The Precision and Repeatability of Media Quality Comparisons: Measurements and New Statistical Methods

Margaret H. Pinson

***Abstract*—This paper calculates confidence intervals for 89 datasets that use the 5-level Absolute Category Rating (ACR) method to evaluate the quality of speech, video, images, and video with audio. This data allows us to compute the subjective test confidence interval (ΔS_{CI}) for 5-level ACR tests. We use a confusion matrix to compare conclusions reached by 88 lab-to-lab comparisons, 22 method-to-method comparisons, and 12 comparisons between expert and naïve subjects. We estimate the differences in conclusions reached by ad hoc evaluations, compared to subjective tests. We recommend using the *disagree* incidence rate to identify lab-to-lab differences (i.e., the likelihood that significantly different stimulus pairs receive opposing rank order from the two labs). *Disagree* incidence rates above 0.31% are unusual enough to warrant investigation and *disagree* incidence rates above 1.0% indicate differences in method, test environment, test implementation, or subject demographics. These incidence rates form the basis for a new statistical method that calculates the confidence interval of a metric (ΔM_{CI}). When ΔM_{CI} is used to make decisions, the equivalence to a video-quality test (EVQT) method determines whether a metric acts similarly to a subjective test. When ΔM_{CI} is not used, the metric is likened to a certain number of people in a video-quality test (PVQT). This information will help users make the better decisions when applying quality metrics. The algorithm code is made available for any purpose. Most of the ratings used in this paper come from open datasets.**

***Index Terms*—Audiovisual quality, CI, confidence interval, confusion matrix, false ranking, image quality, metric, MOS, precision, statistics, subjective test, video quality.**

I. Introduction

SUBJECTIVE tests are the most accurate way to assess audio and video quality. They are also expensive, slow, and thus rare. Quality metrics are less accurate but provide fast insights into video and audio quality. These sweeping statements, while correct, fail to convey a deeper understanding that end-users need—such as the likelihood of choosing the worst quality system by mistake.

Manuscript received 5 October 2022; revised 6 December 2022; accepted 21 December 2022. Date of publication 10 February 2023; date of current version 7 June 2023. This work was supported by the National Telecommunications and Information Administration (NTIA), Institute for Telecommunication Sciences (ITS).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by various review boards, including the NTIA Chief Council, and performed in line with ITU-R Rec. BT.500 and ITU-T Recs. P.910, P.911, P.913, and P.800.

The author is with the National Telecommunications and Information Administration, Institute for Telecommunication Sciences, Boulder, CO 80305 USA (e-mail: mpinson@ntia.gov).

Digital Object Identifier 10.1109/TBC.2023.3236528

Ultimately, the choice between subjective tests and quality metrics is a false dichotomy. Ad hoc evaluations are more common—like engineers relying on their own observations. Some of these people know that ad hoc evaluations are inaccurate but have no viable alternative.

ITU-T Rec. P.1401 identifies best practices for comparing quality metrics to subjective tests. P.1401 focuses on the needs of standards organizations to assess the accuracy of quality metrics. By contrast, companies must choose between relying on ad hoc evaluation, deploying a quality metric, or conducting a subjective test. They would like to understand how this choice impacts the likelihood of making an error. ITU-T Rec. P.1401 cannot answer this question.

This paper offers solutions. We begin by examining previously proposed statistics that characterize the accuracy, repeatability, and precision of subjective tests and objective metrics (Section II). We then define terms and summarize the subjective tests that we will be analyzing (Section III). We complete our introduction by describing our strategy for designing new statistical methods (Section IV).

Section V analyzes the observed precision of subjective tests. We defined the *subjective test's confidence interval* (ΔS_{CI}) as the minimum difference in mean opinion scores (MOS) at which 95% of the pairs will be statistically different (according to the Student's *t*-test using a 95% confidence level). We focus on the most popular rating method—the 5-level Absolute Category Rating (ACR) method—to maximize the available data. We calculate ΔS_{CI} for 91 datasets to indicate trends within a single test.

Section VI compares ratings from multiple labs to analyze similarities and differences in their conclusions. We will use a confusion matrix to classify the conclusions reached by two labs. We define the *disagree* incidence rate as the likelihood of two tests reaching opposing conclusion on the quality ranking of stimuli **A** and **B**, when the paired Student's *t*-test indicates that these stimuli have different quality (using a 95% confidence level). We examine *disagree* incidence rates from 88 lab comparisons, 22 method comparisons, and 12 comparisons between naïve and expert subjects.

This allows us to estimate expected *disagree* incidence rate for well-designed and carefully conducted subjective tests. We conclude that two subjective tests can be considered equivalent when the *disagree* incidence rate is below 0.31%. *Disagree* incidence rates above 0.31% are unusual enough to warrant investigation and *disagree* incidence rates above 1.0%

indicate significant differences in method, test environment, test implementation, or subject demographics.

Section VII infers the *false ranking* incidence rates for ad hoc evaluations. We propose *false ranking* as a logical extension of the *disagree* incidence rate. *False ranking* occurs when the ad hoc evaluation indicates that media **A** has *better* quality than media **B**, but a subjective test concludes that media **A** has significantly *worse* quality than media **B**. *False ranking* explains the ad hoc evaluation or metric's performance in terms that an end user can understand: the odds of making a mistake.

Section VIII compares the decisions reached by a subjective test (using the Student's *t*-test) with conclusions reached by the ad hoc evaluation, where any change in quality is significant. We will use Monte Carlo simulations to divide ≈70 subjects into a formal subjective test with 24 subjects and an ad hoc assessment of 1, 2, or 3 subjects. We will repeat this analysis for pilot tests of 6, 9, or 12 subjects. This lets us estimate the expected *false ranking* incidence rate for ad hoc evaluations.

Then we switch to objective metrics and repeat these analyses, in the reverse order. When confidence intervals (CIs) are *not* used to make decisions, we can equate the performance of the quality metric to a certain number of *people in a video-quality test* (PVQT). We do this by computing the *false ranking* incidence rate of the metric (relative to a subjective test) and comparing it to our observed *false ranking* incidence rates of ad hoc evaluations and pilot tests.

Section IX establishes the *metric's confidence interval* (ΔM_{CI}). This formula applies our previous observations (e.g., ΔS_{CI} and *disagree* incidence rates) to calculate a confidence interval at which the metric has error rates similar to a subjective test. ΔM_{CI} allows users to perform significance tests on metric values—instead of carelessly assuming that any change in metric value is significant.

Section IX also establishes the *equivalence to a video-quality test* (EVQT) to determine whether a metric acts similarly to a subjective test. We will use a confusion matrix to classify the conclusions reached by a lab with conclusions reached by the metric (using ΔM_{CI}). We will compare this lab-to-metric data to our lab-to-lab data. If the correct decision and error incidence rates fall within the expected behavior of lab-to-lab comparisons, then we can conclude that the metric is acting like a subjective test lab.

Section X summarizes our findings and draws the reader's attention to important Tables.

This paper builds upon research presented in [1]. Some details have been changed, for improved clarity when these methods are used in later publications. Code implementing the new statistical methods is available in the GitHub repository NRMetricFramework [2], but none of these new statistics are suitable for comparing metrics and deciding which is more accurate.

II. Background

Let us begin by examining previously proposed statistics that characterize the accuracy, repeatability, and precision of subjective tests and objective metrics. After introducing the differences between video and speech tests, we will proceed chronologically.

A. Differences Between Video- and Speech-Quality Tests

ITU-R Rec. BT.500 addresses the specialized needs of broadcasters and contribution-quality television to conduct subjective video-quality tests. BT.500 recommends a minimum of 15 subjects, as it has for decades. By contrast, ITU-T Rec. P.913 describes best practices for Internet video, mobile devices, and new video technologies. P.913 recommends a minimum of 24 subjects. In both Recommendations,[1] MOSs are calculated for each media (an image, silent video, or audiovisual file).

ITU-T Rec. P.800 describes speech-quality tests. The test methods are very similar, but the number of listeners per file is often quite low (e.g., 6 or 8). MOS is calculated by averaging all subject ratings for all speech samples associated with one system. When the same test is conducted at multiple labs, each lab uses different speech recordings (e.g., phonetically-balanced sentences in the native language of that area).

This paper uses the video convention. Each MOS describes one media file, which we will refer to as a stimulus. Due to this difference, we will focus on statistics for video-quality analysis (VQA) and image-quality analysis (IQA). Most of these statistical methods compare subjective test methods and objective metrics

B. Statistical Analyses of Subjective Ratings

Cermak and Fay [3] use Analysis of Variance (ANOVA) in their 1994 Contribution to the T1A1 subcommittee of the American National Standards Association (ANSI) accredited Alliance for Telecommunications Industry Solutions (ATIS). Cermak and Fey analyze the distribution of rating differences when subjects view and rate the same stimuli twice. ANOVA rarely appears in later publications, due to the constraints on experiment design.

Using this same T1A1 dataset, Webster [4] proposed no-reference (NR) metrics spatial information (SI) and temporal information (TI). Metrics SI and TI characterize the coding complexity of videos in a subjective test. SI and TI appear in ITU-T Rec. P.910 and continue to be widely used.

Over the next decade, Pearson correlation became the de facto standard for lab-to-lab comparisons. Pinson and Wolf [5] has the most data of this sort, with three tables that each compare MOSs from 13 or 14 labs. This report also provides scatter plots and linear fits (gain and offset).

The Student's *t*-test can be used to compare individual systems, as per the comparison of H.264 and MPEG2 in [6]. P.913 recommends the Student's *t*-test for such analyses.

In 2010, Tominaga et al. [7] evaluated the pros and cons of eight subjective rating methods. They compare the MOSs produced by different methods using Pearson correlation and Spearman correlation. They also compute statistics for each subjective method separately: (1) a histogram showing the distribution of MOSs, (2) the range of MOSs in the test (MOS range), (3) the mean of the 95% confidence intervals

[1]We will ignore ITU-T Rec. P.910 and P.911 because they remain focused on standard-definition televisions and cathode ray tube (CRT) monitors. The Video Quality Experts Group (VQEG) is developing a proposal to merge and update P.910, P.911, and P.913. This proposal is expected to harmonize BT.500 and P.913 by identifying use cases for 15 subject tests.

(MCI), (4) MCI normalized by MOS range (MCI_{norm}), (5) the total assessment time required, and (6) the ease of evaluation, from a questionnaire. Tominaga et al. conclude that MOS_{norm}, assessment time, and ease of evaluation were the most sensitive to differences between methods. They recommend the 5-level ACR method.

Huynh-Thu et al. [8] compared different subjective methods and ultimately recommend ACR. They use rating distributions, confidence interval distributions, lab-to-lab scatter plots, ANOVA, the correlation between subjects (as a measure of repeatability), the fraction of media pairs that significantly differ, and linear fits (e.g., transforming from one scale to another). After these two studies were published, the 5-level ACR method notably increased in popularity, and other methods decreased in popularity.

Hoßfeld et al. [9] propose a statistic that characterizes the overall relationship between MOS and standard deviation of scores (SOS). This statistic is called α in [9], but we will use the authors' initials instead (HSE), because α is commonly used in equations. HSE uses the level of agreement or disagreement among subjects to estimate the reliability of the test data. Like SI and TI, HSE can help the reader understand a subjective test's characteristics.

Pinson et al. [10] present a battery of statistical analyses to explore the impact of multiple labs and test environments on MOSs. These statistics include Pearson correlation, the Kruskal-Wallis test, confidence intervals, random subsets of subjects, and a confusion matrix. Pinson et al. conclude that MOS is relative—we expect the ordering of impairments and relative distances to be replicable. Since MOSs are not absolute, we cannot rely upon MOS thresholds.

Le Moan et al. [11] compare the impact of side-by-side and one-after-another presentation methods on image-quality tests. Both are implemented with the 3-level Pair Comparison method (e.g., better, same, worse). They consider assessment time and a confusion matrix to compare each subject's ratings from these two methods (e.g., prefer **A**, prefer **B**, or tie), and then aggregate over subjects and content type.

Kumcu et al. [12] propose seven statistics for analyzing subjective methods, including four statistics taken from Mantiuk et al. [13] in identical or modified form. Two of the seven statistics are simple—assessment time (as per [7]) and Cohen's D—and the rest are too complex to be summarized here. In broad terms, they examine retrospective power, the rank order of stimulus pairs, and the probability that the subjective test will detect significant differences between stimulus pairs.

Nehmé et al. [14] calculate the accuracy of a subjective test as the fraction of stimulus pairs that are significantly different, given the unpaired two-sample Wilcoxon test and a sub-set of the available subjects. This produces a range of accuracy estimates for any given number of subjects. They plot this range of accuracy estimates (on the y-axis) as a function of the number of subjects (on the x-axis). Sample plots are shown for two subjective methods and two labs.

Most of these papers evaluate the impact on MOSs when there are changes to the test environment, subject demographics, number of subjects, or rating method. The exceptions are SI, TI, MCI_{norm}, and HSE, which generally characterize the subjective test and are intended to be compared across different publications.

C. Analyzing Objective Metrics

In the early 1990s, subject matter experts from T1A1 brainstormed improved methods to express the accuracy and precision of video-quality metrics. These discussions culminated in two proposed solutions.

The first proposed solution, resolving power, is described in Brill et al. [15], ATIS T1.TR.72, and ITU-T Rec. J.149. Loosely described, resolving power is the threshold at which 95% of all stimulus pairs are significantly different. Resolving power is calculated for a specific subjective dataset. Either the metric is mapped to the subjective test or vice versa. Resolving power was rejected by industry and subject matter experts. Resolving power yields large thresholds and a pessimistic conclusion that even the best metric has minimal practical value. See [5] for example data.

Resolving power has three design flaws. First, resolving power uses the Student's *t*-test when comparing subjective data but CIs when comparing metric data. This predisposes the measurement sensitivity in favor of subjective testing. Second, resolving power is calculated for a specific dataset and greatly influenced by quirks of its experiment design (e.g., MCI_{norm}, number of subjects, and MOS range). Third, resolving power assumes (without proof) that subjective tests are perfect.

The second proposed solution is based on a confusion matrix and appears in [15] and ATIS T1.TR.72. In a nutshell, the idea is to use a confusion matrix that classifies the conclusions reached by a subjective test with the conclusions reached by the metric, measured as a function of the change in metric value. This confusion matrix idea also failed to gain traction. One problem is that the method proposes many options for the metric's CI without recommending how to choose among them.

From 1994 to around 2010, objective metric analyses and subjective test analyses used similar statistics. For example, Cermak and Fay [3] use regression and ANOVA. Over this period, the Video Quality Experts Group (VQEG) considered various statistics and eventually settled on Pearson correlation, root mean square error (RMSE), and outlier ratio, with CIs and significance tests for each. Nonlinearities in the subjective data are removed by fitting the objective data to the MOSs using a 3rd order monotonic polynomial fit. These techniques are described in ITU-T Rec. P.1401. Sample analyses and code can be found in [5].

Pearson correlation, RMSE, Outlier Ratio, and variations like Spearman correlation fail to acknowledge or accommodate the variation in opinion among subjects. For example, Pinson et al. [10] identifies a flaw when Pearson correlation is used to analyze MOSs. Pearson correlation is proportional to the fraction of the rating scale that is spanned by the dataset's MOSs (MOS range). Thus, lower values of Pearson correlation do not necessarily indicate an inaccurate metric—they could instead be caused by a narrow range of MOSs. This problem led to the proposal of two new statistics.

One solution is epsilon insensitive RMSE, which is abbreviated RMSE*. Epsilon insensitive RMSE is a simple variant of RMSE that considers the confidence interval of the individual MOS scores.[2] Epsilon insensitive RMSE is described in ITU-T Rec. P.1401, and data can be found in Appendix I of ITU-T Rec. P.863.

Another solution is proposed by Krasula et al. [16]. They divide the dataset into pairs of media and compare the distance between metric values (Δ_{model}) with the conclusions reached by the subjective data (better, worse, or tie). For values of Δ_{model} near zero, the subjective data always indicates a tie; and for values of Δ_{model} far from zero, the subjective data always indicates a difference in quality. Krasula et al. focus on the overlap: values of Δ_{model} that could either be associated with ties or better/worse quality. Krasula et al. propose a set of three statistical tests that indicate whether one metric performs better than, the same as, or worse than another metric. One test examines whether the metric correctly distinguishes between media that are statistically identical vs statistically different, and the other two tests examine rank ordering.

Tiotsop et al. [17] adopt a completely different approach of analyzing the differences among multiple objective metrics, without any subjective data. If multiple metrics agree, the commonly indicated ratings are likely to be accurate. If multiple metrics disagree, subjective testing is recommended.

III. Datasets

We will begin by describing our datasets and providing definitions related to subjective testing. People who are uninterested in these details are encouraged to read Section III-A before skipping ahead.

A. Overview, Terms, and Definitions

To understand the precision and repeatability of subjective tests, we collected individual subject ratings from 31 studies of video quality, audio quality, audiovisual quality, and speech quality. When aggregated, these studies contain 95 datasets conducted with the 5-level ACR method, 86 lab-to-lab comparisons (where two labs conducted the same subjective test), 20 rating method comparisons, and 12 comparisons of expert and naïve subjects. Expert subjects have specialized experience that naïve subjects lack (e.g., physicians vs a random sampling of people when rating ultrasound images).

In this section, we will describe each study and dataset. These summaries omit information that is not pertinent to our CI and lab-to-lab analyses, such as the media content, type of impairments, media duration, and test environment. If there is a discrepancy between the published study description and the freely available data, we will describe the available data. These rare discrepancies are not marked.

Through this paper, we will use the term "dataset" to refer to a set of media files that were rated by a common pool of subjects. The subjects may be associated with multiple labs, and different subjects may use different ratings method. However, all subjects must rate all media files.

[2]RMSE* appears to have been proposed during meetings of ITU-T Study Group 12. The Chairs were unable to recommend a reference.

If a study's experiment design does not conform to this definition, then we will split the study into multiple datasets. This will allow us to compare stimulus pairs within a single dataset using the two sample Student's *t*-test (i.e., compare the rating distributions of any two stimuli using the same pool of subjects). An occasional missing rating is acceptable. When this occurs, the two samples for the Student's *t*-test will have slightly different sizes (e.g., 16 ratings for stimulus **A** and 17 ratings for stimulus **B**).

For crowdsourcing, we will omit the restriction that all subjects must rate all media files in the dataset. Unlike lab studies, crowdsource subjects are not expected to rate all media files. Thus, we will analyze the crowdsource ratings as they are intended to be analyzed: as a single pool of subjects.

These studies were conducted according to ITU-R Rec. BT.500, ITU-T P.913, or ITU-T P.800 and thus follow best practices for subjective testing and the ethical treatment of subjects. Most of these studies use standard rating methods and scales: Absolute Category Rating (ACR) method, the Double Stimulus Comparison Scale (DSIS), the Double Stimulus Continuous Quality Scale (DSCQS), the Comparison Category Rating (CCR), Single Stimulus Continuous Quality Evaluation (SSCQE), and Forced Choice (FC). Unless otherwise stated, ACR is implemented as a 5-level scale, CCR as a 7-level scale (-3 to 3), SSCQE as a 100-level scale, and FC as a 2-level scale.

The remaining studies compare standard methods with experimental rating methods. Boolean is a single-stimulus method where subjects rated whether the video quality was acceptable for public safety applications. The Content-Immersive Evaluation of Transmission Impairments (CIETI) method simulates realistic viewing conditions with a 5-level scale, longer video sequences, and changing impairment levels. Preference (PREF) is a variant of CCR with alternate instructions and a -50 to 50 scale. Dissimilarity (DISSIM) is a double-stimulus method where subjects rated the similarity of two sequences in terms of quality on a 100-level scale.

Most of these studies use conventional experiment designs. For speech-quality studies, this means a balance of phonemes and talkers for each impairment. For image and video studies, this means a set of high-quality recordings (original) are impaired identically (e.g., a set of compression bit-rates). Thus, each source medium is repeatedly played to the subjects. Video-quality studies with unrepeated source experiment designs are noted. Unrepeated source experiment designs avoid re-using the same source media. Strategies include photographing the same scenes with different cameras or culling a particular camera impairment from a large pool of content.

We will present datasets in the following order: studies by standards developing organizations (SDO), lab studies, field studies, and private studies. Most of the individual subject ratings and media files for SDO, lab, and field studies are available on the Consumer Digital Video Library (CDVL, www.cdvl.org). Exceptions are noted.

B. Datasets From Standards Developing Organizations

The studies described in this sub-section were conducted by SDOs. These tests represent an ideal of carefully designed and

executed lab studies. These datasets were designed with the aid of a considerable number of experts in the field in addition to the organizations identified. We identify organizations and countries to demonstrate that these datasets include diverse countries, cultures, and native languages.

1) VQEG FRTV Phase I [18], [19]: In 1999–2000, the Video Quality Experts Group (VQEG) conducted the Full-Reference Television (FRTV) Phase I validation test. The goal was independent validation of the performance of video-quality metrics that assess the quality of standard-definition television (625-line and 525-line). Independent test labs created four datasets, divided by bit-rate quality (high vs low) and video format (625-line vs 525-line). Each dataset contains 100 videos that were rated by either 67 or 70 subjects using the DSCQS method.

The subjective testing was carried out by eight labs. Each lab contributed ≈25% of the subjects in a particular test. Thus, each FRTV Phase I dataset enables six lab-to-lab comparisons for each test (choose two of four labs). Among all four tests, there are 24 lab-to-lab comparisons.

The 525-line tests (high and low quality) were contributed by Berkom (France); the Canadian Research Centre (CRC, Canada); Fondazione Ugo Bordoni (FUB, Italy); and Nippon Hoso Kyokai (NHK, Japan). The 625-line test subjects were contributed by Centre commun d'études de télévision et télécommunications (CCETT, France); Centro Studi e Laboratori Telecomunicazioni (CSELT, Italy); Department of Communications, Information Technology and the Arts (DCITA, Australia); and Radiotelevisione Italiana (RAI, Italy).

Two issues were identified with the FRTV Phase I datasets. First, each dataset contained a narrow range of quality, so it was difficult to differentiate between the performance of the objective metrics. Second, the ratings include occasional scoring inversion errors. Subjects watched both videos and then marked ratings on a paper scoring sheet. Accidents occurred where subjects wrote their rating of the earlier video where they should have written their rating of the later video, and vice versa.

2) VQEG FRTV Phase II [20], [21]: In 2002–2003, VQEG conducted the FRTV Phase II validation test. As with Phase I, the goal was independent validation of objective metrics for standard-definition television. FRTV Phase II conducted two subjective tests: a 625-line dataset with 70 videos rated by 27 subjects from FUB (Italy), and a 525-line dataset with 63 videos rated by 32 subjects each from CRC (Canada) and Verizon (USA), for a total of 64 subjects. The FRTV Phase II datasets enable one lab-to-lab comparison (CRC vs Verizon). Many other organizations contributed to other aspects of this endeavor. As with Phase I, the FRTV Phase II data likely contains scoring inversions errors, since these tests also used DSCQS and paper scoring sheets. Only the individual subject ratings are available on CDVL.

3) VQEG RRNR-TV [22]: In 2008–2009, VQEG conducted the Reduced-Reference and No-Reference television (RRNR-TV) validation test. The goal was independent validation of in-service video-quality metrics for standard-definition television. RRNR-TV produced two datasets that used the 5-level ACR method. The 625-line dataset contains 168 videos rated by 34 subjects: 18 from FUB (Italy) and 16 from the National Telecommunications and Information Administration (NTIA, USA). The 525-line dataset also contains 168 videos rated by 32 subjects: 16 from NEC (Japan) and 16 from Yonsei University (Republic of Korea). The RRNR-TV dataset enables two lab-to-lab comparisons (FUB vs NTIA and NEC vs Yonsei). CRC (Canada) was the principal investigator who designed these experiments. Only the individual subject ratings are available on CDVL.

4) VQEG HDTV [23]: In 2009–2010, VQEG conducted the high-definition television (HDTV) validation test. The goal was independent validation of in-service and out-of-service video-quality metrics for high-definition television. VQEG HDTV produced six datasets, numbered 1 to 6. Each dataset contains 168 videos and ACR ratings from 24 subjects from a single lab. Part of one dataset was discarded due to low-quality source material, so HD3 contains only 152 videos. The organizations most directly involved in creating these datasets and running subjects were University of Ghent (Belgium), NTIA (USA), the University of Nantes (France), Ericsson (Sweden), Acreo (Sweden), AGH University of Science and Technology (Poland), CRC (Canada), Psytechnics (United Kingdom), Deutsche Telekom (DT, Germany), and FUB (Italy). Many other organizations made significant contributions to the HDTV datasets. Some of the HDTV videos cannot be redistributed.

5) VQEG Hybrid [24]: In 2013–2014, VQEG conducted the Hybrid Perceptual/Bit-stream validation test. The goal was independent validation of video-quality metrics that supplement the decoded video with information extracted from the encoded bit-stream. The Hybrid test produced 10 datasets. Each dataset has between 114 and 184 videos. Each dataset has 5-level ACR ratings from 24 subjects from a single lab.

Each dataset included 24 sequences that overlapped with other Hybrid tests of the same video resolution. Due to the small size of this subset, we will only make lab-to-lab analyses for the five datasets with HD resolution. Because each comparison yields minimal data, we will aggregate the results of all ten lab-to-lab comparisons (i.e., all combinations of five labs) to obtain a more robust estimate. Thus, Table IV lists one lab-to-lab comparison for VQEG Hybrid.

The organizations most directly involved in creating these datasets were Acreo (Sweden), DT (Germany), RT-RK (Serbia), AGH University (Poland), University of Ghent (Belgium), Yonsei University (Republic of Korea), SwissQual (Switzerland), OPTICOM GmbH (Germany), FUB (Italy), NTIA (USA), University of Nantes (France), and Intel (USA). Many other organizations made significant contributions to the Hybrid datasets. The videos are not available on CDVL.

6) VQEG Multimedia Phase II [10]: In 2010–2011, VQEG conducted the Multimedia Phase II (MM2) test. Sixty audio-visual sequences were rated by six labs in diverse viewing conditions. The goal was to understand the impact of environmental variables on subject ratings. All labs used the 5-level ACR method.

All six labs ran subjects in controlled environments. Four labs also ran subjects in uncontrolled environments (e.g., restaurant, break area, or hallway). Thus, the MM2

dataset enables 45 lab-to-lab comparisons. The labs who contributed subjects were NTIA (USA), Intel (USA), OPTICOM (German), AGH University (Poland), Université de Nantes (France), and Technicolor R&D (France).

MM2 concluded that the number of subjects was the most important control variable. More subjects were needed in an uncontrolled environment to replicate the lab-to-lab statistics of controlled environments. However, the native language, speech comprehension, country of origin, and translation of the ACR scale labels did not seem to matter—or at least mattered so little that the difference was obscured by human factors. The results from this study were pivotal in the development of ITU Rec. P.913, which gives researchers more freedom to conduct subjective tests in the uncontrolled environments where mobile devices are likely to be used.

7) ITU-T P.Sup23 [25]: ITU-T Rec. P.Sup23 contains three speech-quality experiments that were created during the characterization phase tests for the ITU-T G.729 speech codec and made available on the ITU website.[3] Although each of the three experiments were conducted in multiple labs, lab-to-lab comparisons are not possible. Each lab used speech samples in their country's native language, as is the convention in speech-quality testing. This confounding factor would cause problems for the two sample Student's t-tests. Therefore, we will treat each lab's subjects as a separate dataset.

P.Sup23 experiment 1 contains three datasets, each with 176 speech samples and 5-level ACR ratings from 24 subjects; P.Sup23 experiment 2 contains three datasets, each with 136 speech samples and CCR ratings from 48 subjects; and P.Sup23 experiment 3 contains four datasets, each with 200 speech samples and 5-level ACR ratings from 24 subjects. These datasets were created by AT&T (USA), CNET (France), CSELT (Italy), Nortel (formerly Bell-Northern Research, Canada), and Nippon Telegraph and Telephone (NTT, Japan).

8) SDO Emulation [1]: We will use the VQEG HDTV and VQEG Hybrid studies to emulate the response of 15, 9, and 6 subjects, by randomly selecting subsets of the available subjects and aggregating the response of all 16 datasets. This random selection is repeated ten times.

9) VQEG Multimedia [26]: In 2007–2008, VQEG conducted the multimedia validation test. Agreements between the participants prohibit the use of these datasets for most purposes, but we obtained permission to use the anonymized ratings and metric values to develop improved methods for analyzing metric performance. We will use the 13 VGA datasets to demonstrate our new methods. Each dataset contains 166 videos and 5-level ACR ratings from 24 subjects from a single lab.

C. Lab Studies

The studies described in this sub-section were conducted by various academic and industry researchers in lab settings. These studies were carefully designed and executed. Each lab study involved fewer organizations and experts than the SDO studies—but the use cases and media are much more varied.

1) Ghent Denoising [12]: Ghent Denoising is a video-quality study conducted with expert subjects and naïve subjects using three different methods: FC, PREF, and DISSIM. The FC dataset contains 70 videos rated by 19 experts and 18 naïve subjects. The PREF dataset contains 70 videos rated by 18 experts and 20 naïve subjects. The DISSIM data contains 70 videos rated by 18 experts and 20 naïve subjects. The Ghent Medical study will only be used to compare expert and naïve subjects, enabling three such comparisons.

2) Ghent Medical [12]: Ghent Medical is a video-quality study conducted with expert subjects and naïve subjects using four different methods: 100-level ACR, FC, PREF, and DISSIM. The same subjects were used for all four methods, but some subjects did not rate all methods. The ACR dataset contains 20 videos rated by 9 experts and 16 naïve subjects. The FC dataset contains 70 videos rated by 10 experts and 17 naïve subjects. The PREF and DISSIM datasets contains 70 videos rated by 10 experts and 16 naïve subjects. The Ghent Medical study will only be used to compare expert and naïve subjects, enabling four such comparisons.

3) ITS 2010 [27]: ITS 2010 is an audiovisual study that explores the relationship between audio quality and video quality in the overall audiovisual quality. Conducted in 2010 by NTIA, this dataset contains 240 media, rated by ∼26 subjects on a 5-level ACR scale. This dataset is named after our laboratory within NTIA: the Institute for Telecommunication Sciences (ITS).

4) ITS AV-Sync 2010 [28]: ITS AV-Sync 2010 is an audiovisual study that explores the relationship between audio quality, video quality, and delay in the overall audiovisual quality. Conducted in 2010 by NTIA, this dataset contains 407 media that were rated on a 5-level ACR scale. The media were divided into overlapping subsets, each containing 297 media. These two datasets were rated by 12 and 16 subjects.

5) Public Safety #1 [29]: Public Safety #1 (PS1) studied the quality required by first responders for their video systems. Conducted by NTIA in 2005–2006, 16 first responders rated 400 videos on the 5-level ACR scale, and then rated whether the video quality depicted was acceptable for public safety applications, on a Boolean scale.

6) Public Safety #2 [30]: Public Safety #2 (PS2) builds on PS1. Conducted by NTIA in 2006, 19 first responders rated 576 videos on the 5-level ACR scale, and then rated whether the video quality depicted was acceptable for public safety applications, on a Boolean scale.

7) SJTU 4K [31]: The Shanghai Jiao Tong University (SJTU) 4K dataset was created in 2016 in response to the need for research-free datasets with 4K video content (3840 × 2160, 30fps). SJTU 4K aided the study adaptive bit rate (ABR) bitrate estimation. SJTU 4K contains 60 videos with HEVC/H.265 compression that were rated by 42 subjects on the DSIS scale. Individual subject ratings are only available for 30 videos rated by 28 subjects.

[3]ITU-T P.Sup23 constrains the use of the speech files to the development of new and revised ITU-T Recommendations. This paper does not use the speech files. Also, our goal is to develop and socialize new analysis techniques for potential inclusion in ITU-T Rec. P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective-quality prediction models."

8) UPM-Acreo [32]: The UPM-Acreo study compares 5-level ACR with an experimental method, CIETI, which produces single-stimulus ratings on a 5-level scale. Conducted in 2015 by Universidad Politécnica de Madrid (Spain) and Acreo (Sweden), the study includes 132 audiovisual sequences. Three subject pools rated the media differently: 20 subjects rated the video (no audio) on the 5-level ACR scale, 22 subjects rated the video (no audio) on the CIETI scale, and 21 subjects rated the audiovisual sequences on the CIETI scale. This allows three lab-to-lab comparisons that include the confounding factor of rating method. Only the individual subject ratings are available on CDVL.

D. Field Studies and Crowdsourcing

This section describes field studies and crowdsourcing studies. These studies were as rigorous as the lab studies, but they include unconventional elements (e.g., unrepeated scene designs, and camera capture impairments). This section includes studies with experimental designs, prototype tests with few subjects, and real-world impairments with confounding factors such as camera capture.

1) AGH/NTIA/Dolby [33]: This 2015 video-quality study is named after the three labs who contributed subjects: AGH University (Poland), NTIA (USA), and Dolby (USA). The goal was to examine experiment designs that do not re-use source content, including their impact on subject scoring behaviors. The same 230 videos were rated on the 5-level ACR method, with an uneven distribution of subjects among the three labs (31, 22, and 18).

2) CCRIQ [34]: Consumer Content Resolution and Image Quality (CCRIQ) study analyzes the image quality produced by consumer cameras: smartphones, tablets, compacts, and digital single lens reflex (DSLR). CCRIQ uses an unrepeated experiment design where the same scenes were photographed with 23 different cameras. Three labs contributed subjects to this 2014–2015 experiment: Intel (USA), NTIA (USA), and the University of Ghent (Belgium). The images and each lab's subjects were divided into two non-overlapping pools. Thus, CCRIQ contains two datasets: 221 images rated by 26 subjects on the ACR scale, and 171 images rated by 27 subjects on the ACR scale. Each lab ran 8 or 9 subjects for each image pool. This enables three lab-to-lab comparisons for each dataset, for a total of six lab-to-lab comparisons.

The CCRIQ test was conducted on two identical monitors, placed side-by-side. One was configured with 4K resolution (3840 $\times$ 2160) and the other HD resolution (1920 $\times$ 1080). This enables two method-to-method comparisons for HD vs 4K monitors (one for each pool). For all other analyses, we will follow the protocol of the original study and pool the HD and 4K monitor ratings.

3) CCRIQ2 and VIME1 [35]: This 2018 image quality-study was conducted by AGH University (Poland) and NTIA (USA) to analyze unrepeated scene experiment designs using camera capture impairments. This study contains two datasets: one with images left over from CCRIQ (created but not used) and the other with images photographed by VQEG's *Video and Image Models for Consumer Content Evaluation* (VIME) project. The CCRIQ2 dataset contains 88 images that were rated by 19 subjects using the 5-level ACR scale. The VIME1 dataset contains 101 images that were rated by 22 subjects using the 5-level ACR scale.

4) ITS4S [36]: ITS4S, ITS4S2, ITS4S3, and ITS4S4 form a series of four studies conducted by ITS. These image and video-quality studies were designed specifically to enable no-reference metric development. All four studies use unrepeated scene designs.

The first study, ITS4S, was conducted in 2017–2018 as a proof of concept for many novel design choices, including 4-second video sequences, the "skip" rating option, unrepeated scenes, and a few videos where the original production quality is poor or worse. The full test contains 813 videos rated by 27 subjects on the 5-level ACR scale. A subset of 212 videos were later rated by 24 subjects at AGH University (Poland). This enables one lab-to-lab comparison.

5) ITS4S2 [37]: ITS4S2 was conducted in 2018–2019, using similar techniques as ITS4S. ITS4S2 is an image-quality test that contains a diverse selection of images with camera impairments. Some of these images were collected by VIME. The dataset contains 1,429 images that were rated by 16 subjects on the 5-level ACR scale.

6) ITS4S3 [38]: ITS4S3 was conducted in 2018–2019. ITS4S3 contains six sessions, each with 99 videos that depict camera impairments in the context of a first-responder application (e.g., fireground, crime scene, search & rescue). The videos in each session were rated by different subjects at a public safety conference, using the 5-level ACR scale. Each dataset was rated by between 13 and 19 subjects. For three of the six datasets, the subject pool contains enough first responders to enable comparisons between naïve subjects and expert subjects.

7) ITS4S4 [39]: ITS4S4 was conducted in 2019. ITS4S4 contains 196 videos depicting camera pans, real and simulated, that were rated by 26 subjects on the 5-level ACR scale. The ratings sessions for ITS4S3 and ITS4S4 occurred at a large meeting venue, in a quiet room.

8) 401, 501, 701 [40]: Datasets 401, 501, and 710 are crowdsource studies published in 2020. Dataset 401 (by Psytechnics) contains 1,152 speech files, each rated by 8 subjects. Dataset 501 (by SwissQual) contains 200 speech files, each rated by 24 subjects. Dataset 701 (by Dolby) contains 1,152 speech files, each rated by 8 subjects. All three datasets use the ACR scale. These datasets were intended to be analyzed per condition instead of per file, with 192, 96, and 128 conditions, respectively.

E. Private Studies

The remaining datasets are unpublished. Only limited information can be made available.

1) Private Speech #1: Private Speech #1 contains results from a 5-level ACR test conducted on narrowband speech codecs using simulated wireless channels. This dataset contains 1,359 speech files, each rated by $\approx$11, $\approx$22, or 43 subjects on the 5-level ACR scale. This lab study was intended to be analyzed per speech condition, with either 344 or 440 ratings per condition.

2) Private Speech #2: Private Speech #2 contains results from proprietary subjective ACR tests conducted on narrow-band speech codecs using wireline and simulated wireless channels. This dataset contains 2,432 speech files, each rated by 8 subjects on the 5-level ACR scale. This lab study was intended to be analyzed per speech condition, with 512 ratings per condition.

3) Private Speech #3: Private Speech #3 contains four datasets, each with 288 speech files rated by 16 or 18 subjects on the 5-level ACR scale. Each dataset has ratings from two labs, usually 8 subjects per dataset per lab. One lab instead contributed 10 subjects to one of these tests. This enables four lab-to-lab comparisons.

4) Private Video #1: The Private Video #1 contains two video-quality subjective tests. The first is a lab study where 15 experts rated 75 videos on the 100-level ACR scale. The second is a crowdsourcing test where 61 subjects rated 112 videos on a 100-point scale using the SSCQE method.

5) Private Video #2: The Private Video #2 contains 60 video sequences, 1 minute in duration, that were rated by 30 subjects on the 5-level ACR scale. This subjective test was conducted by OPTICOM (Germany).

6) Private Video #3: The Private Video #3 has been referred to as Netflix Quality Variation 2017 in presentations to VQEG. The 320 video sequences were rated on the 5-level ACR scale on two different monitors and rated on a 5-level SSCQE scale at a second lab. Due to the way the subjects and videos were divided, this study yields five datasets: four with 180 video sequences and 48 to 51 subjects (ACR), and one with 320 video sequences and 41 subjects (SSCQE). These datasets enable five method-to-method comparisons.

7) Private Image: Private Image is an image-quality study conducted with expert and naïve subjects using two different methods: DISSIM and FC. Each dataset contains 352 images rated by 8 experts and 9 naïve subjects. This enables two naïve-to-expert comparisons.

IV. Statistical Method Design Strategy

Our goal is to establish new statistical methods that let users understand the precision of objective metrics, relative to the subjective tests (the gold standard) and ad hoc evaluations (the de facto standard). The nature of subjective tests and metrics provides us with several challenges.

As Janowski and Pinson conclude in [41]: "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."

The nature of outliers is very different for subjective tests and metrics. Subjective tests assume that there is an underlying "true quality" for each media, and studies such as [41] support this assumption.[4] While MOS values differ from one test to another, we expect the ordering of impairments and relative distances to be replicable [10]. When we replicate a subjective test in two labs, we expect the residuals to have a Gaussian distribution.

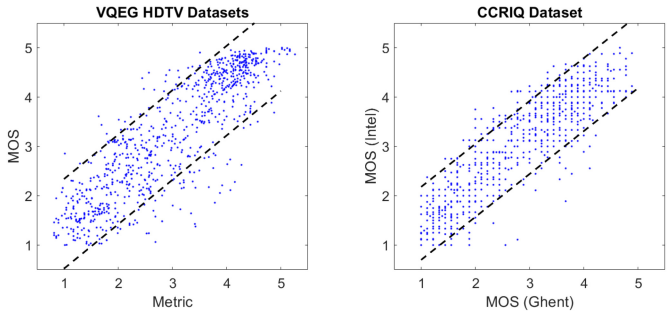

Fig. 1. Scatter plots compare a high performing metric to MOSs (left) and MOSs from two labs (Right). The dashed lines are at 1.5 × std(R), where R is the residuals of a linear fit. Notice the irregular distribution of data for the lab-to-metric comparison (left) and the orderly distribution of data for the lab-to-lab comparison (right).

By contrast, metrics are deterministic. When a metric fails to grasp the quality impact of an impairment, the residuals (between metric and MOS) will have an irregular distribution with clumps of data points and far-flung outliers (see Fig. 1). A statistical method that assumes a Gaussian distribution of residuals is biased in favor of subjective testing.

We want to estimate two factors: precision and repeatability. We will use CIs to measure precision. CIs are simple to use and can be applied identically to MOSs and metric values. We will use a confusion matrix to measure repeatability. CIs and confusion matrices let us re-frame the problem as "what is the likelihood of erroneous decisions."

Section II includes precedents for both choices and guidance on best practices for the statistical analysis of subjective and objective metrics. To ensure fair comparisons, we will use the same techniques previously described to analyze the subjective ratings and the objective metric data. To ensure reliable comparisons, our baseline performance of subjective tests will be based on measurements of nearly 100 datasets.

V. Confidence Interval of a Subjective Test (ΔSCI)

A. Calculating the Confidence Interval of a Subjective Test

We will define the *subjective test's confidence interval* (ΔS_{CI}) as the minimum difference in MOS at which 95% of the stimulus pairs will be statistically different. Fundamentally, the data we have to work with are individual subject ratings, pairs of stimuli (**A**, **B**), and the absolute value of the distance between the MOSs of **A** and **B**, which we will call $\Delta_{A,B}$. We have two choices when using this information to calculate the CI of a subjective test.

[4]Private discussions among subject matter experts indicate rare instances where a media's subject ratings have a bimodal distributions. Opinions on the "true quality" of a media genuinely seem to diverge.

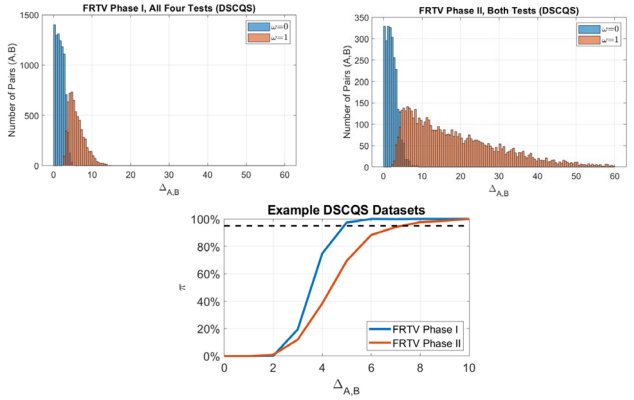

Fig. 2. For two DSCQS datasets, this figure shows the relationship between $\Delta_{A,B}$ and conclusions reached by the Student's t-test ($\boldsymbol{\omega}$). On the histograms, blue means that **A** and **B** are statistically identical ($\boldsymbol{\omega}$=0), orange means that **A** and **B** are statistically different ($\boldsymbol{\omega}$=1), and the overlap is brown. The line plot summarizes these histograms as a percent of stimulus pairs that are statistically different ($\boldsymbol{\pi}$). The rise from $\boldsymbol{\pi}$=0% to $\boldsymbol{\pi}$ =100% corresponds to the brown overlap on the histogram. The dashed line at $\boldsymbol{\pi}$=95% marks ΔS_{CI}. These ΔS_{CI} values are 0.5 and 0.7.

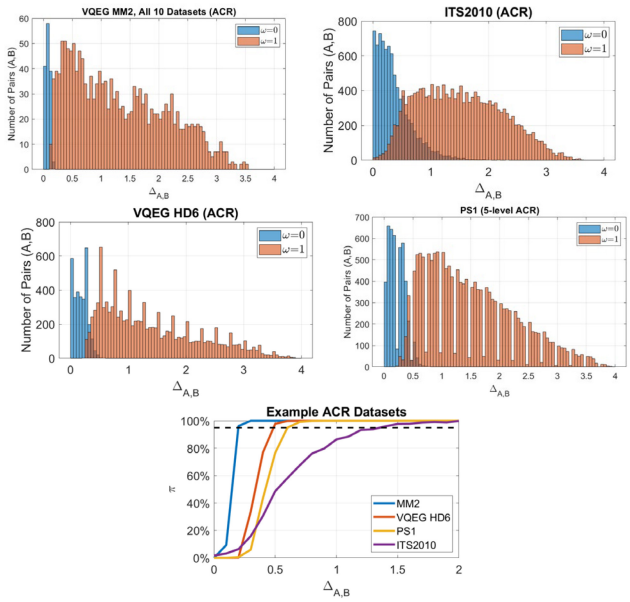

Fig. 3. For four ACR datasets, this figure shows the relationship between $\Delta_{A,B}$ and conclusions reached by the Student's t-test (ω). The ΔS_{CI} values are 0.2, 0.5, 0.6, and 1.3.

Our first choice is how to determine whether the quality of **A** and **B** are significantly different (e.g., Student's t-test, F-test, 1.96 $\times$ standard error). We will use the two-sample Student's t-test, which uses the individual subject ratings from a specific (**A**, **B**) pair to test the hypothesis that **A** and **B** have the same mean. MOSs have a normal distribution [41], so the requirements for the Student's t-test are met. The two population variances can be, and likely are, unequal.

Our second choice is how to aggregate across all pairs of stimuli in the dataset. We are concerned that an uneven distribution of $\Delta_{A,B}$ could impact aggregation statistics like mean and 95th percentile. For example, consider the fraction of stimulus pairs with $\Delta_{A,B}$ close to zero. The fraction will be larger if most of the dataset's media have MOSs near the top or bottom of the rating scale, where SOS is low (like *FRTV Phase I* and *II*); and this fraction will be smaller for datasets that have an even distribution of MOSs (like *VQEG MM2* and *VQEG HD6*).

We will divide the stimulus pairs (**A**, **B**) into subsets that have similar values for $\Delta_{A,B}$. This will let us reach conclusions based on localized data, without being overly influenced by the distribution of $\Delta_{A,B}$ across the entire dataset.

Let us explore the relationship between conclusions reached by the Student's t-test and $\Delta_{A,B}$. This allows us to calculate ΔS_{CI}, a new measure of the precision of a subjective test that is calculated as follows.

Given a subjective test, we will choose all pairs of stimuli (**A**, **B**), where both stimuli were rated by the same subjects and the stimuli are drawn from the same dataset. An occasional missing rating is acceptable. For each pair of stimuli, we will measure $\Delta_{A,B}$, the absolute value of the distance between the MOSs of **A** and **B**. Most of our datasets use the 5-level ACR method, where MOS ranges from 1 to 5 and $\Delta_{A,B}$ ranges from 0 to 4.

We will use the paired stimuli Student's t-test to compare the rating distributions for **A** and **B** at the 95% confidence level. We will record 1 if the conclusion is that **A** and **B** are significantly different and 0 otherwise. We will tally these comparisons in a new binary variable, $\boldsymbol{\omega}$.

Each dataset has few stimulus pairs, so we must bin $\Delta_{A,B}$ by MOS intervals. We will bin 5-level ACR data by 0.1 MOS intervals (0 $\pm$0.05, 0.1 $\pm$0.05, 0.2 $\pm$0.05, ...), 100-level DSCQS data by 0.5 MOS intervals, etc.

We must quantify the relationship between $\Delta_{A,B}$ and $\boldsymbol{\omega}$ as a threshold. Let us compute $\boldsymbol{\pi}$ as the average response of $\boldsymbol{\omega}$ for stimulus pairs in each bin. We will express $\boldsymbol{\pi}$ as a percentage (i.e., average the $\boldsymbol{\omega}$ responses and multiply by 100). Thus, $\boldsymbol{\pi}$ ranges from 0 to 100 where 0% means that all pairs of stimuli (**A**, **B**) in the bin have equivalent quality, and 100% indicates that all pairs of stimuli in the bin have significantly different quality (measured at the 95% confidence level). Finally, we will calculate ΔS_{CI} as the $\Delta_{A,B}$ that comes closest to producing $\boldsymbol{\pi}$ = 95%. ΔS_{CI} is the value of $\Delta_{A,B}$ where the $\boldsymbol{\pi}$ curve crosses the 95% threshold.

Fig. 2 and Fig. 3 show this data presented as histograms and line plots ($\boldsymbol{\pi}$ as a function of $\Delta_{A,B}$). Fig. 2 shows two DSCQS datasets, and Fig. 3 shows four ACR datasets. The histograms are similar to those used in [16]. In Fig. 2, we see that the VQEG FRTV Phase I tests (top left) has a much narrower range of quality than the VQEG FRTV Phase II tests (top right). In Fig. 3, all ten datasets of the VQEG MM2 study are aggregated (top left) and the large number of subjects causes $\boldsymbol{\omega}$=0 to span a very narrow range. The ITS2010 study (top right) has an unusually large overlap between the $\boldsymbol{\omega}$=0 and $\boldsymbol{\omega}$=1 curves, probably because the stimuli include very different levels of audio and video quality. The VQEG HD6 and PS1 histograms (middle) have an uneven distribution of $\Delta_{A,B}$. This pattern would disappear if we used larger bins.

B. ΔSCI Observed for 5-Level ACR Datasets

Although this technique could be used for any rating scale, most of our datasets were conducted with 5-level ACR. Table I

TABLE I
SUMMARY OF 5-LEVEL ACR DATASETS FOR ΔS_{CI}

Video Study	**Dataset Size (Files, Subjects)**
AGH/NTIA/Dolby	(230, 71), *(230, 31), (230, 22), (230, 18)*
ITS4S	(813, 27), (212, 24)
ITS4S3	(99, 14), (99, 17), (99, 14), (99, 15), (99, 13), (99,19)
ITS4S4	(196, 26)
Public Safety #1	(400, 16)
Public Safety #2	(576, 19)
Private Video #2	(60, 30)
Private Video #3	(180, 51), (180, 51) (180, 48), (180,50)
SJTU 4K	(30, 28)
UPM-Acreo	(132,20)
VQEG HDTV	(168, 24), (168, 24), (152, 24), (168, 24), (168, 24), (168, 24),
VQEG Hybrid	(184, 24), (184, 24), (184, 24), (184, 24), (184, 24), (114, 24), (194, 24), (184, 24), (194, 24), (120, 24)
VQEG RRNR-TV	(168, 32), (168, 31)
Image Study	**Dataset Size (Files, Subjects)**
CCRIQ	(221, 27), (171, 26), *(171,9), (171,9), (171,9), (221, 9), (221, 9), (221, 8)*
CCRIQ2 & VIME1	(88,19), (101,21)
ITS4S2	(1429,16)
Audiovisual Study	**Dataset Size (Files, Subjects)**
ITS 2010	(240,26)
ITS AV-Sync 2010	(297,12), (297,16)
VQEG MM2	(60,213), *(60,28), (60,9), (60,34), (60,25), (60,25), (60,24), (60,24), (60,14), (60,15), (60,15)*
Speech Study	**Dataset Size (Files, Subjects)**
ITU-T P.Sup23	(176, 24), (176, 24), (176, 24), (200, 24), (200, 24), (200, 24), (200, 24)
Private Speech #1	(1278,≈11), (33,≈22), (48,43)
Private Speech #2	(2432,8)
Private Speech #3	(288,10), (288,8), (288,8), (288,8), (288,8), (288,8), (288,8), (288,8)
401 (crowdsource)	(192, 227)
501 (crowdsource)	(96,115)
701 (crowdsource)	(128,144)
Emulation	**Dataset Size (Files, Subjects)**
SDO Emulation	(2718,15) (2718,9), (2718,6)

lists the 91 datasets that will be used to calculate ΔS_{CI} for a typical 5-level ACR test. The first column identifies the media type and a name attributed to the entire study. The second column divides the study into datasets, presented as the number of files and number of subjects in parentheses. In a few cases, subjects from multiple labs allow us to estimate each lab's CI (presented in italics) and to estimate the overall CI (presented in regular font). Details of the "SDO Emulation" dataset will be presented below.

C. Analysis

Fig. 4 shows the resulting relationship between ΔS_{CI} and the number of subjects in a 5-level ACR dataset. The area of the dot increases linearly with the number of datasets that produce this result.

Table II shows the observed trend connecting ΔS_{CI} and the number of subjects. The bottom row shows the range of ΔS_{CI} observed for the 16 SDO datasets (with all 24 subjects) and extrapolated ΔS_{CI} for fewer subjects. These points are near

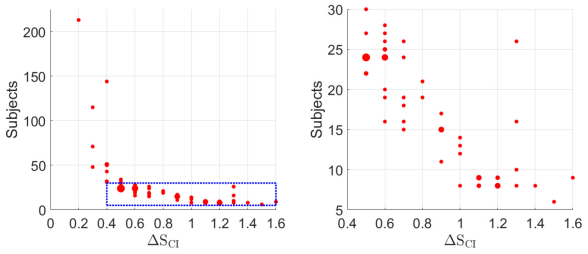

Fig. 4. Relationship between ΔS_{CI} and number of subjects in a 5-level ACR dataset, presented as a 2-D histogram. The right side enlarges the region outlined in a blue box on the left.

TABLE II
EXPECTED CONFIDENCE INTERVAL FOR A 5-LEVEL ACR TEST

Number of Subjects	24	15	9	6
Typical ΔS_{CI}	0.5 to 0.7	0.7 to 1.0	1.0 to 1.4	≥ 1.5
SDO baseline ΔS_{CI}	0.5 to 0.6	0.7	1.1	1.5

the lower edge of the response curve in Fig. 4. We cannot accurately predict trends for more than 24 subjects, because datasets rarely have more than 30 subjects. $\Delta S_{CI} \leq 0.3$ is very difficult to achieve, and the number of subjects would need to be increased dramatically above numbers typically used today.

We divided the tests into sub-sets based on test type (SDO, lab environment, field tests, and tests in a public environment) and stimulus type (video, audiovisual, image, and speech). The following factors yielded data across the entire response curve with no apparent bias: lab studies, field studies, video stimuli, and image stimuli. Ten or fewer subjects seem to have too much noise to observe trends based on test type or stimulus type.

The following factors impacted the range of ΔS_{CI}. The SDO and speech datasets had lower values for ΔS_{CI}. The former can be explained by the extra oversight: perhaps 20 to 50 experts contribute to these experiment designs. The latter can be explained by relative homogeneity of speech, which uses a limited set of phonemes. The audiovisual datasets had higher values for ΔS_{CI} and included the two outliers at $\Delta S_{CI} = 1.3$. This can be explained by stimuli where the audio quality and video quality were unrelated.

The relationship in Table II appears to be a characteristic of the 5-level ACR method itself, not a characteristic of the media or the test environment. The bottom and top of the ΔS_{CI} range indicate the most favorable and least favorable performances of well-designed and carefully conducted subjective tests. Higher values within this range do not indicate that something is wrong. The analyses above imply that complex impairments contribute to higher ΔS_{CI}. For example, the background quality could be poor while the foreground quality is good.

The number of subjects and ΔS_{CI} in Table II are calculated after removing outliers. Some researchers eliminate subjects with noisy data, typically by applying somewhat arbitrary thresholds. Other researchers retain subjects with noisy data, since rating noise occurs randomly and is unlikely to indicate poor behavior on the part of the subject. Outlier data are rarely distributed, so we could not study this variable.

TABLE III
CONFUSION MATRIX BETWEEN TWO SUBJECTIVE TEST LABS

		Lab 1		
		Better	**Equivalent**	**Worse**
Lab 2	**Better**	Agree Ranking	Unconfirmed	Disagree
	Equivalent	Unconfirmed	Agree Tie	Unconfirmed
	Worse	Disagree	Unconfirmed	Agree Ranking

VI. SUBJECTIVE TEST DECISION COMPARISONS

A. Comparing Decisions Reached by Two Subjective Tests

Now that we understand the precision of subjective tests, let us examine their repeatability. We will use confusion matrices to compare conclusions reached by subjective tests. We will denote the MOS for stimuli **A** and **B** as MOS_A and MOS_B, respectively.

When a subjective test is repeated in multiple labs, we observe different MOSs and different rating distributions for the same stimuli. This allows us to calculate the repeatability of subjective testing, based on the frequency with which the same or different conclusions are reached by different labs.

Given a subjective test conducted identically at two or more different labs, we will choose all pairs of stimuli, **A** and **B**, where both stimuli were rated by the same subjects and the stimuli are drawn from the same dataset. An occasional missing rating is acceptable. We will use the paired stimuli Student's *t*-test to compare the rating distributions for **A** and **B** at the 95% confidence level. For each lab's subjects, we will decide whether **A** is better than, equivalent to, or worse than **B**. We will perform this procedure for all 88 lab-to-lab comparisons, 22 method-to-method comparisons, and 12 comparisons of expert and naïve subjects.

Table III shows the confusion matrix when two labs perform the same subjective test. Each lab uses the Student's *t*-test to determine whether stimulus **A** is better than, equivalent to, or worse than stimulus **B**.

We cannot trust one lab's results more than the other lab's results, so the outcomes are phrased in terms of agreement and disagreement: *agree ranking*, *agree tie*, *unconfirmed*, and *disagree*. An *agree ranking* incident occurs when both labs conclude that **A** is statistically better than **B**. An *agree tie* incident occurs when both labs conclude that **A** and **B** have the same quality. An *unconfirmed* incident occurs when one lab concludes that **A** is statistically different than **B**, but the other lab concludes that **A** and **B** are statistically equivalent. A *disagree* incidence occurs when one lab concluded that **A** is statistically *better* than **B**, but the other lab concluded that **A** is statistically *worse* than **B**.

B. Observed Incidence Rates

Table IV lists datasets where individual subject ratings are available from two or more labs, with other aspects of the subjective test held constant. We will use these datasets to establish a baseline for the repeatability of subjective tests. The columns identify the name of the subjective test, the subjective test method, the number of media, and the number of subjects in each lab, presented as pairs of numbers in parentheses. Each pair of numbers in parentheses enables a single comparison.

TABLE IV
SUMMARY OF LAB-TO-LAB COMPARISONS

Study	**Method**	**Media**	**Subjects (Lab 1, Lab 2)**
AGH/NTIA/Dolby	ACR	230	(31, 22), (31,18), (22,18)
CCRIQ	ACR	171	(18, 18), (18, 17), (18, 17)
		212	(18, 18), (18, 17), (18, 17)
ITS4S	ACR	212	(27, 24)
RRNR-TV	ACR	168	(16, 16), (16, 15)
VQEG FTTV Phase II	DSCQS	64	(32, 32), (32, 32)
VQEG Hybrid HD	ACR	24 (×10)	(24, 24)
VQEG MM2	ACR	60	(28,34), (28,25), (28,25), (34,25), (34,25), (25,25), (28,24), (28,24), (34,24), (34,24), (25,24), (25,24), (25,24), (25,24), (24,24), (28,15), (28,15), (34,15), (34,15), (25,15), (25,15), (25,15), (24,15), (24,15), (24,15), (24,15), (15,15), (14,15), (14,15), (28,14), (34,14), (25,14), (25,14), (24,14), (24,14), (9,34), (9,25), (9,25), (9,24), (9,24), (9,15), (9,15), (9,14), (28,9)
Private Speech #3	ACR	288	(10,8), (8,8), (8,8), (8,8)
VQEG FRTV Phase I	DSCQS	90	(18,18), (18,18), (18,18), (18,18), (18,18), (18,18), (18,18), (18,17), (18,18), (18,18), (18,16), (18,16), (18,16), (18,16), (18,16), (18,16), (18,16), (18,16), (18,16), (18,16), (17,16), (17,16), (17,16), (16,16)

Table V lists datasets where individual subject ratings are available for two or more rating methods. Table VI lists datasets where individual subject ratings are available for expert subjects and naïve subjects. We will use these datasets to understand how repeatability degrades when subjects may genuinely disagree, due to differences in presentation method or task.

We will tally the frequency of the four possible classification types in the Table III confusion matrix for the lab-to-lab comparisons listed in Table IV. We will split the lab-to-lab comparisons into two categories: narrow range of quality (from the VQEG FRTV Phase I datasets) and wide range of quality. The resulting incidence rates appear in Table VII. Similarly, Table VIII shows method-to-method comparisons (from Table V) and expert-to-naïve comparisons (from Table VI).

An outlier at 1.82% is omitted from Table VII to avoid skewing the statistics. This outlier may be caused either by scoring inversions (see Section III-B) or known differences in how the labs implemented DSCQS. When this outlier is included, the overall average lab-to-lab *disagree* incidence rate rises to 0.32%.

C. Disagree Incidence Rates

Table VII indicates that the *disagree* incidence rate does not depend upon the distribution of MOSs in the dataset.

TABLE V
SUMMARY OF METHOD-TO-METHOD COMPARISONS

Study	Factor	Media	Subjects
Private Video #3	Monitor	180	(51,51), (50,48)
Private Video #3	Rating method	180	(51,41), (50,41), (51,41), (48,41)
UPM-Acreo	Rating method	132	(21,22), (21,20), (22,20)
Private Video #3	Monitor	180	(25,25), (25,25)
Private Video #3	Monitor	180	(26,26), (23,25)
Private Video #3	Rating method	180	(25,26), (25,23), (25,26), (25,25)
Public Safety #1	Rating method	400	(16,16)
Public Safety #2	Rating method	576	(19,19)
Ghent Medical	Rating method	70	(10,16)
CCRIQ	Monitor	171 221	(53, 53) (53, 53)

TABLE VI
SUMMARY OF EXPERT VS NAÏVE SUBJECT

Study	Expertise	Media	Subjects
Ghent Denoising	Image and video processing	70	(18,20), (18,20), (19,18)
Ghent Medical	Laparoscopic surgeons	20	(9,16)
		70	(10,16), (10,16), (10,17)
Private Image	Industry employees	352	(8,9), (8,9)
ITS4S3	First responders	99	(7,7), (6,11), (10,8)

TABLE VII
LAB-TO-LAB CLASSIFICATION INCIDENCE RATES

	Wide Range of Quality		Narrow Range of Quality	
Outcome	**Mean**	**Range**	**Mean**	**Range**
Agree Ranking	64%	47% to 77%	46%	24% to 65%
Agree Tie	17%	10% to 29%	28%	17% to 48%
Unconfirmed	19%	10% to 31%	26%	19% to 38%
Disagree	0.15%	0% to 0.94%	0.25%	0% to 0.91%

TABLE VIII
METHOD-TO-METHOD AND EXPERT-TO-NAÏVE CLASSIFICATION INCIDENCE RATES

	Method-to-Method		Expert-to-Naïve	
Outcome	**Mean**	**Range**	**Mean**	**Range**
Agree Ranking	68%	44% to 82%	22%	3% to 40%
Agree Tie	13%	4% to 23%	47%	13% to 83%
Unconfirmed	17%	9% to 31%	32%	14% to 50%
Disagree	2.06%	0.0% to 7.78%	0.16%	0% to 0.53%

This matches our expectations that the *disagree* incidence rate depends on the experiment design (e.g., rating method, subject demographics, and the system used to reproduce the media). The two largest *disagree* incidence rates are for the method-to-method comparisons for the PS1 and PS2 datasets, where the 5-level ACR method is compared to the experimental Boolean method.

Fig. 5 shows a histogram with the *disagree* rates for all three types of comparisons combined. The distribution of *disagree* incidence rates has a main cluster below 0.31% with a long tail of higher values. The *disagree* incidence rate is very small for all lab-to-lab comparisons (except for the outlier), as well as most of the method-to-method and expert-to-naïve subject comparisons.

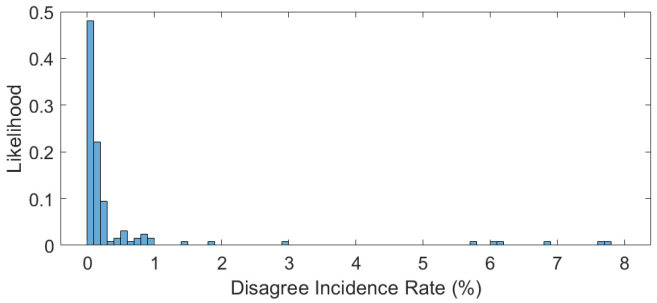

Fig. 5. Histogram of *disagree* incidence rates for lab-to-lab, method-to-method, and expert-to-naïve subjective test comparisons.

We saw no evidence that experts and naïve subjects rank order of stimuli differently, but experts may be more sensitive to quality differences. The *unconfirmed* incidence rates for expert-to-naïve subjects are unusually higher than the *unconfirmed* incidence rates for lab-to-lab and method-to-method comparisons. The root cause is differences in likelihood that experts and naïve subjects will conclude that **A** and **B** are equivalent. For the Ghent Medical and Ghent Denoising datasets, the *equivalent* incidence rate is 20% to 40% lower for expert subjects than naïve subjects. For the ITS4S3 dataset, the *equivalent* incidence rates are roughly the same for expert and naïve subjects—but the naïve subjects worked in related fields and had some understanding of first responder needs.

We conclude that *disagree* incidence rates above 0.31% are unusual enough to warrant investigation and disagree incidence rates above 1.0% indicate a method-to-method difference or lab-to-lab difference.

D. Applying the Disagree Thresholds

Disagree incidence rates above 1.0% occur for one lab-to-lab comparison and seven method-to-method comparisons. The lab-to-lab comparison outlier was noted earlier. *UPM-Acreo* indicates differences between the ACR (no audio) and SSCQE (with audio) methods. *PS1* and *PS2* indicate differences between the ACR and Boolean methods. The remaining four comparisons (from *Private Video #3*) indicate differences between the ACR and SSCQE methods. This matches our expectations that the test method has a major impact on conclusions.

The *disagree* incidence rate can only detect patterns that impact the relative ranking of stimuli. The two *CCRIQ* method-to-method comparisons produced 0.0% and 0.02% *disagree* incidence rates. This means the HD and 4K monitors agree on the quality ranking of the 24 consumer cameras in this study. However, analyses of *CCRIQ* in [34] indicate ≈0.2 differences between the MOSs of HD and 4K monitors for high quality images and no significant difference for low quality images. This means 4K monitors are slightly better than HD monitors—a phenomenon the *disagree* incidence rate cannot detect.

TABLE IX
CONFUSION MATRIX BETWEEN A SUBJECTIVE TEST AND AN AD HOC EVALUATION OR PILOT TEST

		Subjective Test		
		Better	**Equivalent**	**Worse**
Ad hoc Evaluation or Pilot Test	**Better**	Correct Ranking	False Distinction	False Ranking
	Worse	False Ranking	False Distinction	Correct Ranking

VII. AD HOC EVALUATIONS AND PILOT TESTS

A. Comparing Decisions Reached by Ad Hoc Evaluations and Subjective Tests

Ad hoc evaluations dispense with the scientific method in favor of convenience. For example, a management team encodes select test sequences with hardware video encoders on loan from competing vendors, discusses the quality differences, and makes a purchase decision. A bit of subjective testing knowledge lets us add a minimum of structure and formality. For example, two or three researchers watch or listen to each stimulus in a proposed experiment and write down their ratings.

Pilot tests are similar to ad hoc evaluations, from a statistical analysis standpoint, in that MOSs are compared deterministically. ITU-T Rec. P.913 recommends eight to twelve subjects for pilot studies to indicate trending. Section V indicates that ΔS_{CI} for nine subjects will range from 1.0 to 1.4, so the Student's *t*-test is unlikely to reach any interesting conclusions.

Table IX shows the confusion matrix when a pilot test or ad hoc evaluation is compared to a subjective test. The subjective test uses the Student's *t*-test, as above. Ad hoc evaluations and pilot tests produce MOSs with low precision that must be compared deterministically (>, <, =). We will ignore ties since our ad hoc or pilot test data seldom produce identical MOSs. We will evaluate the ad hoc evaluations and pilot tests relative to the subjective test's more accurate assessments, so the outcomes are phrased in terms of correctness: *correct ranking*, *false distinction*, and *false ranking*.

B. Simulating Ad Hoc and Pilot Test Ratings

We want to establish the relationship between the decisions reached by an ad hoc evaluation and the decisions reached by a formal subjective test. The ad hoc subjects are likely to have an over-inflated sense of the accuracy of their judgements and to make an occasional error due to miscommunication. The interesting questions probably span a narrow range of quality.

We will simulate likely behaviors using the VQEG FRTV Phase I datasets. These four datasets have a narrow range of quality, continuous scale (from 0 to 100), and the occasional rating error. These characteristics seem appropriate and realistic for ad hoc decisions. Each FRTV Phase I dataset has 16 to 18 ratings from four different labs. This will allow us to assign one lab the role of ad hoc evaluation (or pilot test) and the other labs their actual role of a formal subjective test.

TABLE X
ESTIMATED FALSE RANKING RATES FOR AD HOC & PILOT TESTS

Subjects	**1**	**2**	**3**	**6**	**9**	**12**
Mean	11.4%	8.5%	6.8%	4.4%	3.5%	3.0%
Minimum	3%	2%	1%	1%	1%	0%
Maximum	30%	26%	21%	17%	13%	10%

TABLE XI
ESTIMATED INCIDENCE RATES FOR AD HOC & PILOT TESTS THAT SPAN A NARROW RANGE OF QUALITY

Subjects	**1**	**2**	**3**	**6**	**9**	**12**
Correct Ranking	36%	40%	41%	44%	45%	46%
False Distinction	49%	51%	51%	51%	51%	51%
False Ranking	12%	9%	7%	5%	4%	3%

C. Error Incidence Rates

To frame discussions, we will establish six performance levels: ad hoc evaluation with one, two, or three people; and pilot tests with six, nine, or twelve subjects. The former may consist of nothing more than verbal discussion, while the latter are presumed to follow the standard methods in an ITU Recommendation.

Let us now compare conclusions reached by the subjective test with the ad hoc evaluation or pilot test (see Table IX). We will simulate a typically subjective test by drawing 24 subjects at random from three labs. All possible stimulus pairs will be compared using the Student's *t*-test. To simulate the ad hoc evaluation or pilot test, we will draw 1, 2, 3, 6, 9, or 12 subjects at random from the fourth lab; these MOSs will be compared deterministically.

Table X lists the *false ranking* incidence rates, computed on all four FRTV Phase I tests. As mentioned above, the *false ranking* incidence rates are not impacted by the range of quality in the test.

Table XI lists the average incidence rates across all trials, using the dataset that has the narrowest range of quality (39% of the 100-level scale). In ACR scale language, this test covers from "excellent" to part way between "good" and "fair." This represents the worst-case scenario: a critical business decision about similar video systems.

Compare Table VII and Table VIII with Table X and Table XI to understand the quantitative superiority of subjective tests over ad hoc evaluations. Essentially, one person can use a single media to identify the higher-quality system, but a subjective test would only support 36% of their conclusions, and about 12% of their conclusions will be erroneous (i.e., chose the lower-quality system). Private communications with industry anecdotally support the high error rates of ad hoc evaluations.

With a subjective test, the odds of choosing the lower-quality system drops to 0.15% on average, with 0.31% as the expected worst case. Moreover, statistical methods can be used to aggregate media MOSs into system MOSs, which increases the likelihood of choosing the best system. This is not possible with ad hoc evaluations.

TABLE XII
CONFUSION MATRIX BETWEEN SUBJECTIVE TEST AND METRIC DECISIONS MADE USING DETERMINISTIC MATH

		Subjective Test (MOS)		
		Better	**Equivalent**	**Worse**
Metric ($\widehat{MOS}$)	**Better**	Correct Ranking	False Distinction	False Ranking
	Worse	False Ranking	False Distinction	Correct Ranking

TABLE XIII
RANGE OF FALSE RANKING RATES WHEN EQUATING A QUALITY METRIC TO A NUMBER OF PEOPLE IN AN AD HOC EVALUATION OR PILOT TEST

PVQT	Meaning	Metric's False Ranking
1 PVQT	1-person ad hoc evaluation	13% > *false ranking* ≥ 10%
2 PVQT	2-person ad hoc evaluation	10% > *false ranking* ≥ 8%
3 PVQT	3-person ad hoc evaluation	8% > *false ranking* ≥ 6%
6 PVQT	6-person pilot test	6% > *false ranking* ≥ 4%
9 PVQT	9-person pilot test	4% > *false ranking*

VIII. COMPARING QUALITY METRICS WITH VIDEO-QUALITY SUBJECTIVE TESTS

We want to compare the conclusions reached by a metric with the conclusions reached by people. Quality metrics can be considered as substitutes—or proxies—for subjective quality ratings. For this reason, we denote the metric value for a certain stimulus, **A,** as $\widehat{MOS}_A$.

We will begin with the simplest case where the user makes decisions based on deterministic comparisons between $\widehat{MOS}_A$ and $\widehat{MOS}_B$. Like an ad hoc evaluation or pilot test in Section VIII, any increase or decrease in metric value is significant.

A. Comparing Deterministic Decisions Reached by Metrics and Subjective Tests

We will compare conclusions reached by the metric with conclusions reached by a subjective test. As noted earlier, we must avoid predisposing our measurement sensitivity in favor of subjective testing. Thus, when comparing MOSs, we will use the CI of a 24-subject test conducted by an SDO ($\Delta S_{CI} = 0.5$) from Table II.

Table XII contains a confusion matrix that describes the possible outcomes when a metric is compared to a subjective test. Like the confusion matrix in Table IX, CIs are used for MOS comparisons, but $\widehat{MOS}$ comparisons are made deterministically. Outcomes are phrased in terms of correctness: *correct ranking*, *false distinction*, and *false ranking*.

We will ignore the possibility of identical metric values, which would impose a "Metric Equivalent" row, as per Table III. The *metric equivalent* incidence primarily occurs if the metric clips values at a maximum (or minimum) value. That data must be discarded err it distort our measurements. Metric equivalence scarcely occurs otherwise, since computations typically use double-precision floating-point numbers.

B. Equating a Metric to a Number of People

We want to liken the metric to a number of people in video-quality test (PVQT). We will use the statistics from ad hoc evaluations, because people make direct comparisons between metric values (without CIs). PVQT will let us make simple statements like, "This metric is analogous to two people monitoring the quality of your video stream." The analogy assumes the use of CIs to compare MOSs and deterministic math to compare $\widehat{MOS}$s.

The confusion matrix is presented in Table XII. Due to the similarity between Table IX and Table XII, we can compare the *false distinction* incidence rate of an ad hoc evaluation and a metric. If the *false distinction* rates are similar, we can conclude that the metric is behaving similarly to an ad hoc assessment. We will not constrain *correct ranking* and *false distinction* because we do not have a figure of merit to remove the confounding factor (range of quality examined). *False ranking* is the most egregious type of error, and thus the most important to constrain.

The *false ranking* rates in Table X overlap. An argument could be made for choosing 30% as the maximum *false ranking* incidence rate for one PVQT. However, that low level of performance would make the "one person" analogy meaningless.

The *false ranking* incidence rates get closer together as the number of subjects increases. The average *false ranking* rates for 9 and 12 subjects only differ by 0.5%. This is probably within the measurement uncertainty of the Table X estimates. Therefore, we will limit the upper end of PVQT at nine people.

Table XIII identifies the range of a metric *false ranking* incidence rates that are associated with 1, 2, 3, 6, and 9 people. We rounded these thresholds to the nearest integer, for ease of use. This rounding also reflects the limited accuracy of our Table X estimates. The lower threshold (larger value) is included in the range, while the upper threshold is excluded.

C. Applied Results

Let us begin with MOSs and metrics evaluated by VQEG during Multimedia [26] validation tests. These validation tests provide a robust set of well-designed subjective datasets to calculate equivalence to a number of people. All eight metrics except PSNR are referred to by randomly assigned letters (A to H) and linearly mapped onto a 1 to 5 scale. The VGA metrics were analyzed against thirteen datasets: twelve with 166 videos and one with 142 videos. Decisions reached by each dataset's stimulus pairs are pooled into a single tally of incidence rates.

Table XIV shows the metric letter, correlation between the metric and MOS (when stimuli from all thirteen datasets are pooled), PVQT, and the incidence rates. Metrics E and C are less accurate than a single person, with 19% and 22% *false ranking,* respectively.

Equating a metric to a number of people in a video-quality test will help naïve users easily understand the effective precision of a metric. However, the discrete nature of these values and the absolute thresholds make this method unsuited for comparing accuracy between metrics. A small change in *false ranking* incidence rates could produce a large change in PVQT.

TABLE XIV
PVQT FOR VQEG MULTIMEDIA VALIDATION TEST METRICS

Metric	Correlation	PVQT	Correct Ranking	False Ranking	False Distinction
H	85%	3	67%	6%	27%
A	84%	3	66%	7%	27%
B	81%	3	66%	7%	27%
G	79%	3	66%	7%	27%
F	81%	3	67%	7%	27%
PSNR	74%	1	63%	11%	27%
D	75%	1	62%	11%	27%
E	50%	NA	55%	19%	27%
C	39%	NA	51%	22%	27%

TABLE XV
CONFUSION MATRIX BETWEEN A SUBJECTIVE TEST AND METRIC DECISIONS MADE USING CIS

		Subjective Test (MOS)		
		Better	**Equivalent**	**Worse**
Metric ($\widehat{MOS}$)	**Better**	Correct Ranking	False Distinction	False Ranking
	Equivalent	False Tie	Correct Tie	False Tie
	Worse	False Ranking	False Distinction	Correct Ranking

The drawback with likening a metric to a number of people in a video-quality test is that any change in $\widehat{MOS}$ is treated as significant. We know this is false. $\widehat{MOS}$s are imprecise, and we would like to understand their precision.

We cannot recommend *false ranking* for comparisons between metrics. However, like a scatter plot, *false ranking* can help metric researchers understand the behavior of a single metric.

IX. CONFIDENCE INTERVALS FOR METRICS

We would like to compare the conclusions reached by a metric to the conclusions reached by a subjective test when all comparisons use CIs. This will let us compute a CI for the metric, such that the metric's error rate will not exceed the error rate of a subjective test when this CI is used to make decisions. We can also test whether all of the metric's decision incidence rates are equivalent to a subjective test's decision incidence rates, when this CI is used to make decisions.

A. Metric's Confidence Interval (ΔMCI)

Table XV shows the confusion matrix when a quality metric is compared to a subjective test. CIs are used to determine the significance of MOS comparisons and $\widehat{MOS}$ comparisons. The subjective test provides our ground truth, so the outcomes of the metric are phrased in terms of correctness: *correct ranking*, *correct tie*, *false tie*, *false distinction*, and *false ranking*. Note the similarities between Table XV and Table III. We will tally the frequency of the four possible classification types in the Table XV confusion matrix.

Philosophically, we want a metric confidence interval that can be used to make decisions that yield error rates similar those seen in the lab-to-lab comparisons. From most egregious to least egregious, these error categories are *false ranking*, *false distinction*, and *false tie*.

The *disagree* category (in Table III) corresponds to the *false ranking* category (in Table XV). Our observations in Section VI indicate that the *disagree* incidence rate should be $\leq 1\%$. This is the maximum observed *disagree* incidence rate, based on all lab-to-lab comparisons that had no known problems with implementation.

The *unconfirmed* category (in Table III) corresponds to the *false tie* and the *false distinction* categories (in Table XV). In terms of incidence rates:

$$unconfirmed = false\ tie + false\ distinction \tag{1}$$

Thus, we will limit the *false distinction* incidence rate to half of the *unconfirmed* incidence rate. The *unconfirmed* incidence rate depends on the range of quality in the subjective test. Since the metric is likely to be applied to various media, we will use the "wide range of quality" statistics from Table VII. The *unconfirmed* incidence rate should be $\leq 31\%$, so our threshold will be 15.5% (i.e., 31%/2).

We will ignore the *false tie* category, which is arguably the least offensive type of error a metric can make. Pragmatically, we lack defensible limits. To limit *false ranking* and *false discrimination* incidence rates, we must allow the *false tie* incidence rate to increase.

We could place separate limits on the *false ranking* and *false distinction* incidence rates. That solution has two problems. First, metrics have a higher relative rate of *false ranking* compared to *false distinction*, due to imperfect modeling of human perception. So, in practice, the CI would only depend on the *false ranking* rate. Second, the CI would be too large for most applications. Users are willing to tolerate a low level of *false ranking* incidents to improve the *correct ranking* incidence rate.

Instead, we will define the overall *metric error* incidence rate to be the sum of *false ranking* and *false distinction* incidence rates. This will allow *false ranking* to become a larger proportion of *metric error* incidence rate.

Let us calculate the ΔM as:

$$\Delta M = \widehat{MOS}_A - \widehat{MOS}_B. \tag{2}$$

We then calculate the *metric's confidence interval,* ΔM_{CI}, as the value of ΔM where the *metric error* incidence rate is equal to the sum of *false ranking* (comparable with 1% *disagree*) and *false distinction* (comparable with 15.5% *unconfirmed*), or 16.5%.

B. Agree Ranking, Agree Tie, and Figure of Merit Concur

When we use ΔM_{CI} to make decisions, the *false ranking* and *false distinction* incidence rates are, by definition, constrained to fall within subjective testing incidence rates. To demonstrate equivalence to a subjective test, we must also ensure that *agree ranking*, *agree tie*, and *false ranking* fall within the expected range.

As we would expect, the *agree ranking*, *agree tie*, and *unconfirmed* incidence rates in Table VII depend on the range of quality in the subjective tests. In general, when the range of quality is smaller, *agree ranking* is reduced and *agree tie* becomes greater, consistent with intuition. Setting expectations

for *agree ranking* and *agree tie* is thus confounded by the spread of quality in the test.

To gain independence from the spread of quality, we observe that spread drives a strong and reliable trade-off between *agree ranking* and *agree tie*, expressed as fractions. This trade-off is described by:

$$\widehat{r} = (-1.2 \times \textit{agree tie} + 1.0)^2 \tag{3}$$

where *agree ranking* and *agree tie* are expressed as fractions, r is the measured *agree ranking*, and $\widehat{r}$ describes the predicted *agree ranking* as a function of *agree tie*. We computed $\widehat{r}$ using linear regression (see the red line in Fig. 6, left). A square root is needed to remove a non-linearity.

The relationship in (3) explains our observed measurements and establishes the relationship between *agree_tie* and *agree_ranking*. Next, given a comparison between two unforeseen tests, we must assess whether their incidence rates fall within expectations. That is, if we measure *agree_tie* and *agree_ranking* and add this point to our plot, would it lie within the scatter around the red fit line. Motivated by (1), we now define a new statistic *concur* as:

$$\begin{aligned} \textit{concur} &= \sqrt{\textit{agree ranking}} - \sqrt{\widehat{r}} \\ &= \sqrt{\textit{agree ranking}} + 1.2 \times \textit{agree tie} \end{aligned} \tag{4}$$

where *concur* measures the residual between $\widehat{r}$ and *agree_ranking*.

Concur directly measures whether the observed trends within our subjective datasets reoccur in another subjective test or objective metric. Note that *concur* takes a value of 1.0 when the approximation in (3) is exact, and it deviates about 1.0 for our data. *Concur* ranges from 0.91 to 1.05 and histograms of *concur* are shown in Fig. 6. The range of values (0.91 to 1.05) was calculated empirically. This mathematical function of *agree ranking* and *agree tie* allows us to remove the influence of the dataset's quality range. This is demonstrated by the fact that the two classes of subjective tests largely overlap in Fig. 6.

Concur is a single figure-of-merit for comparing the results of two tests (subjective-to-subjective, or subjective-to-metric). Larger values of *concur* indicate higher levels of agreement. The downside of this convenience is that *concur* is a bit more abstract than *agree ranking* or *agree tie.* Also note that *concur* is unitless.

C. Showing Equivalence to Subjective Testing

We would like to determine whether a metric's decision incidence rates are equivalent to a subjective test's decision incidence rates, when ΔM_{CI} is used to make decisions. We will use the figure-of-merit *concur*, as defined above, and set a threshold that *concur* must be no less than 0.91. Thus, we define a metric to be equivalent to a subjective test when it produces *concur* $\geq$ 0.91.

We will not place limits on the final category, *false tie*, for three reasons. We lack defensible limits; the *false tie* rate is arguably the least offensive type of error a metric can make; and *false tie* incidence rates are inherently limited by the other four factors.

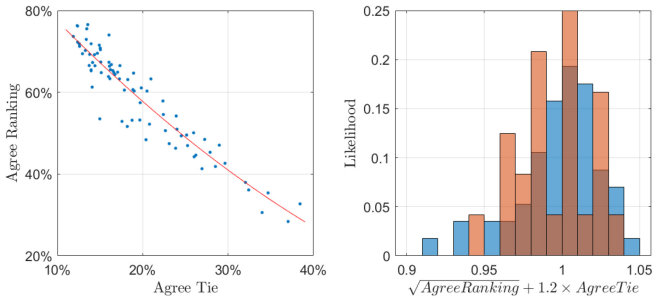

Fig. 6. The scatter plot (left) shows the relationship between *agree ranking* and *agree tie* incidence rates, with (1) plotted in red. The histogram (right) shows the distribution of (2) for wide range of quality tests in blue and tests with narrow range of quality tests in orange. The overlap is brown.

TABLE XVI
AD HOC TEST EQUIVALENCE FOR VQEG MULTIMEDIA VALIDATION TEST METRICS

Metric	Correlation	ΔM_{CI}	EVQT	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie	Concur
H	0.85	0.48	Yes	59%	2%	14%	12%	13%	0.925
A	0.84	0.52	Yes	57%	3%	13%	14%	14%	0.915
B	0.81	0.40	Yes	56%	2%	13%	15%	14%	0.910
F	0.81	0.36	No	56%	2%	13%	16%	14%	0.905
G	0.79	0.36	No	55%	2%	12%	16%	14%	0.908
D	0.75	0.72	No	49%	4%	12%	20%	15%	0.880
PSNR	0.74	0.52	No	48%	3%	12%	23%	15%	0.866
E	0.50	0.64	No	32%	6%	10%	36%	17%	0.771
C	0.38	1.44	No	24%	7%	9%	43%	18%	0.700

In conclusion, *equivalence to a video-quality test* (EVQT) is true when:

- *Concur* $\geq$ 0.91
- ΔM_{CI} is used to make decisions

D. Applied Results

We will analyze the same metrics that were evaluated by VQEG during Multimedia validation tests. Decisions reached by each dataset's stimulus pairs are combined into a single tally of incidence rates. Table XVI shows the metric name, correlation between the metric and MOS, ΔM_{CI}, EVQT, the incidence rates, and metric *concur*. The table contains rounded incidence rates for ease of comprehension.

ΔM_{CI} and EVQT must not be used for comparisons between metrics. The metrics are sorted from most accurate to least accurate, but ΔM_{CI} both increases and decreases. The problem is that ΔM_{CI} depends on the distribution of $\widehat{MOS}$, which is a unique characteristic of the metric.

X. CONCLUSION

Our goal is to provide the information and techniques that people need to make better decisions. To accomplish this, we provide new statistical methods that measure and compare the decisions reached by subjective tests and metrics that

assess media quality. ΔS_{CI}, *disagree*, and *false ranking* characterize the expected precision and repeatability of subjective tests. Expected values for these statistics become the basis of PVQT, ΔM_{CI}, and the EVQT equivalence test. They measure the precision and repeatability of metrics against the baseline performance of subjective tests. Our six new statistical methods are as follows:

- ΔS_{CI}—subjective test's confidence interval
- *Disagree*—likelihood that two labs agree that stimuli **A** and **B** are significantly different, but disagree about their quality ranking ($MOS_A < MOS_B$ vs $MOS_A > MOS_B$)
- *False ranking*—likelihood that a metric or ad hoc evaluation will rank stimuli in the opposite order to a subjective test
- PVQT—equates the metric to a number of people in a video-quality test
- ΔM_{CI}—metric's confidence interval, at which the metric has error rates similar to a subjective test
- EVQT—determines whether the metric responds similarly to a subjective test, when ΔM_{CI} is used to make decisions

These new statistical methods have a few limitations. We were only able to characterize ΔS_{CI} for the 5-level ACR method. These statistical methods cannot be used to compare the performance of multiple metrics; PVQT and ΔM_{CI} could be particularly misleading if used in that way.

These new statistical methods are designed for quality assessments of individual media files. Additional work would be needed to extend PVQT, ΔM_{CI}, and EVQT to speech quality, where the goal is to assess the quality of a system (e.g., aggregating results from multiple speakers). Additional work would also be needed to characterize ΔS_{CI} for the system quality, as measured by most speech-quality tests (i.e., aggregating ratings from multiple speakers).

This paper provides important insights into the precision and repeatability of subjective tests. Key measurements are:

- The range of ΔS_{CI} for the 5-level ACR method in Table II
- The likelihood of two subjective tests reaching different conclusions in Table VII and Table VIII
- A mapping between the *false ranking* incidence rate of a metric and the number of people in an ad hoc evaluation or pilot test in Table XIII

We recommend the *disagree* incidence rate for lab-to-lab comparisons and to determine whether experimental subjective test protocols differ significantly from proven protocols. We conclude that *disagree* incidence rates above 0.31% are unusual enough to warrant investigation and *disagree* incidence rates above 1.0% indicate difference in method, test environment, test implementation, or subject demographics.

Our Monte Carlo simulations indicate that ad hoc evaluations have dramatically higher *false ranking* incidence rates than the lab-to-lab *disagree* incidence rates noted above. Typical *false ranking* rates for an ad hoc evaluation are 13% to 10% for one person, 10% to 8% for two people, and 8% to 6% for three people. However, ad hoc evaluation repeatability is erratic, with the worst case being approximately three times higher (e.g., 30% for one person).

We welcome future discussions on these statistical methods and proposals for improved techniques. Code implementing PVQT, ΔM_{CI}, and the EVQT equivalence test is freely available in [2]. See function ci_calc.m and ci_calc.py. Many of the subject ratings used in this analysis are available on the Consumer Digital Video Library (CDVL, www.cdvl.org).

REFERENCES

[1] M. Pinson, "Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality," NTIA, Washington, DC, USA, Rep. TR-21-550, Oct. 2020. [Online]. Available: https://www.its.ntia.gov/publications/3253.aspx

[2] "NR metric framework." National Telecommunications and Information Administration, Institute for Telecommunication Sciences. Accessed: Jul. 27, 2020. [Online]. Available: https://github.com/NTIA/NRMetricFramework

[3] G. Cermak and D. Fay, "T1A1.5/94-148: Correlation of objective and subjective measures of video quality." Sep. 20, 1994. [Online]. Available: https://www.vqeg.org

[4] A. Webster, "Two criteria for video test scene selection," Working Party 2, Study Group 12, Question 22, document 35-E, ITU, Geneva, Switzerland, Dec. 1994. [Online]. Available: https://www.its.ntia.gov/publications/2598.aspx

[5] M. Pinson and S. Wolf, "Techniques for evaluating objective video quality models using overlapping subjective datasets," NTIA, Washington, DC, USA, Rep. TR-09-457, Nov. 2008. [Online]. Available: https://its.ntia.gov/publications/2494.aspx

[6] 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.

[7] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Proc. 2nd Int. Workshop Qual. Multimedia Exp. (QoMEX)*, 2010, pp. 82–87, doi: 10.1109/QOMEX.2010.5517948.

[8] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011, doi: 10.1109/TBC.2010.2086750.

[9] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Proc. 3rd Int. Workshop Qual. Multimedia Exp.*, 2011, pp. 131–136, doi: 10.1109/QoMEX.2011.6065690.

[10] M. H. Pinson et al., "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[11] S. Le Moan, M. Pedersen, I. Farup, and J. Blahová, "The influence of short-term memory in subjective image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 91–95, doi: 10.1109/ICIP.2016.7532325.

[12] A. Kumcu, K. Bombeke, L. Platiša, L. Jovanov, J. Van Looy, and W. Philips, "Performance of four subjective video quality assessment protocols and impact of different rating preprocessing and analysis methods," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 48–63, Feb. 2017, doi: 10.1109/JSTSP.2016.2638681.

[13] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, Aug. 2012.

[14] Y. Nehmé, J. Farrugia, F. Dupont, P. Le Callet, and G. Lavoué, "Comparison of subjective methods for quality assessment of 3D graphics in virtual reality," *ACM Trans. Appl. Percept.*, vol. 18, no. 1, pp. 1–23, Jan. 2021.

[15] M. H. Brill, J. Lubin, P. Costa, and J. Pearson, "Accuracy and cross-calibration of video-quality metrics: New methods from ATIS/T1A1," in *Proc. Int. Conf. Image Process.*, 2002, p. 3, doi: 10.1109/ICIP.2002.1038897.

[16] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proc. 8th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, 2016, pp. 1–6, doi: 10.1109/QoMEX.2016.7498936.

[17] L. F. Tiotsop et al., "On the link between subjective score prediction and disagreement of video quality metrics," *IEEE Access*, vol. 9, pp. 152923–152937, 2021, doi: 10.1109/ACCESS.2021.3127395.

[18] P. Corriveau and N. Walch. "VQEG subjective test plan, full reference phase 1." Video Quality Experts Group (VQEG). Jan. 18, 1999. [Online]. Available: https://vqeg.org/publications-and-software/publications

[19] A. M. Rohaly et al. “Final report from the video quality experts group on the validation of objective models of video quality assessment.” Video Quality Experts Group (VQEG). Mar. 2000. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[20] “FR-TV: Full-reference television phase II subjective test plan.” Video Quality Experts Group (VQEG). Sep. 2002. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[21] P. Corriveau and A. Webster. “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II.” Video Quality Experts Group (VQEG). 2003. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[22] A. Webster and F. Speranza. “Validation of reduced-reference and no-reference objective models for standard-definition television, phase I.” Video Quality Experts Group (VQEG). 2009. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[23] A. Webster and F. Speranza. “Report of the validation of video quality models for high definition video content.” Video Quality Experts Group (VQEG). Jun. 2010. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[24] C. Lee, S. Borer, and J. Berger. “Hybrid perceptual/bitstream validation test final report.” Video Quality Experts Group (VQEG). Jul. 2014. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[25] *ITU-T Coded-Speech Database*, Rec. ITU-T P.Sup23, Int. Telecommun. Union, Geneva, Switzerland, Feb. 27, 1998. [Online]. Available: https://www.itu.int/rec/T-REC-P.Sup23-199802-I
[26] A. Webster and F. Speranza. “Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase 1.” Video Quality Experts Group (VQEG). Sep. 2008. [Online]. Available: https://vqeg.org/publications-and-software/publications/
[27] M. H. Pinson, W. J. Ingram, and A. A. Webster, “Audiovisual quality components: An analysis,” *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 60–67, Nov. 2011.
[28] M. H. Pinson, A. A. Webster, and W. J. Ingram, “Preliminary investigation into the impact of audiovisual synchronization of impaired audiovisual sequences,” NTIA, Washington, DC, USA, Technical Memo TM-11-474, Mar. 2011. [Online]. Available: https://www.its.ntia.gov/publications/2549.aspx
[29] M. H. Pinson, S. Wolf, and R. B. Stafford, “Video performance requirements for tactical video applications,” in *Proc. IEEE Conf. Technol. Homeland Security*, May 2007, pp. 85–90.
[30] “Public safety #2 (PS2) video quality dataset,” CDVL. Accessed: Mar. 5, 2022. [Online]. Available: https://www.cdvl.org/members-section/viewfile/?id=3000
[31] Y. Zhu, L. Song, R. Xie, and W. Zhang, “SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding,” in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2016, pp. 1–4, doi: 10.1109/BMSB.2016.7521936.
[32] S. Tavakoli, K. Brunnström, J. Gutiérrez, and N. García, “Quality of experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology,” *Signal Process. Image Commun.*, vol. 39, pp. 432–443, Nov. 2015. [Online]. Available: https://doi.org/10.1016/j.image.2015.05.001
[33] L. Janowski, L. Malfait, and M. Pinson, “Evaluating experiment design with unrepeated scenes for video quality subjective assessment,” *Qual. User Exp.*, vol. 4, p. 2, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s41233-019-0026-4
[34] M. A. Saad et al., “Image quality of experience: A subjective test targeting the consumer’s experience,” in *Proc. Int. Symp. Electron. Imag. Human Vis. Electron. Imag.*, Feb. 2016, pp. 1–6.
[35] J. Nawala, M. H. Pinson, M. Leszczuk, and L. Janowski, “Study of subjective data integrity for image quality data sets with consumer camera content,” *J. Imag.*, vol. 6, no. 3, p. 7, 2020.
[36] M. H. Pinson, “ITS4S: A video quality dataset with four-second unrepeated scenes,” NTIA, Washington, DC, USA, Technical Memo TM-18-532, Feb. 2018. [Online]. Available: https://www.its.ntia.gov/publications/3194.aspx
[37] M. H. Pinson, “ITS4S2: An image quality dataset with unrepeated images from consumer cameras,” NTIA, Washington, DC, USA, Technical Memo TM-19-537, Apr. 2019. [Online]. Available: https://www.its.ntia.gov/publications/3219.aspx
[38] M. H. Pinson, “ITS4S3: A video quality dataset with unrepeated videos, camera impairments, and public safety scenarios,” NTIA, Washington, DC, USA, Technical Memo TM-19-538, Apr. 2019. [Online]. Available: https://www.its.ntia.gov/publications/3220.aspx
[39] M. H. Pinson and S. Elting, “ITS4S4: A video quality study of camera pans,” NTIA, Washington, DC, USA, Technical Memo TM-20-545, Dec. 2019. [Online]. Available: https://www.its.ntia.gov/publications/3233.aspx
[40] B. Naderi, T. Hoßfeld, M. Hirth, F. Metzger, S. Möller, and R. Z. Jiménez, “Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach,” 2020, *arXiv:2003.11300*.
[41] L. Janowski and M. H. Pinson, “The accuracy of subjects in a quality experiment: A theoretical subject model,” *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec. 2015.

Margaret H. 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 Telecommunications and Information Administration, Boulder. She is an Internationally Recognized Expert with 33 years of experience developing improved methods for assessing video quality. Her research includes algorithm development, human testing, and international standards. She contributed to eight national and international efforts of ATIS and the Video Quality Experts Group (VQEG) to independently validate video quality metrics. She led the effort to create ITU-T Rec. P.913, which describes improved subjective test methods for modern video systems. She has written 83 publications. Her current research focuses on NR metrics that predict what people would say is the quality of an image or video.

Ms. Pinson is a VQEG Co-Chair, administers the Consumer Digital Video Library (CDVL), and makes all of her algorithms openly available. She contributes to ITU Recommendations and has led several efforts to independently validate video quality metrics, which is a necessary step of the standards development process. She helped design and conduct four prize challenges, including the *5G Challenge*. Her prior IEEE TRANSACTIONS ON BROADCASTING article shows that NR metrics must be trained on significantly more data, if they are to accurate enough to be deployed by industry.