us denote it by s_{ij}, which is an estimation of the true standard deviation. We propose to minimize the squared difference between the observed variance and the model variance, as given by

$$f(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(s_{ij}^2 - \alpha_i^2 - \beta_j^2\right)^2 \tag{17}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of all α_i and β_j respectively.

The above equation is a function of α_i and β_j, since s_{ij}^2 are obtained from the data and are known. We are looking for the minimum of (17). Note that all α_i and β_j are standard deviations so they cannot be negative. So the minimization problem is

$$\begin{cases} \min_{\alpha_i,\beta_j} f(\boldsymbol{\alpha},\boldsymbol{\beta}) \\ \forall_{i,j}\alpha_i \geq 0, \beta_j \geq 0 \end{cases}. \tag{18}$$

Equation (18) is not the only possible way to estimate α_i and β_j parameters. As a future research topic, we will analyze different options and choose the one that gives the best estimator.

All candidates for the minimum of (18) can be obtained by setting the derivative of (18) with respect to each α_i and β_j to zero. The obtained derivative has the form

$$\frac{\partial f(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \alpha_k} = \sum_{j=1}^{J} 2\left(s_{kj}^2 - \alpha_k^2 - \beta_j^2\right)(-2\alpha_k) \tag{19}$$

and a similar expression is obtained for the derivative with respect to β_k.

The obtained equation has two possible solutions

$$\alpha_k = 0 \tag{20}$$

or

$$\sum_{j=1}^{J}\left(s_{kj}^2 - \alpha_k^2 - \beta_j^2\right) = 0. \tag{21}$$

In the case of derivative over β_k we have

$$\beta_k = 0 \tag{22}$$

or

$$\sum_{i=1}^{I}\left(s_{ik}^2 - \alpha_i^2 - \beta_k^2\right) = 0. \tag{23}$$

The above equations generated many alternative sets of equations that are different by the choice of the indexes of α_i and β_k, which are zeros. From those equations the optimal solution can be found. Since such a procedure would be complicated we proposed a simpler solution. Nevertheless, we need a proof that the above proposed solution is optimal. This is left for future research.

Transforming (21) and (23) we obtain

$$\alpha_k^2 = \frac{1}{J}\sum_{j=1}^{J} s_{kj}^2 - \frac{1}{J}\sum_{j=1}^{J}\beta_j^2 \tag{24}$$

$$\beta_k^2 = \frac{1}{I}\sum_{i=1}^{I} s_{ik}^2 - \frac{1}{I}\sum_{i=1}^{I}\alpha_i^2. \tag{25}$$

The above equations have similar means over certain variances. For simplicity we introduce notation

$$a_k = \frac{1}{J}\sum_{j=1}^{J} s_{kj}^2 \tag{26}$$

and

$$b_k = \frac{1}{I}\sum_{j=1}^{I} s_{kj}^2 \tag{27}$$

which simplifies (24) and (25) to:

$$\alpha_k^2 = a_k - \frac{1}{J}\sum_{j=1}^{J}\beta_j^2 \tag{28}$$

$$\beta_k^2 = b_k - \frac{1}{I}\sum_{i=1}^{I}\alpha_i^2. \tag{29}$$

Additionally we assume that all a_i and b_i are in increasing order. Such an assumption does not change anything, since the numbers that link with a particular subject or PVS are not important from the computation point of view.

Let us assume that all α_i and β_j are greater than zero. In this case (24) and (25) are used to estimate α_i and β_j. Since we have as many equations as variables and one of those equations is a linear combination of the other equations, there is no unique solution. Instead we can find an infinite number of solutions, which are function of a single free parameter. In order to find a solution, we say that α_1 (the smallest α) or β_1 equals x. Let us assume that the smallest value is $\alpha_1 = x$ so based on (24) we can say that

$$\frac{1}{J}\sum_{j=1}^{J}\beta_j^2 = a_1 - x^2. \tag{30}$$

Note that the order of a_i determines the order of α_i. Therefore, the smallest value was used to estimate $\frac{1}{J}\sum_{j=1}^{J}\beta_j^2$.

The final equations for α_i^2 and β_j^2, which depend on x, are

$$\alpha_k^2 = a_k - a_1 + x^2 \tag{31}$$

and

$$\beta_k^2 = b_k - \frac{1}{I}\sum_{i=1}^{I} a_i + a_1 - x^2. \tag{32}$$

If all obtained results are positive, we are looking for a unique solution. A unique solution is obtained by choosing a particular value of x. The value of x can be interpreted as trade between difficult PVSs or inaccurate subjects. Since we cannot decide which situation occurred a simple solution where $x = 0$ is used.

However, (32) can result in negative values. In this case we conclude based on (20) that $\alpha_1 = 0$. For $i = 1$, (25) is not valid anymore since we used alternative equations. Therefore, the set of equations (24) and (32) have new forms. Equation (24) is valid for $k > 1$ and (32) has the form

$$\beta_k^2 = b_k - \frac{1}{I}\sum_{i=2}^{I}\alpha_i^2 \tag{33}$$

where the sum starts from 2, not 1 as it did previously. In this case the set of equations are not linearly related and can be solved. Again some solutions can be negative. In this case the smallest value should be changed to zero and the next set of equations should be solved. The set of positive solutions is found iteratively. We used actual values of s_{ij} from the AGH/NTIA subjective test results to verify the new algorithm described in (24) through (33). Specifically, we applied a commercially available optimization problem solver to (17) and found that it produced the same values of α_i and β_j as the new algorithm in every case. A formal proof of optimality is outside the scope of this paper.

2) Estimation in the Case of a Single Answer Per PVS: Equations (31) and (32) can be computed only if numerous answers per PVS are given by the same subject. Such data are expensive for the researcher and boring for the subject. In a typical subjective experiment a single PVS is scored once by a subject. We want equations for α_k^2 and β_k^2 that require only one answer per PVS per subject, i.e., o_{ijr} without repetitions on o_{ij}. Using the proposed model it is possible.

We cannot derive the equation based on the variance estimation s_{ij}^2 since only one answer per subject i and PVS j is known. Nevertheless, an error given by single subject can be measured assuming that μ_{ϕ_j} and μ_{Δ_i} are known by the equation

$$r_{ij} = o_{ij} - \mu_{\psi_j} - \mu_{\Delta_i} \tag{34}$$

where r_{ij} is a residual of the answer after removing the PVS and subject influence.

According to the additive model, $r_{ij} = \alpha_i X + \beta_j Y$, both X and Y are independent normally distributed variables with mean zero and standard deviation 1. Collecting answers from different subjects and the same PVS will give us a set of random variables from different distributions for which variances are $\alpha_i^2 + \beta_j^2$. Since this result is not obvious it is proved in Appendix B.

Computing the variance of r_{ij} for a fixed $j = k$ is the same as computing the variance of a collection of random variables that have variances $\alpha_i^2 + \beta_k^2$. Since each subject gives one answer $p_1 = p_2 = \cdots = p_I = \frac{1}{I}$. So the obtained variance is $\beta_k^2 + \frac{1}{I}\sum_{i=1}^{I} \alpha_i^2$. This variance can be computed from the data since we have I values of r_{ik}. The variance computed from the data is denoted by $s_{\cdot k}^2$, hence

$$s_{\cdot k}^2 = \beta_k^2 + \frac{1}{I}\sum_{i=1}^{I} \alpha_i^2. \tag{35}$$

Exactly the same reasoning can be made for computing a variance obtained for all PVSs scored by a single subject. Then we obtain

$$s_{k\cdot}^2 = \alpha_k^2 + \frac{1}{J}\sum_{j=1}^{J} \beta_j^2. \tag{36}$$

Since $s_{\cdot j}^2$ and $s_{i\cdot}^2$ are known and we are estimating α_i^2 and β_j^2, the latter have to be derived from (35) and (36). Note that if $s_{i\cdot}^2$ is substituted for a_i and $s_{\cdot j}^2$ for b_j, then the above equation is identical to (28) and (29), the equations for multiple answers. Therefore, the solving algorithm is exactly the same.

C. Estimating Model Parameters

Subsections VI-A and VI-B derive equations that estimate the model parameters μ_{Δ_j}, α_i and β_j.

The precision of the estimation procedure has to be evaluated as well, especially since μ_{Δ_i} is based on μ_{ψ_j}, α_i on both μ_{Δ_i} and μ_{ψ_j}, and finally β_j on all other parameters. Such cascades generate larger and larger errors.

In order to evaluate the estimation precision, a simulation was run. We generated subjects' answers for different ψ_j, Δ_i, α_i and β_j as specified by (16). For each simulation ψ_j spanned from 1.1 to 4.9, β_j from 0.03 to 0.6, Δ_i from -0.6 to 0.6, and α_i from 0.03 to 0.7. Each PVS was described by ψ_j and β_j. For example, if only four of the PVSs were used in the simulation scenario then $(\psi_j, \beta_j) = (1.1, 0.03)$, (1.1, 0.6), (4.9, 0.03), and (4.9, 0.6). If more PVSs were used, the other PVSs were spaced linearly within the area defined by these four points. Similarly, each subject was described by Δ_i and α_i. The same procedure was used to create a set of (Δ_i, α_i) pairs that evenly spanned the available area. This forms a set of values used as an imput to the simulator. The same values should be estimated in the estimation process.

The simulated scores are calculated as

$$o_{ijr} = \|\psi_j + \Delta_i + \alpha_i \xi_1 + \beta_j \xi_2\| \tag{37}$$

where each instance of ξ_1 and ξ_2 is a pseudorandom number drawn from the standard normal distribution, and function $\|x\|$ rounds the value x to the nearest integer within the range [1], [5]. From these simulated scores, we can estimate μ_{ψ_j}, μ_{Δ_j}, α_i, and β_j using (3), (13), (31), and (32). We measure the accuracy of these estimates by computing R-squared (R^2) for the simple linear fit between the true values and the obtained estimates. We are using a linear fit since we are more interested in the relation between the parameters than the exact value of particular parameter.

This simulation and estimation procedure was repeated for different numbers of PVSs and different numbers of subjects. Because we chose (ψ_j, β_j) and (Δ_i, α_i) to evenly span two dimensions, the number of simulated PVSs and the number of simulated subjects must be squares of natural numbers starting from nine (i.e., 9, 16, 25, 36, … 225). To reduce the impact of specific ξ_1 and ξ_2 values obtained by chance, each simulation was run 30 times and the R^2 results averaged.

Fig. 7 uses a color map to indicate the R^2 values produced by simulation runs in the case that the same subject rates the same PVS six times. Fig. 7 shows that both ψ_j and Δ_i can be estimated precisely. The worst case is the rare case of both many subjects and a low number of PVSs but even then R^2 is 0.85. The estimation of α_i and β_j is not so good. Moreover their precision calls for contradicting conditions—i.e., a large number of PVSs and a low number of subjects for α_i estimation and a large number of subjects and low number of PVSs for β_j estimation. This is good news if we are interested only in one of those two parameters.

If we need a good estimation of all parameters the precision can be described by minimum R^2, as shown in Fig. 8. Based on Fig. 8 we can see that reasonable precision can be obtained

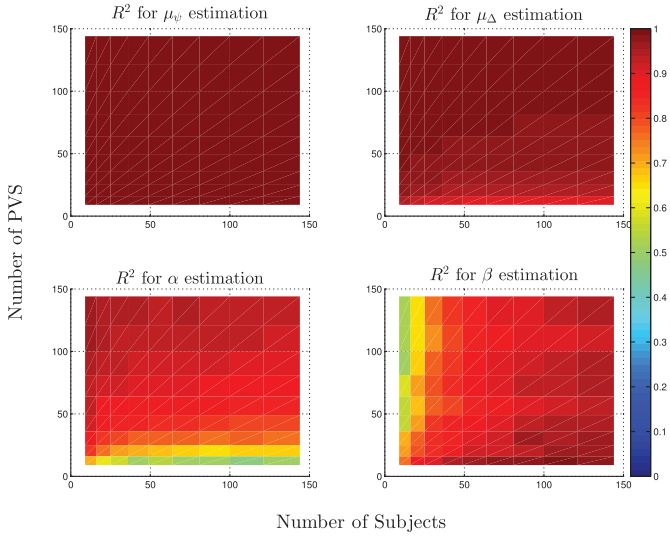

Fig. 7. Precision of estimation of a particular model parameter calculated as R^2 of a linear fit between the true values and the obtained values, for $r = 6$.

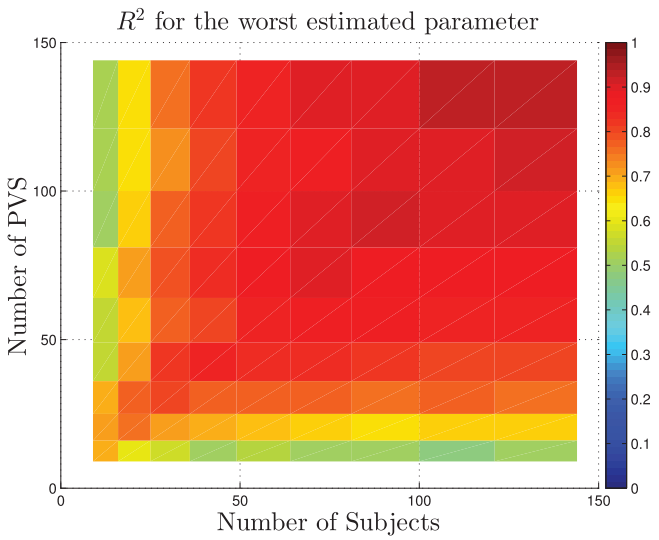

Fig. 8. Precision of estimation of the model parameter for which R^2 of a linear fit between the true values and the obtained values is the lowest, for $r = 6$.

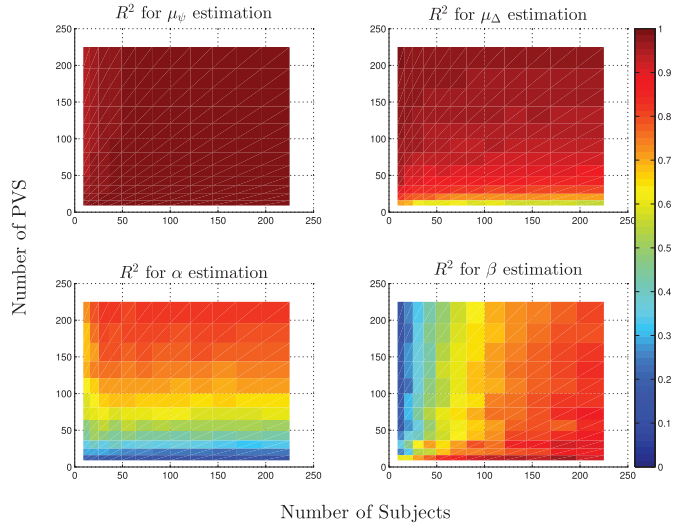

Fig. 9. Precision of estimation of a particular model parameter calculated as R^2 of a linear fit between the true values and the obtained values, for $r = 1$.

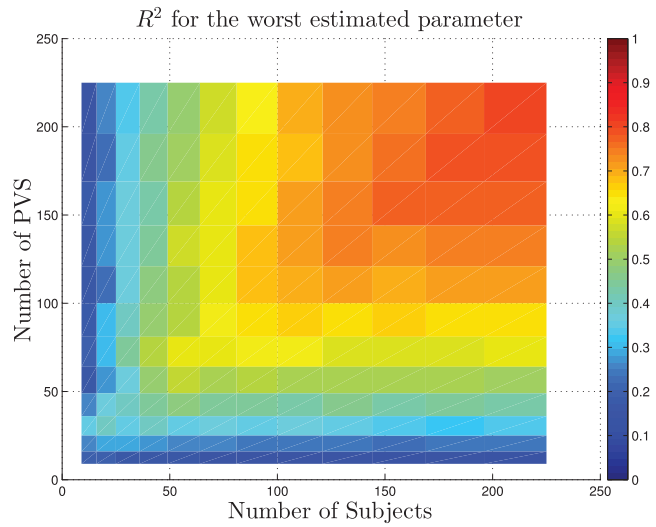

Fig. 10. Precision of estimation of the model parameter for which R^2 of a linear fit between the true values and the obtained values is the lowest, for $r = 1$.

for 25 subjects and 25 PVSs. These 25 PVSs rated six times by each subject gives 180 sequences to watch, which is possible.

A similar analysis was made in the case of single answer estimation (i.e., each subject rates each PVS once.) The obtained results are shown in Figs. 9 and 10. As expected, the obtained precision is worse than in the case of multiple answers per PVS. To precisely estimate all model parameters we need around 180 PVSs, which is a typical number for a subjective experiment, but also around 180 subjects, which is difficult to obtain.

The AGH/NTIA experiment contains 10 PVSs with 6 repetitions scored by 28 subjects. For those sequences, we have two different ways to calculate the error generated by a subject and the error generated by a PVS. The first is discrete analysis, as shown in Section V. We will use the repeatability of the first score as shown in Fig. 2(b). The second way to calculate those errors is using the continuous model and estimating α_i and β_j. We expect that both metrics will correlate strongly, since they both show the same phenomena.

The scatter plots obtained are shown in Figs. 11(a) and 11(b). As we can see, in the case of subject analysis the correlation is weaker. This is a natural consequence of limits in the α_i estimation for a small number of PVSs. As shown in Fig. 7, the expected accuracy measured by R^2 is around 0.5. A much better result is obtained for the PVS analysis. Again, this is in line with the results obtained at the beginning of this section. This comparison between the discrete and continuous analyses shows the correctness of the proposed model.

The overall precision of the proposed estimation algorithm is not sufficient for many real world applications. Calling for 180 subjects is especially unrealistic, and also it would be better if the estimated values are very close to the original by value not only by order. Therefore, further studies are needed on better estimation algorithms to replace (17) through (36).

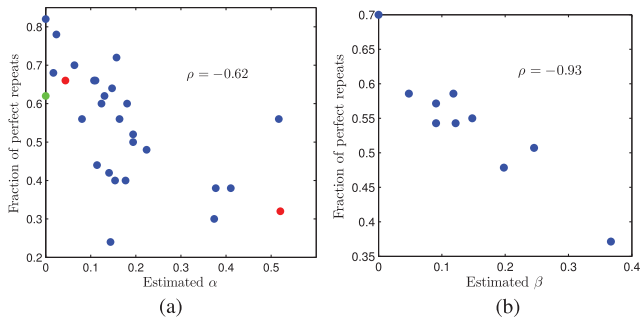

Fig. 11. These two scatter plots compare the discrete analysis estimates (y axis) and continuous analysis estimates (x axis). The Pearson correlation between x and y is identified on each plot. Red dots indicates subjects with incorrect instruction and green dots indicate an expert. (a) Subject error. (b) PVS error.

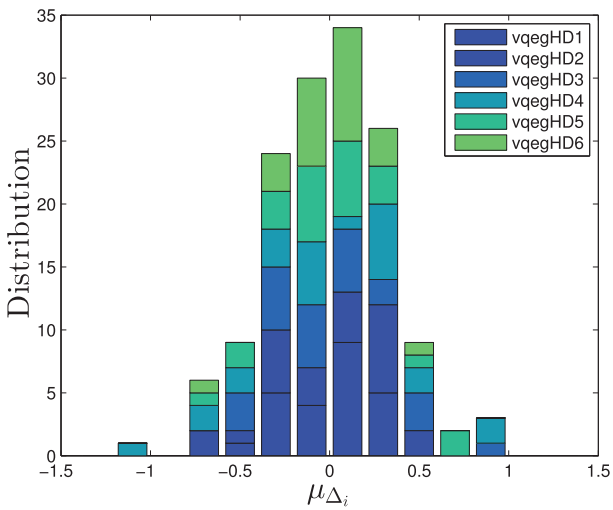

Fig. 12. Distribution of Δ_i for the six VQEG HDTV datasets.

VII. Model Correctness by Deduction

We would like to measure our subject model's ability to explain subjective data. More precisely, we will now show that the model explains certain behaviors observed in real subjective data.

A. Model Correctness by Normalizing Subjective Data

Our analysis of Δ_i in Section VI-A indicates that we should be able to remove the influence of Δ_i from any subjective dataset. This normalization will not impact μ_{ψ_j} and should decrease our estimate for σ_j from s_j (see (4)) to $\hat{s}_j$, given by

$$\hat{s}_j = \underset{i,r}{\sigma}(o_{ijr} - \Delta_i). \tag{38}$$

We expect to see

$$\hat{s}_j < s_j \tag{39}$$

while μ_{ψ_j} remains constant. The consequences should be the ability to differentiate between more pairs of PVSs.

To test this theory, let us examine the six datasets from the VQEG validation of HDTV video quality objective metrics. Fig. 12 shows the distribution of Δ_i values for all datasets. This distribution shows the normal distribution seen in Fig. 5.

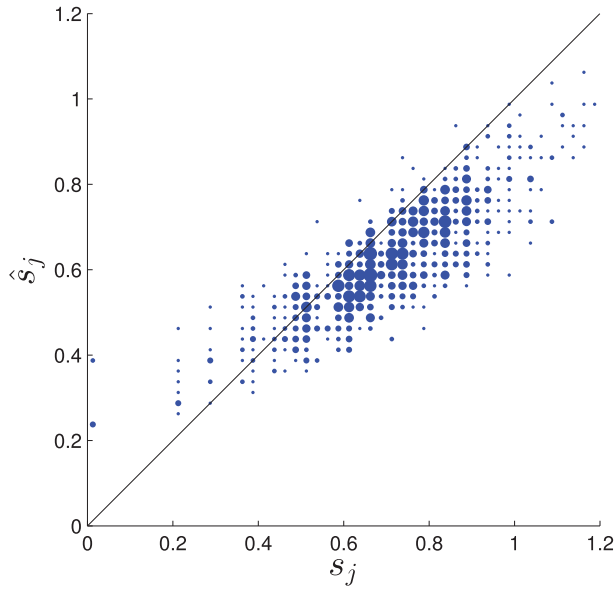

Fig. 13. Shift in standard deviation. Impact of removing Δ_i from the six VQEG HDTV datasets. Larger dots indicate more data.

Fig. 13 shows the difference between s_j and $\hat{s}_j$. The dot size indicates the data density. While $\hat{s}_j$ is generally less than s_j, $\hat{s}_j$ is larger for small values of s_j.

When Δ_i is removed, σ_j decreases by 0.035 to 0.167 on average, depending on the dataset. Removing μ_{Δ_i} usually moves the minimum and maximum σ_j value toward the median, thus eliminating extreme values. This means that for the minimum variance, (39) is not satisfied. We believe that the small values of s_j are caused by clipping of the MOS scale and not an imperfection of the model given in (2).

For all PVS pairs within each of the six VQEG HDTV datasets, a Student's t-test was used to calculate whether or not the PVSs had equivalent MOS at the 95% significance level. This calculation was repeated on the normalized scores (with μ_{Δ_i} removed). Based on the combined results from all six datasets, removing μ_{Δ_i} had the following impact:

- 97.22% no change;
- 2.69% increase in sensitivity (equivalent → different);
- 0.08% decrease in sensitivity (different → equivalent);
- 0% inversions (opposite conclusions; inversions are impossible).

Overall, normalizing each subject's scores by Δ_i seems to improve the ability of these datasets to distinguish between PVS and MOS in a meaningful way. The impact may be more pronounced than is indicated here, because subjective experiments often compare stimuli with similar quality. More examples can be found in Janowski and Pinson [22]. This supports the correctness of our subject model and provides corroborating evidence for the term Δ_i.

B. Model Correctness by Subject Reuse (α)

The traditional experiment design uses a large number of subjects to rate each PVS. However, our final subject model, (16), claims that all subjects agree on quality comparisons, such as stimulus A is better quality than stimulus B. No matter how subjects are divided into subsets, we do not expect to find score

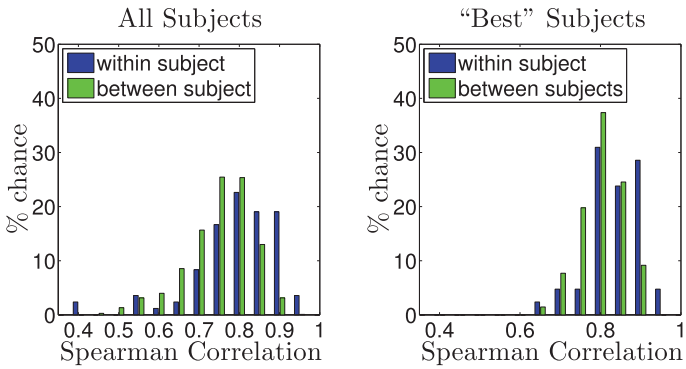

Fig. 14. Voting differences from one subject's repeated votes (blue) compared with the voting differences from one subject to another (light green). The left histogram shows all subjects, and the right histogram shows subjects with a high correlation to MOS.

inversions. Empirical data from Pinson *et al.* [13] supports this hypthesis. Comparisons between five sets of 25 to 34 subjects calculated the probability of inversion at less than 0.03%.

If this is true, then a viable alternative would be to use fewer subjects and have each subject rate each PVS multiple times.

Fig. 14 displays two distributions of Spearman correlations. The dark blue histogram shows correlation within one subject's ratings, between two different sessions. The light green histogram shows correlation between two different subjects for the same session. The left subplot uses data from all subjects; the right subplot uses the 50% of subjects with the highest Pearson correlation to MOS. We see that the within-subject correlations are only slightly higher than the between-subject correlations.[3]

Let us consider only PVSs from the AGH/NTIA dataset with three or six ratings per subject. We randomly chose two disparate subsets:

(A) 18 subjects, retaining only the 1st rating;
(B) 6 subjects, retaining three ratings (the first of each session).

Thus, the two subsets contain an equal number of ratings per sequence, and the subsets do not overlap. We calculated MOS over these two subsets, and compare their MOS to μ, the MOS computed using all ratings from all subjects. This process was repeated 1000 times.

The R^2 between subset (A) and μ is on average 0.984, and at a minimum 0.967. The R^2 between subset (B) and μ is on average 0.966, and at a minimum 0.933. The average R^2 value is quite high, because all subjects in each subset are also used to calculate μ. The minima are strongly impacted by the behavior of aberrant subjects, who were not screened.

The above two subsets are not quite equivalent. This is to be expected, because subset (B) partially ignores the impact of Δ_i on the subject pool—subset (B) does not contain a sufficient variety of Δ_i. If we replace subset (B) with subset (C) and increase the size as follows:

(A) 18 subjects, retaining only the 1st rating;
(B) 7 subjects, retaining three ratings (the first of each session)

[3]Spearman correlation is used, as it is intended for ordinal data. These same histograms can be computed with Pearson correlation, but the trend is identical.

then R^2 between subset (C) and μ is on average 0.983, with at a minimum 0.953. This is similar to the performance of subset (A).

The similar behavior of subsets (A) and (C) supports the theory that that all subjects agree on the relative rankings of stimuli. This provides supporting evidence for our subject behavior model—as opposed, for example, to a hypothetical subject behavior model that allows for opinion motivated score inversions. The similar behavior of subsets (A) and (C) has several practical implications, which we will consider later.

VIII. Conclusion

We have shown that subjective ratings are influenced by subject bias (Δ_i), subject inaccuracy (α_i) and PVS inaccuracy (β_j). These appear to be separate variables for each subject and PVS. There are several practical consequences of this behavior.

First, subjects' scoring is a random process. This is expected behavior that must be accepted; not a flaw or fault that can be eliminated. These error terms explain apparent inconsistencies within a single subject's data and probably cause much of the lab-to-lab differences seen in datasets scored at multiple labs. These error terms also explain why the original video sequence is not rated "imperceptible" by DSIS and other double stimulus subjective methods.

We observe that the Δ_i, α_i, and β_j distributions combined span about $\pm$ 25% of the rating scale. This may be why prior research concluded that a discrete 5-level ACR scale is just as accurate at measuring MOS as a continuous scale [2], [3]. These error terms are so large that rounding to a discrete rating is unimportant. Also this large error could be the reason why subject demographics factors appear irrelevant.

Second, we propose that subjective data should sometimes be normalized by removing μ_{Δ_i}, as per (13). Whether or not to remove μ_{Δ_i} depends upon the type of data analysis.

- When the analysis focuses on MOS comparisons, then μ_{Δ_i} should be removed. Most subjective tests use this type of MOS analysis. The sensitivity of statistical comparisons between stimuli usually improves but the cost of the subjective test does not change.
- When the analysis compares objective and/or subjective data with user descriptions (e.g., from blogs, forums, or questionnaires), then MOS and subject bias should be retained.
- When the analysis focuses on subject behavior, then the analysis could focus only on μ_{Δ_i}. The vqegMM2 dataset [13] provides an example.

Third, the number of subjects in an experiment can be reduced, if each subject scores each PVS multiple times. We saw that one rating from each of 18 subjects yields approximately the same accuracy as three ratings from each of seven subjects. By extension, one rating from each of 24 subjects should yield approximately the same accuracy as three ratings from each of nine subjects. This ratio is of interest, because Pinson *et al.* [13] recommend 24 subjects for an ACR experiment. This technique would not be appropriate when the goal of the experiment is to accurately characterize the magnitude of ψ_j, because the smaller subject pool allows for less averaging of subject biases.

Fourth, we cannot support the use of repeated sequences to screen subjects. Subjects are unable to perfectly repeat their

prior score. Obtaining an accurate estimate of ϵ_{ijr} for a subject would require a large number of repeated scores, but proposals for screening subjects by repeated scores use a small number of PVSs. Inaccuracies can occur randomly and are thus unlikely to indicate poor behavior on the part of the subject.

Similarly, subject screening techniques need to be rethought. For example, the screening technique from ITU-R Rec. BT.500 and thresholds based on Pearson correlation tends to reject subjects with large α_i. The problem is with the confidence interval. Since we estimate the expected value, we would like to have as good an estimation as possible. The 95% confidence interval is approximately

$$\pm 2 * \sigma / \sqrt{N} \tag{40}$$

where N is the sample size. Removing inaccurate subjects decreases σ_j yet also decreases N. This can increase the confidence interval.

Fifth, when the subject pool for a single experiment is split among two or more labs, the raw scores should be pooled. That is, when all subjects observe and rate an identical set of stimuli, then the subjects represent the larger pool of all people. Thus, their scores can be mingled without applying any scaling or fitting function. This is neither surprising nor novel, since this procedure is in common use (e.g., by T1A1 and VQEG).

Sixth, if you want to detect subject demographics and laboratory environmental factors, then the number of subjects needs to increase dramatically over what is used today. This would explain why researchers had difficulty identifying subject demographics and laboratory environmental factors that explain lab-to-lab differences.

Seventh, the error distributions of subjects and PVSs can be characterized by the following:

- M_i defined in (11) and shown in Fig. 2(c); and
- P_j defined in (12) and shown Fig. 3.

IX. Future Work

We know that (2) is imperfect. It assumes that each subject has an identical and symmetrical error distribution. We know this is not true, but these differences may be small enough that they can be safely ignored. It assumes that each subject's bias remains constant over time, yet the Mann Whitney U test indicates that six of our 28 subjects had some significant differences in μ_{Δ_i} depending upon the session number. Equation (2) assumes that subject bias is limited to an additive factor, yet we observe an occasional subject who chooses to use ACR as a 4-level or 3-level scale (e.g., ignoring the "excellent" category, the "bad" category or both). Equation (2) ignores time dependencies within the session (i.e., sequence order). A more complicated model of subject behavior would be an interesting topic for future investigation.

Our model is based on the 5-level scale and video quality evaluation. Based on results presented in [2] and [3], we expect the main concept of the model should not change if ACR is replaced with another scale (e.g., DSCQS, DSIS, ACR11, or PC). We hope to generalize the obtained model for different services, such as image or sound quality, and to include terms that explain lab-to-lab differences. These generalizations are left as a future work, which should be made in cooperation with other researchers.

In the AGH/NTIA experiment, we just touched the problem of irrelevant subjects. An experiment focusing on different aspects of why subjects are irrelevant can be proposed. The proposed model can be both validated if it is relevant to detect specific subjects behavior and eventually extended to a more perfect model.

We are also planning to work on better model parameter estimation methods that do not call for such large and impractical experiments. We will focus on methods that do not call for repeated scores.

The subject model should be used by the community in order to evaluate different experiments and better understand particular subjective results. We are planning to evaluate more subjective data which are available online, especially data in [26] and Qualinet databases.[4]

Appendix A
Parameters Estimation

The MATLAB code estimates μ_{ψ_i}, μ_{Δ_i}, α_i, and β_j from the raw subjective data can be found at http://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx. Both codes reflecting a simplified for of (16) that assumes $r = 1$ and $r > 1$ are available. Also any update of the estimation method will be described at this web page.

Appendix B
Variance of Random Variable Drown From Different Distributions

If a random sample is drawn from N independent distributions with mean zero and finite variance, then the variance of the obtained sample is a weighted sum of the variances of the distributions used, where the weight is the probability of choosing a particular distribution.

Proof: Z is drawn from different distributions having probability density functions $f_i(x)$ and variance σ_i^2; each distribution $f_i(x)$ generates random variable X_i. Let us assume that all distributions are independent, $E(X_i) = 0$ for all i, and each variable X_i is drawn with probability p_i. Then according to the total probability rule Z has distribution $g(z) = \sum_{i=1}^{N} p_i f_i(z)$. The variance of any random variable is given by $E((Z - E(Z))^2)$. Nevertheless, we assumed that all variables X_i have mean zero, so $E(Z) = 0$. Therefore, both X_i and Z variances can be computed using $E(X_i^2)$ or $E(Z^2)$ respectively. Let us compute $E(Z^2)$

$$\begin{aligned} E(Z^2) &= \int z^2 g(z) dz = \int z^2 \sum_{i=1}^{N} p_i f_i(z) dz \\ &= \sum_{i=1}^{N} p_i \int z^2 f_i(z) dz \\ &= \sum_{i=1}^{N} p_i E(X_i^2) = \sum_{i=1}^{N} p_i \sigma_i^2 \end{aligned} \tag{41}$$

which finishes the proof.

[4][Online] Available: http://dbq.multimediatech.cz/

ACKNOWLEDGMENT

The authors would like to thank B. Ćmiel for important remarks on the parameter estimation and S. Voran for his review of the equations.

REFERENCES

[1] J.-S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 564–571, Feb. 2014.
[2] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Proc. 2nd Int. Workshop Quality Multimedia Experience*, Jun. 2010, pp. 82–87.
[3] Q. Huynh-Thu and M. Ghanbari, "Modelling of spatio-temporal interaction for video quality assessment," *Image Commun.*, vol. 25, no. 7, pp. 535–546, Aug. 2010.
[4] S. Voran and S. Wolf, "The development and evaluation of an objective video quality assessment system that emulates human viewing panels," in *Proc. Int. Broadcast. Conv.*, Jul. 1992, pp. 504–508.
[5] M. Pinson, N. Staelens, and A. Webster, "The history of video quality model validation," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process.*, Sep. 2013, pp. 458–463.
[6] *T1A1.5 video quality project: GTE labs analysis*, Contribution to ANSI Committee T1, Sep. 1994.
[7] *Multiple Stimulus With Hidden Reference and Anchor (MUSHRA)*, ITU-R BS.1534, 2003.
[8] D. Shi, D. Xi, and H. Ruimin, "Bit-rate reduction using psychoacoustical masking model in frequency domain linear prediction based audio codec," in *Proc. 2nd Int. Conf. Indust. Mechatronics Automat.*, May 2010, vol. 2, pp. 229–232.
[9] R. K. Muralimanohar, C. Kronen, K. Arehart, J. Kates, and M. K. Pichora-Fuller, "Quality of voices processed by hearing aids: Intra-talker differences," in *Proc. Meetings Acoust.*, 2013, vol. 19, no. 1, p. 3388.
[10] C. Hoene, J. Valin, K. Vos, and J. Skoglund, Internet Engineering Task Force (IETF), Nov. 2013, "Summary of opus listening test results draft-ietf-codec-results-03," Accessed on: May 01, 2014.
[11] S. Tavakoli, J. Gutierrez, and N. Garcia, "Subjective quality study of adaptive streaming of monoscopic and stereoscopic video," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 684–692, Apr. 2014.
[12] P. Kovacs, K. Lackner, A. Barsi, A. Balazs, A. Boev, R. Bregovic, and A. Gotchev, "Measurement of perceived spatial resolution in 3D light-field displays," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 768–772.
[13] M. H. Pinson, "The influence of subjects and environment on audiovisual subjective tests: An international study," *J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, 2012.
[14] A. Ostaszewska and S. Żebrowska Łucyk, "The method of increasing the accuracy of mean opinion score estimation in subjective quality evaluation," in *Wearable and Autonomous Biomedical Devices and Systems for Smart Environment*, ser. Lecture Notes in Elect. Eng., A. Lay-Ekuakille and S. Mukhopadhyay, Eds. Berlin, Germany: Springer, 2010, vol. 75, pp. 315–329.
[15] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Proc. SPIE, Adv. Image Video Commun. Storage Technol.*, Feb. 1995, vol. 2451, pp. 90–101.
[16] T. Hobfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *Proc. 3rd Int. Workshop Quality Multimedia Experience*, Sep. 2011, pp. 131–136.
[17] M. Pinson and L. Janowski, "AGH/NTIA: A video quality subjective test with repeated sequences," Nat. Telecommun. and Inf. Admin., Boulder, CO, USA, Tech. Rep. TM-14-505, Jun. 2014.
[18] A. Catellier and L. Connors, "Web-enabled subjective test (WEST) research tools manual," Nat. Telecommun. and Inf. Admin., Boulder, CO, USA, Tech. Rep. HB-14-501, Jan. 2014.
[19] M. Pinson *et al.*, "Report on the validation of video quality models for high definition video content," Nat. Telecommun. and Inf. Admin., Boulder, CO, USA, Tech. Rep., 2010 [Online]. Available: http://www.its.bldrdoc.gov/media/4212/vqeg_hdtv_final_report_version_2.0.zip
[20] M. H. Pinson, C. Schmidmer, L. Janowski, R. Pepion, Q. Huynh-Thu, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "Subjective and objective evaluation of an audiovisual subjective dataset for research and development," in *Proc. 5th Int. Workshop Quality Multimedia Experience*, 2013, pp. 30–31.
[21] M. Pinson, S. Wolf, and G. Cermak, "HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss," *IEEE Trans. Broadcast.*, vol. 56, no. 1, pp. 86–91, Mar. 2010.
[22] L. Janowski and M. Pinson, "Subject bias: Introduction a theoretical user model," in *Proc. 5th Int. Workshop Quality Multimedia Experience*, to be published.
[23] P. Le Callet, S. Maller, and A. Perkis, "Qualinet white paper on definitions of quality of experience (2012)," Qualinet, Boulder, CO, USA, 2012 [Online]. Available: http://www.qualinet.eu/images/stories/whitepaper_v1.1_dagstuhl_output_corrected.pdf
[24] D. Kahneman, *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux, 2011.
[25] A. Głowacz, M. Grega, P. Gwiazda, L. Janowski, M. Leszczuk, P. Romaniak, and S. P. Romano, "Automated qualitative assessment of multi-modal distortions in digital images based on GLZ," *Special Ann. Telecommun. Quality Experience Socio-Economic Issues Netw.-Based Services*, vol. 65, no. 1–2, pp. 3–17, Feb. 2010.
[26] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

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 Department 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.

Margaret Pinson received the B.S. and M.S. degrees in computer science from the University of Colorado at Boulder, Boulder, CO, USA, in 1988 and 1990, respectively.

Since 1988, she has been with the Institute for Telecommunication Sciences, National Telecommunication and Information Administration (NTIA/ITS), Boulder, CO, USA, where she is currently a Co-Chair of the Video Quality Experts Group (VQEG), Co-Chair of the VQEG Independent Lab Group (ILG), Co-Chair of the VQEG Audiovisual HD (AVHD) project, and an Associate Rapporteur of Questions 2 and 12 in ITU-T Study Group 9.

Mrs. Pinson was the recipient of the Best Paper Award at the 2014 Workshop on Quality of Multimedia Experience (QoMEX) for a proposed model of rating behavior based on subject bias and subject error.