NTIA Report 21-550

Confidence Intervals for Subjective Tests and Objective Metrics That Assess Image, Video, Speech, or Audiovisual Quality

Margaret H. Pinson

report series

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Report 21-550

Confidence Intervals for Subjective Tests and Objective Metrics That Assess Image, Video, Speech, or Audiovisual Quality

Margaret H. Pinson

U.S. DEPARTMENT OF COMMERCE

Adam Candeub
Acting Assistant Secretary of Commerce for Communications and Information
National Telecommunications and Information Administration

October 2020

DISCLAIMER

Certain products, technologies, and corporations are mentioned in this report to describe aspects of the ways that digital images and videos are created, modified, transmitted, and consumed at present or may be in the future. The mention of such entities should not be construed as any endorsement, approval, recommendation, prediction of success, or that they are in any way superior to or more noteworthy than similar entities that were not mentioned.

CONTENTS

FIGURES

TABLES

CONFIDENCE INTERVALS FOR SUBJECTIVE TESTS AND OBJECTIVE METRICS THAT ASSESS IMAGE, VIDEO, SPEECH, OR AUDIOVISUAL QUALITY

Margaret H Pinson[1]

This report describes a methodology that measures the precision of objective metrics that assess image quality, video quality, speech quality, or the overall audiovisual quality. We assess the confidence intervals of 60 subjective tests and use a confusion matrix to classify the conclusions reached when two subjective test labs perform the same experiment. This allows us to compute the metric's confidence interval and, when confidence intervals are used to make decisions, to prove whether the metric performs similarly to a subjective test with 15 or 24 subjects. When confidence intervals are not used, the metric's precision is likened to a certain number of people in an ad-hoc quality assessment. The methods in this report are developed and evaluated using speech quality, video quality, image quality, and audiovisual quality datasets.

Keywords: Audiovisual quality, confidence interval, image quality, metric, speech quality, subjective test, video quality

1. INTRODUCTION

Standard statistical techniques fail to answer critical questions raised by standards developing organizations (SDO) when validating quality metrics. The SDOs have agreed upon subjective test methods where panels of people rate the quality of images, speech, video, or audiovisual media and then aggregate these ratings to produce Mean Opinion Scores (MOS). The SDOs have also agreed upon statistical methods to assess the accuracy of metrics that estimate MOSs.

The most common unanswered questions are:

- How do we decide that a metric is good enough?
- When does a metric's performance rise to that of subjective testing?
- How do we explain a metric's precision to naive users?

Through early 2020, lack of verifiable answers caused three problems. First, SDOs assume that metrics cannot rise to the accuracy of subjective testing, which is unproven. Second, metric developers and International Telecommunications Union (ITU) Recommendations fail to provide guidance on metric precision. Consequently, many users assume metrics have infinite precision,

[1] The author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

which we know to be incorrect. Third, SDOs use Peak Signal to Noise Ratio (PSNR) as a minimum performance benchmark.

The PSNR benchmark asks the question, “Would a well-informed user choose PSNR instead of this alternative metric?” A metric is judged worthy if its accuracy is statistically better than PSNR. If the user could not use PSNR, due to the lack of a high quality version of the media to act as a reference, then the metric is judged worthy if its accuracy is statistically equivalent to or better than PSNR. The PSNR performance benchmark is pragmatic yet hotly disputed. For example, PSNR is not based on human perception, performs poorly for transmission errors, has a non-linear relationship to MOS, and has a strong scene bias (e.g., PSNR’s accuracy can be improved by removing an offset for each reference video).

We collected individual subject ratings from 60 subjective tests that were performed in compliance with ITU-R Rec. BT.500, ITU-T Rec. P.913, and ITU-T P.800. When aggregated, 2,331 subjects rated the quality of 17,665 media files, for a total of 433,398 subject ratings. The media files include 2,592 images (15%), 7,959 speech samples (45%), 6,445 silent videos (36%), and 669 audiovisual files (4%). The ratings include 90 lab-to-lab comparisons, where two labs conducted the same subjective test. The ratings also include 13 comparisons between similar subjective tests (e.g., the same stimuli but a different rating method, at the same lab or a different lab). These 60 subjective tests were supplemented by MOSs from 28 video quality subjective tests for which individual subject ratings are not available.

We begin by analyzing the expected behavior of subjective tests. We define ΔS_{CI}, which is a new method to measure the precision of subjective test. Conceptually, ΔS_{CI} is the subjective test’s overall confidence interval (CI). We also establish a confusion matrix that categorizes the conclusions reached by multiple labs conducting the same subjective test.

Using our 60 datasets, we calculate ΔS_{CI} and the relationship between ΔS_{CI} and the number of subjects in the test. We condense our findings into the expected CI of a well-designed and carefully conducted subjective test that uses 24, 15, 9, or 6 subjects and the Absolute Category Rating (ACR) method. Using the confusion matrix, we measure the likelihood that two subjective tests will reach different conclusions. These measurements establish the expected precision of subjective ratings.

We use a second confusion matrix to compare conclusions reached by the metric with conclusions reached by subjective tests. Within that confusion matrix, we use a constant value to categorize the metric’s conclusions (i.e., **A** is better than, equivalent to, or worse than **B**). To be fair, we ignore the subjective test’s rating distributions and use our empirical data to categorize the subjective test’s conclusions, based on the expected CI of a well-designed and carefully conducted subjective test. These design elements prevent unintended bias in favor of subjective tests.

We propose a method for calculating a CI that indicates the metric’s precision. We select the CI where the metric is no more likely to falsely differentiate between two stimuli or falsely rank two stimuli than a subjective test. Our subjective data analyses provide error rate thresholds. We recommend an ideal CI (selected with stringent criteria) and a practical CI (selected with less stringent criteria). We propose a method to show that the metric is equivalent to a subjective test, when these CIs are used to make decisions.

Metric values are often compared directly, without CIs. Therefore, we also propose a method to equate the metric to a certain number of subjects in an ad-hoc quality assessment or pilot test. This method assesses the metric's precision when CIs are not used. We use a confusion matrix, a figure-of-merit based on the level of agreement between subjective tests, and a threshold based on the level of disagreement between subjective tests.

These new methods provide insights into the metric's precision. Our supporting analyses of subjective ratings provide insights into the precision and repeatability of subjective tests. The statistical methods and observed trends can be used to predict future trends when assessing image quality, video quality, speech quality, and audiovisual quality. Code implementing the statistics described in this report can be found in the *NRMetricFramework* GitHub repository [1].

2. BACKGROUND

2.1 Confidence Intervals for Metrics

In the early 1990s, subject matter experts met at NTIA/ITS to brainstorm improved methods to express the precision of video quality metrics. These discussions culminated in ATIS T1.TR.72 [2] and ITU-T Rec. J.149 [3].

The first proposed solution, resolving power, was included in both Recommendations. Brill *et al.* [4] provides the clearest presentation of the resolving power statistic. Pinson and Wolf [5] demonstrates resolving power on subjective test data. 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 (e.g., a subjective dataset conducted to validate the metric). Either the metric is mapped to the subjective test or vice versa.

Resolving power has been quietly 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.

The second proposed solution, based on a confusion matrix, appears in [6] 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 that has been deemed to be significant. When two stimuli **A** and **B** are compared, we conclude that either **A** is better than **B**, **A** is equivalent to **B**, or **A** is worse than **B**. We can compute such comparisons for all pairs of stimuli in a dataset to calculate overall statistics on, for example, how often the subjective test and the metric reach the same conclusion.

This confusion matrix idea also failed to gain traction. One problem is that the method analyzes many CI options without recommending how to choose among them. The method proposed in this report is inspired by the ATIS T1.TR.72 confusion matrix method, yet includes major changes to the underlying algorithm.

3. SUMMARY OF SUBJECTIVE DATASETS

Table 1 summarizes the subjective datasets that we will use in this report. Occasionally, this simplified presentation may be confusing or misleading. For example, if this report uses only part of a dataset, then Table 1 does not describe the unused parts of the dataset. To fully understand the experiment designs, refer to the original publications [7] to [34].

Reading these references should not be necessary. The omitted information is not used by this report's analyses (e.g., media subject matter, impairments, test environment, and subject demographics). We would need significantly more data to develop a general rule that characterizes the influence of these secondary factors.

Table 1 uses horizontal lines to distinguish between experiments. Most of these experiments contain one subjective test, but the following experiments contain two or more distinct subjective tests: ITS4S3, ITU-T P.Sup23, Private Video Dataset #1, and all of the VQEG validation tests. The tally of 60 subjective tests in the introduction accounts for this distinction.

Several of these experiments contain a complex division of stimuli among subjects at different labs. We define a stimulus to be one file containing an image, video, or speech sample. If the stimuli are divided into subsets and rated by different subject pools, then Table 1 contains one line for each subset. This line specifies the number of labs, stimuli, and subjects associated with that subset.

Table 1 contains the following columns:

- Dataset Name of the dataset
- Ref. Reference to a publication that describes the dataset
- Open Access Whether the dataset is available to researchers (as of August 2020)
- Study Type "Crowdsource" for crowdsourcing experiments;
"Field" for field studies (e.g., exploratory designs, real-world impairments with confounding variables, prototype tests with few subjects);
"Lab" for controlled lab studies; or
"SDO" for controlled lab studies conducted by standards developing organizations according to rigorous test plans
- Media Whether the dataset contains images, speech, video (without sound), or audiovisual media (video with sound)
- Method Subjective test method
- Scale [1..5] for the discrete 5-level scale: excellent, good, fair, poor, and bad;
[0..100] for the continuous 100-level scale;
[-3..3] for the 7-level comparison scale
[0,1] for a Boolean scale (acceptable, or unacceptable)

- Subsets — The names of media subsets (assessed by distinct subject pools); or "—" if all subjects rated all stimuli
- Labs — Number of labs that contributed subject ratings
- Stimuli — Number of stimuli
- Subjects — Number of subjects;
  "—" if individual subject ratings are not available; or
  ($Y = X_1 + X_2 + \ldots + X_N$) if subjects at multiple labs rated the same stimuli, where Y is the total number of subjects, X_i is the number of subjects from one lab, and N is the number of labs
- Available — "Ratings" if individual subject ratings are available;
  "MOS" if only mean opinion scores (MOS) are available; or
  "Simulated ratings" if a technique other than subjective testing was used to simulated MOSs (see [13])

Most of these datasets use the ACR method, where subjects view and rate each stimulus in isolation. One dataset uses the Comparison Category Rating (CCR) method, which is also known as the Double Stimulus Comparison Scale (DSCS) or Pair Comparison (PC). CCR presents two stimuli in random order, using a 7-level scale. We removed the random ordering before our analyses, so most of the data is in the range [0..3]. Several datasets use the Double Stimulus Continuous Quality Scale (DSCQS) method. DSCQS subjects view the source (SRC) video, view the processed video sequence (PVS), and then rate the SRC and PVS separately on [0..100] scales. The Difference of Scores (DOS) is calculated from these separate ratings. We will analyze all three sets of ratings (SRC, PVS, and DOS). These subjective methods are described in [35].

Several datasets use non-standard methods. Public safety #1 and Public Safety #2 asked subjects to rate all stimuli on both a 5-level ACR scale and a Boolean scale. The latter assessed whether the video quality presented was acceptable for public safety practitioner tasks (e.g., tactical response, observation, video recordings). The UPM-Acreo dataset compares 5-level ACR with the Content-Immersive Evaluation of Transmission Impairments (CIETI) method. CIETI produces single stimulus ratings on the ACR's [1..5] scale. For details on the CIETI method, see [25] and its references. The Private Video Dataset #3, Netflix Quality Variation 2017, uses both 5-level ACR and Single Stimulus Continuous Quality Evaluation (SSCQE).

The CCRIQ [10] dataset's subjects rated each image on an HD monitor and a 4K monitor. An increase of ≈0.2 MOS was associated with the 4K monitor for images with MOS greater than 3.0 on a [1..5] scale, while the 4K and HD monitor MOSs were equivalent for images with MOS less than 3.0. We will merge the 4K and HD ratings into a single subject pool with twice the number of subjects (i.e., one for the 4K monitor and another for the HD monitor). Thus, CCRIQ demonstrates a subjective test where there is a small but statistically significant difference of opinion among subjects (due to the monitor used).

Some of the datasets are distributed with individual subject ratings, which makes more analyses possible. Most of these datasets are distributed through the Consumer Digital Video Library (CDVL, www.cdvl.org). The Supplement 23 to the ITU-T P Series (P.Sup23) datasets are available on the ITU website (see [16]). Some of the datasets are distributed with MOSs only. More information on these datasets, including the download link, can be found in the *NRMetricFramework* GitHub repository [36].

The remainder of the datasets in Table 1 cannot be distributed. The private speech dataset #1 contains results from proprietary subjective ACR tests conducted on narrowband speech codecs using simulated wireless channels. The private speech dataset #2 contains results from proprietary subjective ACR tests conducted on narrowband speech codecs using wireline and simulated wireless channels. The private video dataset #1 contains two subjective tests conducted on the 100-point scale: one rated by experts and the other by crowdsourcing. The private video dataset #2 contains 1-minute video sequences.

The Video Quality Experts Group (VQEG) Multimedia experiment designs are published in [32], but the ratings, metric data, and videos are protected by a multiple party non-disclosure agreement. ITS obtained permission from all participants to use the VQEG Multimedia ratings to develop improved statistical analysis methods for objective metrics. The original analysis produced a technique to use common subsets of stimuli to map multiple datasets onto a single scale, as proposed in [5]. Common subsets of stimuli are available for the VQEG Multimedia, High Definition Television (HDTV), and Hybrid Perceptual Bit-stream (Hybrid) datasets. These common sets are not noted in the table below.

Table 1 contains datasets that assess the quality of images, videos, audiovisual media, and speech. The subjective test methods and rating tasks are fundamentally alike. However, the experiment designs contain major differences due to the treatment of impairments, which are referred to as conditions by speech quality researchers and Hypothetical Reference Circuits (HRC) by video quality researchers.

The image and video datasets use similar experiment designs, because images are a simplified case of “video without motion” when both are presented on digital monitors. Different scenes have a major impact on image and video impairments, which results in a wide variance of MOSs. While experiment designs focus on impairments, analyses focus on individual PVSs (i.e., rating distributions for a single image or video file). Therefore, each file is typically rated by 15 or 24 subjects. Audiovisual datasets follow these same trends.

Speech tests are organized differently. Different talkers and utterances produce similar MOSs. Analyses typically ignore the file response (e.g., distribution of ratings for a single speech file) in favor of condition response (e.g., a particular speech codec, bit-rate, and packet loss rate). Consequently, speech quality datasets typically contain fewer subject ratings per file (e.g., 8 to 12) and more files per condition (i.e., more source stimuli for each impairment).

Because of this difference, we will occasionally analyze the condition response of speech quality datasets. Still, speech quality metrics predict the quality of a single stimulus (file), just like image and video quality metrics. Therefore, our ultimate goal for both subjective test and metric analyses is to analyze individual files (i.e., an image, short video, or speech utterance).

When a video quality test is conducted in multiple labs, all subjects view and rate the same video files. By contrast, when a speech quality test is conducted in multiple labs, each lab typically selects original speech utterances in their native language to associate with each impairment. That is, subjects from different labs analyze the same condition using different files. The ITU-T Rec. P.Sup23 dataset is designed this way, so we cannot use it for lab-to-lab comparisons. The only speech dataset that we can use for lab-to-lab comparisons is private speech dataset #3.

Table 1. Subjective Tests with Published Subject Rating

Dataset	Ref.	Open Access	Study Type	Media	Method	Scale	Subsets	Labs	Stimuli	Subjects	Available
401	[7]	Yes	Crowdsource	Speech	ACR	[1..5]	—	1	1152	8 (192 per condition)	Ratings
501					ACR	[1..5]	—	1	200	24 (96 per condition)	
701					ACR	[1..5]	—	1	1152	8 (128 per condition)	
AGH/NTIA/Dolby	[8]	Yes	Field	Video	ACR	[1..5]	—	3	230	71 = 31 + 22 + 18	Ratings
BID	[9]	Yes	Lab	Image	ACR	[1..5]	—	1	582	—	MOS
CCRIQ	[10]	Yes	Field	Image	ACR	[1..5]	Blue Red	3	221 171	26 = 9 + 8 + 9 27 = 9 + 9 + 9	Ratings
CCRIQ2	[11]	Yes	Field	Image	ACR	[1..5]	—	1	88	19	Ratings
CID2013	[12]	Yes		Image	ACR	[1..5]	—	1	474	—	MOS
DIQA	[13]	Yes	Field	Image	Objective	—	Fine Reader Omni Tesseract	—	175	—	Simulated ratings
ITS 2010	[14]	Yes	Lab	Audio & video	ACR	[1..5]	—	1	240	≈26	Ratings
ITS AV-Sync 2010	[15]	Yes	Lab	Audio & video	ACR	[1..5]	—	1	297	28 = 12 + 16	Ratings
ITU-T P.Sup23	[16]	Yes[2]	SDO	Speech	ACR CCR ACR	[1..5] [-3..3] [1..5]	EXP1 EXP2 EXP3	3 3 4	176 136 200	72 = 24 + 24 + 24 144 = 48 + 48 + 48 96 = 24 + 24 + 24 + 24	Ratings
ITS4S	[17]	Yes	Field	Video	ACR	[1..5]	Full Partial	1 2	813 212	27 51 = 27 + 24	Ratings
ITS4S2	[18]	Yes	Field	Image	ACR	[1..5]	—	1	1429	16	Ratings
ITS4S3	[19]	Yes	Field	Video	ACR	[1..5]	CS SR CW VW	1 1 1 1	99 99 99 99	14 17 14 15	Ratings

[2] ITU-T P.Sup23 constrains use of the speech files to the development of new and revised ITU-T Recommendations. This report 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.”

Dataset	Ref.	Open Access	Study Type	Media	Method	Scale	Subsets	Labs	Stimuli	Subjects	Available
							MPR FG	1 1	99 99	13 19	
ITS4S4	[20]	Yes	Field	Video	ACR	[1..5]	—	1	196	26	Ratings
KoNVid-1k	[21], [22]	Yes	Crowdsource	Image	ACR	[1..5]	—	1	961	—	MOS
LIVE-Wild	[23]	Yes	Crowdsource	Image	ACR	[0..100]	—	1	1,153	—	MOS
Private Speech Dataset #1	—	No	Lab	Speech	ACR	[1..5]	—[3]	1	1,359	≈11, ≈22, or 43 per file 344 or 440 per condition	Ratings
Private Speech Dataset #2	—	No	Lab	Speech	ACR	[1..5]	—	1	2,432	8 per file 512 per condition	Ratings
Private Speech Dataset #3	—	No	Lab	Speech	ACR	[1..5]	A B C D	2	288 288 288 288	18 = 10 + 8 16 = 8 + 8 16 = 8 + 8 16 = 8 + 8	Ratings
Private Video Dataset #1	—	No	Lab Crowdsource	Video	ACR	[0..100]	A B	1	75 112	15 experts 61 crowdsource	Ratings
Private Video Dataset #2, OPTICOM	—	No	Lab	Video	ACR	[1..5]	—	1	60	30	Ratings
Private Video Dataset #3, Netflix Quality Variation 2017	—	No	Lab	Video	ACR, SSCQE	[1..5]	Subset 1 Subset 2	2 2	180 180	102 = 51 + 51 98 = 50 + 48	Ratings
Public Safety #1	[24]	Yes	Lab	Video	ACR	[1..5] 0 or 1	—	1	400	16 first responders	Ratings
Public Safety #2	—	Yes	Lab	Video	ACR	[1..5] 0 or 1	—	1	576	19 first responders	Ratings

UPM-Acreo	[25]	Ratings	Lab	Audio & Video	ACR, CIETI	[1..5]	ACR(V)
							CIETI(V)
							CIETI(AV)
								2	132	20
										22
										21
											Ratings

[3] The files and subjects were divided into overlapping subsets.

Dataset	Ref.	Open Access	Study Type	Media	Method	Scale	Subsets	Labs	Stimuli	Subjects	Available
VIME1	[11]	Yes	Field	Image	ACR	[1..5]	—	1	101	21	Ratings
VQEG FRTV Phase I	[26], [27]	Yes	SDO	Video	DSCQS	[0..100]	525-line, low	4	100	70 = 18 + 18 + 18 + 16	Ratings
							525-line, high	4	100	70 = 18 + 18 + 18 + 16	
							625-line, low	4	100	70 = 18 + 18 + 18 + 16	
							625-line, high	4	100	67 = 18 + 17 + 16 + 16	
VQEG FRTV Phase II	[28], [29]	Ratings	SDO	Video	DSCQS	[0..100]	525-line	2	63	64 = 32 + 32	Ratings
							625-line	1	70	27	
VQEG HDTV	[30]	Yes (Mostly)	SDO	Video	ACR	[1..5]	HD1	1	168	24 each	Ratings
							HD2	1	168		
							HD3	1	152		
							HD4	1	168		
							HD5	1	168		
							HD6	1	168		
VQEG Hybrid	[31]	Ratings	SDO	Video	ACR	[1..5]	HD1	1	184	24 each	Ratings
							HD2	1	184		
							HD3	1	184		
							HD4	1	184		
							HD5	1	184		
							VGA1	1	114		
							VGA2	1	194		
							VGA3	1	184		
							WVGA1	1	194		
							WVGA2	1	120		
VQEG Multimedia (MM)	[32]	No	SDO	Video	ACR	[1..5]	13 VGA resolution subsets	13	166 each	—	MOS
VQEG Multimedia 2 (MM2)	[33]	Yes	Lab & Field	Video	ACR	[1..5]	—	10	60	213 = 28+9+34+25+25+24+24+14+15+15	Ratings
VQEG RRNR-TV	[34]	Ratings	SDO	Video	ACR	[1..5]	525-line	2	168	32 = 16 + 15	Ratings
							625-line	2	168	31 = 16 + 15	

4. CONFIDENCE INTERVALS FOR SUBJECTIVE TESTS

Let us begin by analyzing the precision of subjective tests. The intermediate data for this section appears in Appendix A. Section 3 contains information on where to obtain the subjective ratings, if these are publicly available.

Pinson *et al.* [33] proves that MOSs are relative, not absolute. That is, we expect the ordering and relative distances between MOSs to be replicable when we conduct the same subjective test in two labs. Conversely, we do not expect multiple labs to produce identical MOSs. We will accept the theorem that MOSs are relative for the remainder of this report.

One of the most common analyses performed on subjective data is to compare the MOSs of two stimuli. For example, we encode the same sequence at two different bit-rates and apply the paired sample Student's t-test to determine if the MOSs are statistically different, given the distribution of subject ratings for each stimulus.

Video quality metrics traditionally estimate MOS, which we will refer to as $\widehat{MOS}$. Very few metrics estimate the distribution of subject ratings, so we cannot use the Student's t-test to compare $\widehat{MOS}s$. We have two options to determine whether the difference between two $\widehat{MOS}$s is significant. First, we can use a CI. Second, we can use deterministic math (e.g., any difference in metric values is assumed to be significant).

Either way, we have confounding factors:

- Source of data (subjective ratings versus objective metric)
- Type of data (collection of ratings versus $\widehat{MOS}$)
- Method of comparison (statistical test, CI, or deterministic math)

This predisposes our measurement sensitivity in favor of subjective testing. Philosophically, this is undesirable. Our lack of understanding of the sensitivity of subjective test data can lead users to draw unwarranted negative conclusions about the accuracy of metrics. When we evaluate the accuracy and precision of subjective tests, we use probability theory with the understanding that quality ratings include random processes. However, we do not take the final step of describing the precision of the subjective test's MOSs in a way that can be easily applied to $\widehat{MOS}$s. Therefore, we will begin by investigating the precision of subjective tests, measured as a constant CI. We will then delve deeply into the differences that occur when a subjective test is repeated at two different labs. We will conclude by investigating the impact of the rating scale on these lab-to-lab differences.

4.1 From Student's *t*-test to Confidence Interval

In this section, we will explore the relationship between conclusions reached by the Student's t-test and the distance between two MOSs. This prepares us to define $\Delta S_{CI,}$ which is a new measure of the precision of subjective test.

Given a subjective test, 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. For each pair of stimuli, **A** and **B**, we will measure ΔS, the absolute value of the distance between the MOSs of A and B (i.e. MOS(A) – MOS(B)). We will also use the paired stimuli Student's *t*-test to compare the rating distributions for **A** and **B** at the 95% confidence level.[4] We will record 0 if the conclusion is that the stimuli are equivalent, and 1 if the conclusion is that **A** and **B** are different. We will tally these comparisons in a new binary variable, **EQ**.

We will bin ΔS by 0.1 MOS intervals (0 ±0.05, 0.1 ±0.05, 0.2 ±0.05, …) and compute π, the average response for that population, expressed as a percentage (i.e., average the 0/1 responses and multiply by 100).

$$\pi = \mathrm{mean(EQ)} \times 100 \qquad (1)$$

Note that π ranges from [0..100] where 0% means that all pairs of stimuli (**A**, **B**) have equivalent quality, and 100% indicates that all pairs of stimuli have significantly different quality (measured at the 95% confidence level).

Figure 1 plots π as a function of ΔS. For this computation, subjects from different labs are pooled together. If the dataset contains subsets of stimuli rated by different subsets of subjects, then the pairs (**A**, **B**) are constrained such that both **A** and **B** must appear in the same subset of subjects. The left sub-figure contains VQEG HDTV and VQEG Hybrid datasets, which include data from 16 subjective tests. These lab studies were conducted according to VQEG validation test plans as a critical element of the standards development process. These datasets use conventional experiment designs (e.g., a full matrix of scenes and impairments). The right sub-figure contains field studies (i.e., AGH/NTIA/Dolby, CCRIQ, CCRIQ2, ITS4S2, ITS4S3, ITS4S4, and VIME1). These datasets seek increased realism at the cost of reduced control and more confounding factors in the experiment design. AGH/NTIA/Dolby explores novel experiment designs, while the other datasets analyze photographs from commercial cameras. The VQEG MM2 field study is omitted from Figure 1 due to the abnormally large number of subjects. Notice how compact the lab studies are compared to the field studies. This indicates the value of carefully constructed experiment designs that control as many variables as possible.

Each dataset is plotted separately as narrow blue lines. Datasets with 24+ subjects are plotted as solid blue lines, and datasets with 23 to 13 subjects are plotted as dashed-dotted blue lines. The heavy black line aggregates the pairwise conclusions from all datasets on that sub-plot, further constrained that all pairs (**A**, **B**) must be drawn from the same dataset. Thus, the heavy black line indicates the median response.

Figure 2 plots π as a function of ΔS for four datasets that were conducted in multiple labs: CCRIQ, ITS4S4, AGH/NTIA/Dolby, and VQEG MM2. Each thin line (blue or pink) contains data from one lab's subjects. Blue indicates the subjects were in a controlled environment; purple

[4] The other logical choice would be 100%, thus ensuring all stimulus pairs can be distinguished. Subject ratings are impacted by random processes, so we must expect unusual outliers: stimulus pairs that cannot be distinguished despite being separated by a relatively large ΔS. We do not want to give these outliers undue influence. We will apply this same philosophy to objective metric outliers.

indicates the subjects were in a public environment (e.g., cafeteria, patio, or hallway). CCRIQ subdivides each lab into the two subsets (Red and Blue), for a total of six sets of subjects. Solid lines indicate 24+ subjects, dot-dashed lines indicate 23 to 13 subjects, and dashed lines indicate 9 to 8 subjects. The thick black line includes all subjects (i.e., pairs (**A**, **B**) use ratings from all labs). Thus, the heavy black line shows a low ΔS that reflects the conclusions reached by the entire dataset, when all subjects are pooled.

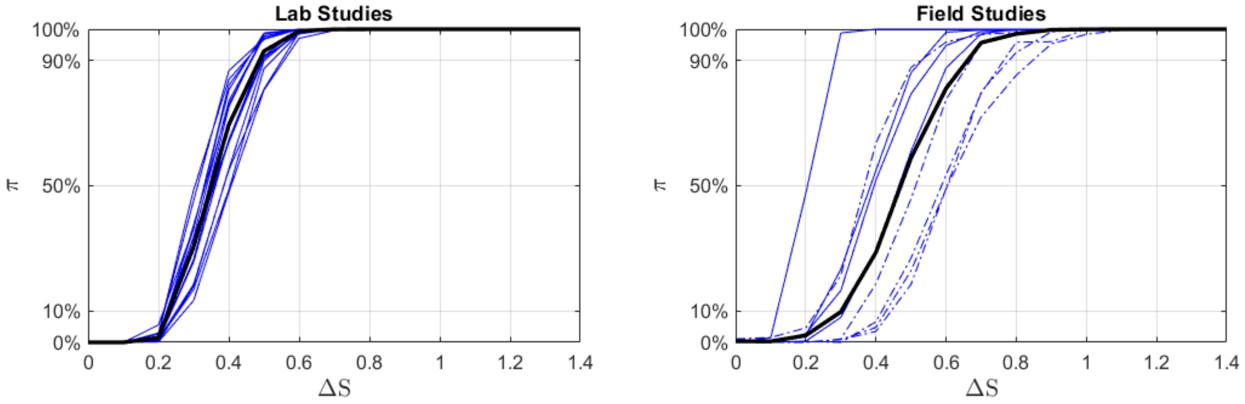

Figure 1. Relationship between π and ΔS for lab studies (left) and field studies (right).

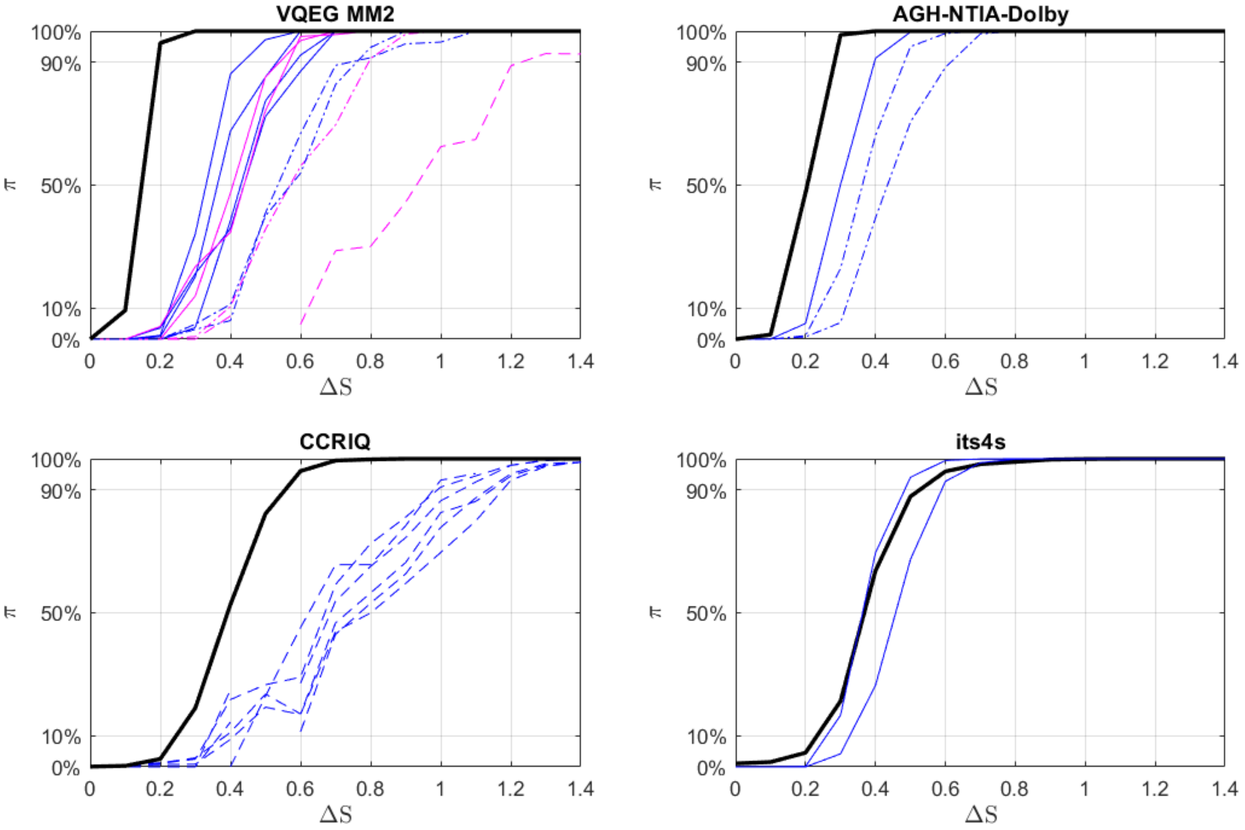

Figure 2. Relationship between π and ΔS for subjective tests conducted at multiple labs.

Figure 2 shows us that the ΔS needed to distinguish between stimulus pairs increases as the number of subjects decreases. ITU-R Rec. BT.500 recommends 15 subjects and has recommended so for decades. ITU-T Rec. P.913 recommends 24 subjects, based on the experience of subject matter experts shared during discussions at VQEG meetings and analyses of the VQEG MM2 multiple lab study [33]. To summarize [33], 24 subject experiments yield a much more stable and repeatable experiment result across different test labs than 15 subject experiments; and the averaged lab-to-lab correlations are always 0.96 or greater, which indicates a well conducted experiment. Also note that two of the four experiments in Figure 2 have an abnormally large number of subjects: 213 for CCRIQ and 71 for AGH/NTIA/Dolby.

Caution must be taken to not extrapolate other facets of subjective tests based on the trends shown in Figures 1 and 2. The VQEG MM2 data in Figure 2 plots labs with different test environments and different numbers of subjects. When we examine only the solid lines or only the dot-dashed lines (containing a similar number of subjects), the test environment (controlled or public) has minimal impact on the relationship between π and ΔS. However, the analyses of the VQEG MM2 experiment presented in [33] concluded that 35 subjects are needed in a public environment to imitate the precision of 24 subjects in a laboratory environment. That analysis asked a very different question (i.e., the percentage of stimulus pairs that could be differentiated using the Student's *t*-test). Similarly, Figures 1 and 2 change by less than 1% when subject bias is removed, despite reducing the standard deviation of scores (see [38] and [37]). The aggregation required to calculate the ΔS versus π curves masks these phenomena.

We now define a new measure of the precision of a subjective test, ΔS_{CI}. We call this the test's CI, but this should not be confused with CIs calculated by other statistical processes. We define a subjective test's CI, ΔS_{CI}, as the ΔS level that is closest to distinguishing between 95% of all stimuli pairs. For example, the thick black line in Figure 1 yields the following CIs for 5-level ACR tests: lab studies reach π = 93% for ΔS_{CI} = 0.5, and field studies reach π = 96% for ΔS = 0.7.

4.2 Confidence Interval for a Typical Subjective Test

In this section, we will calculate ΔS_{CI} for many datasets, to establish trends. We calculate ΔS_{CI} as the ΔS that comes closest to producing π = 95% for the datasets in Table 1 for which individual subject ratings are available. Figure 3 shows the resulting relationship between ΔS_{CI} and the number of subjects for 5-level ACR datasets. This figure omits datasets that used other rating scales (e.g., DSCQS, CCR, ACR-HR); these will be discussed later.

Figure 3 presents the same relationship in three different ways: a 2-D histogram and two scatter plots. On the histogram, color indicates the number of datasets in the bin. On the scatter plots, the area of the dot increases linearly with the number of datasets that produce this result. The large blue dot represents the most common result (ΔS_{CI} = 0.5 for 24 subjects). The blue circles mark extrapolations for fewer subjects.

VQEG HDTV, VQEG Hybrid, VQEG RRNR-TV, and ITU-T P.Sup23 were conducted by SDOs using the 5-level ACR scale and 24 subjects. These tests represent an ideal of carefully designed and executed lab studies. Considered individually, 21 of the 27 subjective tests yield ΔS_{CI} = 0.5, and 6 of 27 yield $\Delta S_{CI} = 0.6$. The type of media (speech or video) does not appear to impact ΔS_{CI}.

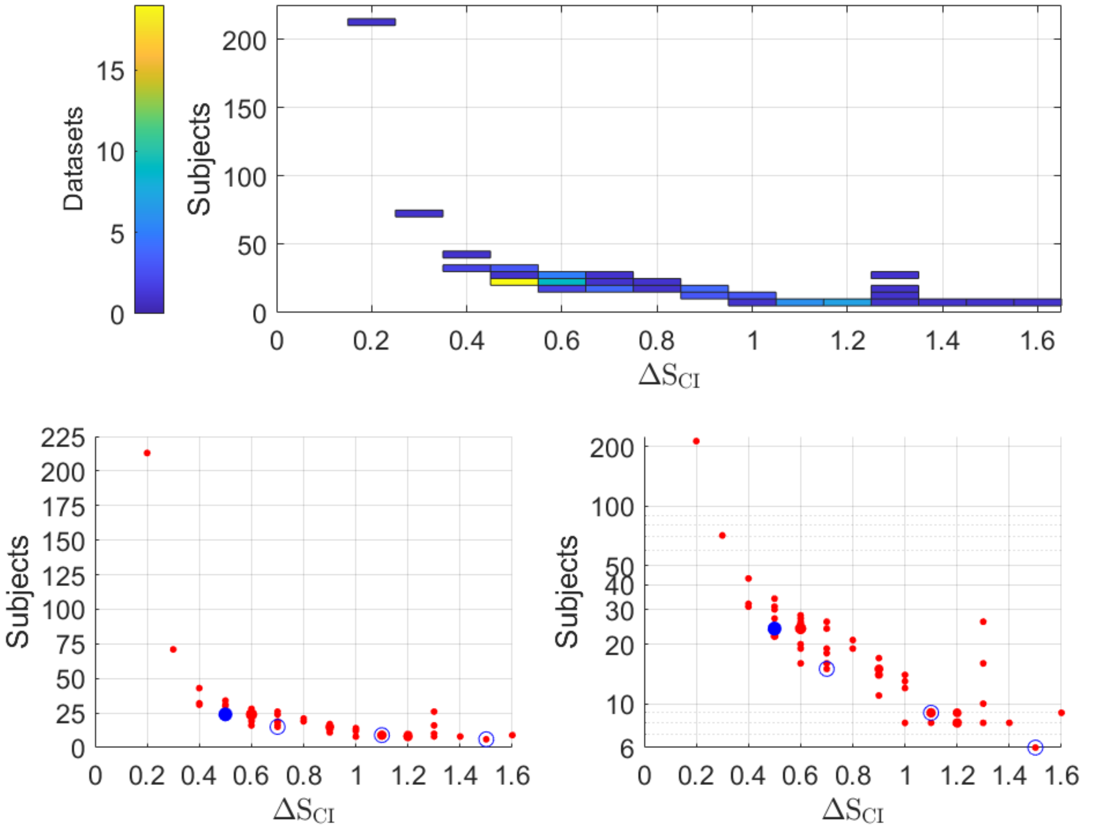

Figure 3. Relationship between ΔS_{CI} and number of subjects in the dataset, presented as a histogram (top), scatter plot (bottom left), and scatter plot with log y-axis (bottom right).

We will now focus on the VQEG HDTV and VQEG Hybrid datasets, because these 16 datasets were conducted according to two test plans with many similar characteristics. When stimulus pairs from these 16 datasets are aggregated, $\Delta S_{CI} = 0.5$. Using these 233,655 stimulus pairs, we will extrapolate ΔS_{CI} values for experiments with 15, 9, and 6 subjects. We selected subjects at random from each dataset, calculated ΔS_{CI}, recorded the results, and repeated with different random selections of subjects a total of six times. The results were stable, so no additional random selection cycles were performed. These results, marked with blue circles on Figure 3, yielded ($\Delta S_{CI} = 0.7$ for 15 subjects), ($\Delta S_{CI} = 1.1$ for 9 subjects), and ($\Delta S_{CI} = 1.5$ for 6 subjects). All plots in this section include these three extrapolated values.

Figure 3 shows that ΔS_{CI} decreases as the number of subject increases, as we expected. The range of ΔS_{CI} values associated with a particular number of subjects is fairly low (± 0.1 MOS). We observe some variations that are worth closer examination on subsequent graphs. Datasets with more than 45 subjects are omitted from these graphs, to highlight differences among datasets with more typical numbers of subjects.

Figure 4 shows the impact of test environment and experiment design on ΔS_{CI}. In these scatter plots, the area of the dot increases linearly with the number of datasets. The current subset is plotted in blue, and the other data are plotted in green. Subjective tests conducted by SDOs (Figure 4 upper-left) have lower values for ΔS_{CI}, near the limit of what we have observed. This is no doubt due to the care taken in crafting and executing the subjective tests. Subjective tests conducted in public environments (Figure 4 lower-right) seem to have larger values for ΔS_{CI}, however all of this data comes from one experiment (VQEG MM2). Lab studies and field studies span the full spread of ΔS_{CI}. This indicates that exploratory designs and real-world impairments do not negatively impact ΔS_{CI}—or that the low precision of our measurements masks a subtle difference between the number of subjects needed in lab studies and field studies. The low precision of our measurements masks other phenomena, such as the analysis presented in [10].

Figure 5 shows the impact of media on ΔS_{CI}. These scatter plots follow the same size and color convention as used in Figure 4. We have relatively few datasets with images, speech, or audiovisual media, so we cannot reach strong conclusions. Speech tests seem to have lower ΔS_{CI} values that are within the observed range of image and video tests. This is plausible, as the variability among phonemes is smaller than the variability of visual subject matter; ears have higher sensitivity than eyes; and our auditory system can listen to multiple sounds simultaneously, while our visual system is influenced by attention (e.g., where we focus within the field of view). Audiovisual media seem to have slightly high ΔS_{CI} irrespective of the number of subjects. This is also plausible, because the spread of ratings is impacted by two variables: the visual quality and the audio quality.

Dataset ITS 2010 yields the worst outlier (26 subjects, $\Delta S_{CI} = 1.3$), and ITS AV-Sync 2010 yields the second worst outlier (16 subjects, $\Delta S_{CI} = 1.3$). These two experiments explored the relationship between audio quality and video quality on the overall audiovisual quality. ITS AV-Sync 2010 added delay as another further variable. These two experiments asked subjects to rate media that in some cases mingle very different impairment levels (e.g., high quality video with low quality audio). These outliers are explained by experiment design. Perhaps more surprising is that the other subset of ITS AV-Sync 2010 was not an outlier (12 subjects, $\Delta S_{CI} = 1.0$).

As expected, tests with more subjects have tighter distributions of votes and this enables smaller differences in MOS to become significant. Overall, we observe the following trend connecting ΔS_{CI} (MOS difference for which 95% of stimulus pairs are significantly different) and the number of subjects:

- $\Delta S_{CI} = 0.5$ 24 subjects
- $\Delta S_{CI} = 0.7$ 15 subjects
- $\Delta S_{CI} = 1.1$ 9 subjects
- $\Delta S_{CI} = 1.5$ 6 subjects

These values provide a lower limit to expected performance, based on well-designed experiments conducted by SDOs. Deviations from this ideal produce larger values of ΔS_{CI} for the given numbers of subjects. The distribution of data in Figure 3 shows variation among datasets, resulting in slightly higher values for ΔS_{CI}. Put another way, we expect $\Delta S_{CI} \geq 0.5$ for 24 subjects; and we will need from 24 to ≈34 subjects to achieve $\Delta S_{CI} = 0.5$. As a rule of thumb, the actual CI will probably fall between the value given above and the value for the next lower category of subjects (e.g., 24 subjects yield ΔS_{CI} from 0.5 to 0.7).

The ΔS versus π curves also indicate a lower bound around $\Delta S = 0.2$, below which quality differences are unlikely to be detected, regardless of the number of subjects used. The ΔS_{CI} versus number of subject curves indicate a soft lower bound. Datasets with $\Delta S_{CI} \leq 0.3$ are very difficult to achieve. The number of subjects would need to be increased dramatically above numbers typically used today. Janowski and Pinson [37] observe that such small differences are masked by the noise (error) associated with subject rating behaviors. Note that the only dataset with $\Delta S_{CI} \leq 0.3$, CCRIQ, provides 213 subject ratings for each medium.

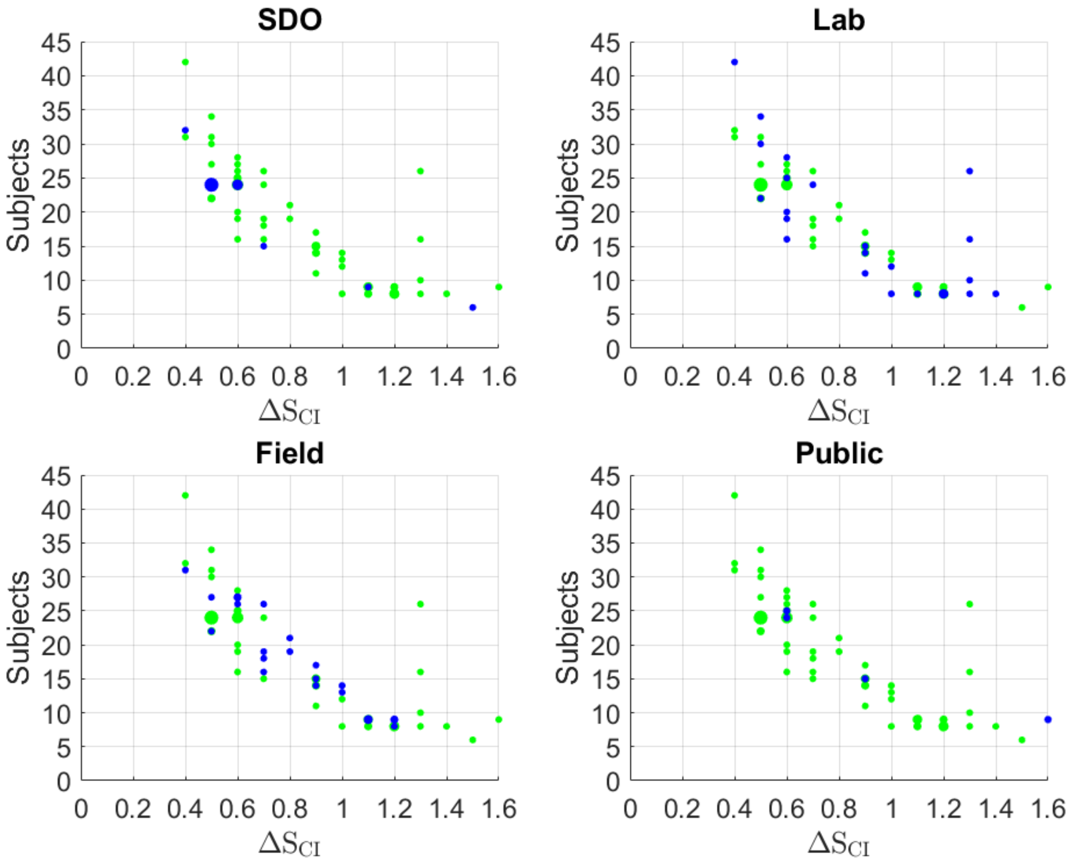

Figure 4. Impact of environment and experiment design on the relationship between ΔS_{CI} and number of subjects in the dataset.

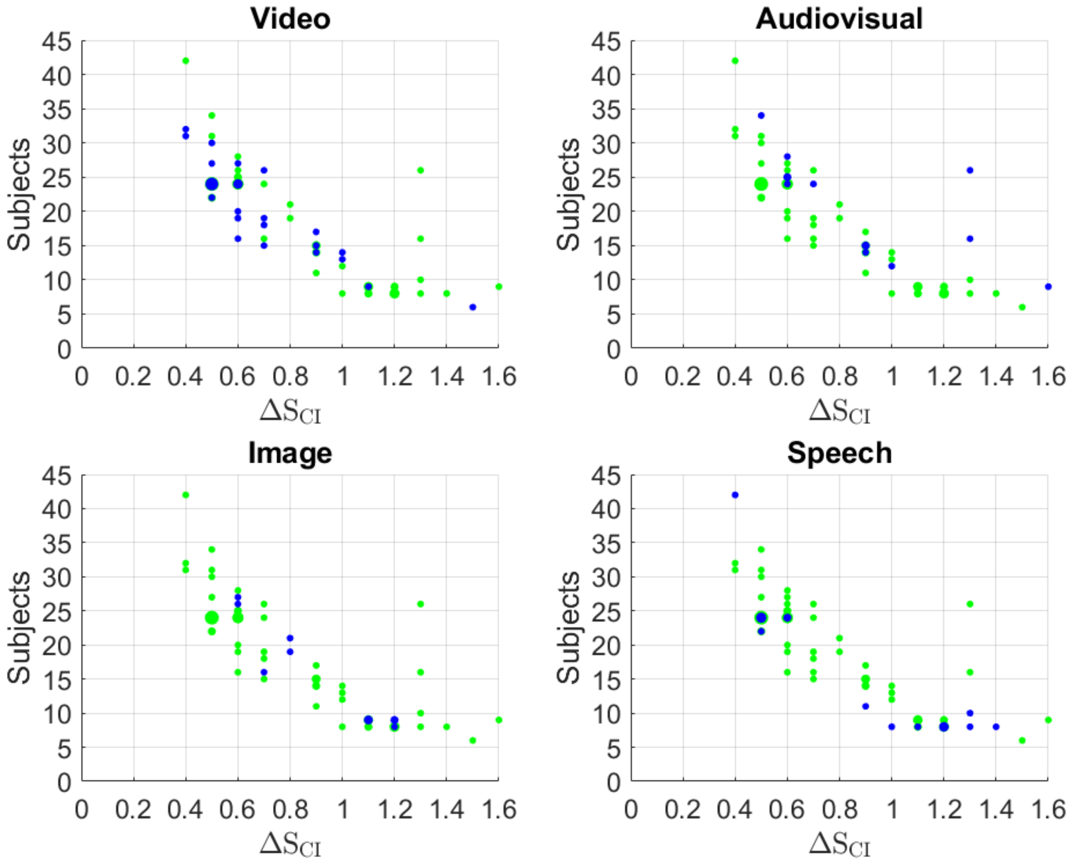

Figure 5. Impact of media type on the relationship between ΔS_{CI} and number of subjects in the dataset.

4.3 Reduced Range

We know that there is a phenomenon close to the end of a rating scale, where MOSs compress and the standard deviation of ratings decreases. How much does this impact ΔS_{CI}?

When we aggregate data from the VQEG HDTV and VQEG Hybrid datasets, we notice that $\Delta S_{CI} = 0.5$. We will now limit the stimuli to high MOSs (i.e., MOS > 4.0), medium MOSs ($2.0 < \text{MOS} \leq 4.0$), and low MOSs ($\text{MOS} \leq 2.0$). Each of these three ranges yields $\Delta S_{CI} = 0.5$. We note that, when datasets are considered individually, the low and high range occasionally have $\Delta S_{CI} = 0.4$. Thus, the expected compression exists; it is simply smaller than the granularity of our measurements.

4.4 Speech Conditions

Figure 6 examines speech datasets where ΔS_{CI} is calculated for conditions instead of files. In these scatter plots, the per-condition speech data are plotted in blue. The other data, including the per-file speech datasets, are plotted in green. This must be plotted on a logarithmic y-axis, due to the very large number of subjects. Three of these are crowdsourcing tests (d401, d501, and d701). For the private speech datasets, Table 2 presents both per-file and per-condition data. Figure 6 indicates there is likely a lower limit to condition ΔS_{CI} of ≈0.3.

Table 2. Speech ΔS_{CI}, File versus Condition

Dataset	**Subjects**	**Evaluate**	**ΔSCI**
d401 **d501** **d701**	227 115 144	**Condition** **Condition** **Condition**	0.3 0.3 0.4
Private Speech Dataset #1	≈11 ≈22 43 344 to 440	File File File **Condition**	0.9 0.5 0.4 0.3
Private Speech Dataset #2	8 512	File **Condition**	1.4 0.3

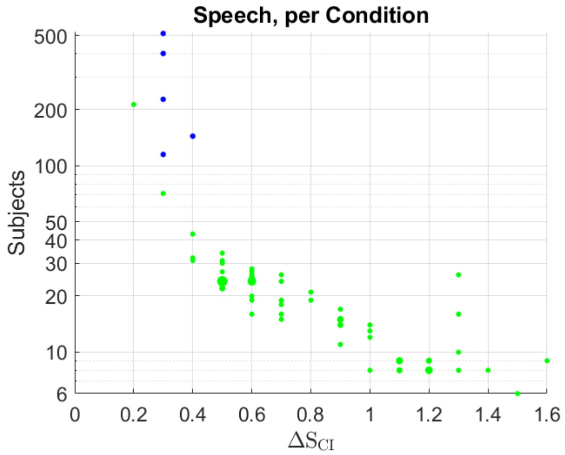

Figure 6. Impact of speech per-condition evaluation the relationship between ΔS_{CI} and number of subjects in the dataset.

4.5 ΔS_{CI} for DSCQS, ACR 100-level, and CCR

We have individual subject ratings from very few experiments conducted with methods other than ACR 5-level. This section reports those results.

Table 3 lists ΔS_{CI} for datasets that use a continuous, 100-level scale. The "SS ΔS_{CI}" column contains data for the single stimulus ratings that constitute intermediate results of the DSCQS ratings. The first number is ΔS_{CI} calculated from the source ratings; and the second number is ΔS_{CI} calculated from the processed video sequence ratings. The ΔS_{CI} for ACR 100-level and DSCQS cannot be directly compared, as we will explain in Section 4.6.

Table 3. ΔS_{CI} for DSCQS and ACR 100-level

DSCQS Dataset	Subset	Scale Used	Subjects	SS ΔS_{CI}	ΔS_{CI}
VQEG FRTV Phase I	Low 525-line	75%	90	5, 5	6
	Low 625-line	56%	79	5, 6	8
	High 525-line	51%	90	4, 5	5
	High 625-line	39%	90	5, 6	6
VQEG FRTV Phase II	625-line	55%	27	6, 8	7
	525-line	57%	64	4, 5	5
ACR 100-level Dataset	**Subset**	**Scale Used**	**Subjects**	**ΔS_{CI}**	
Private Video Dataset #1	Experts	83%	15	11	
	Crowdsource	83%	61	10	

ITU-T Rec. P.Sup23, EXP 2 provides subject data for the CCR method, using speech files and 24 subjects. Subjects enter data on a discrete [-3..3] scale, but after the random presentation is removed, most of the data is on a [0..3] scale. Each of these three datasets yields $\Delta S_{CI} = 0.8$.

Public Safety #1 and Public Safety #2 used both an ACR 5-level scale and a Boolean acceptability scale that can be interpreted as ACR 2-level with range [0..1]. The Boolean data yielded $\Delta S_{CI} = 0.4$, which is 40% of the available scale. By comparison, both datasets have $\Delta S_{CI} = 0.6$ for the ACR 5-level scale, with 16 and 19 subjects respectively. These values are within expectations and are included in the prior analyses. Despite our warnings not to compare the ΔS_{CI} from different methods, it is safe to conclude that the Boolean rating method is flawed (i.e., significant differences are difficult to detect). The team that conducted this study changed tactics and developed a subjective method that later became ITU-T Rec. P.912, “Subjective video quality assessment methods for recognition tasks.”

4.6 Relationship between Method, MOS Range, and Confidence Interval

Most of the datasets in Table 1 use the 5-level ACR method. This is the most popular method today, largely due to a compelling study by Nippon Telegraph and Telephone (NTT) [39]. NTT evaluated the following subjective test methods: ACR with a 5-level scale, ACR with an 11-level scale, Double Stimulus Impairment Scale (DSIS), Double Stimulus Continuous Quality Scale (DSCQS), and Subjective Assessment of Multimedia Video Quality (SAMVIQ). Regarding rating scales and ignoring other differences,

- ACR 5-level and DSIS use a discrete [1..5] scale
- ACR 11-level uses a discrete [1..11] scale
- DSCQS and SAMVIQ use a continuous [0..100] scale

Reference removal is a technique where the original (reference) stimuli are included in the subjective test, and their ratings are subtracted from the ratings of impaired stimuli in post-processing. The DSCQS method always implements reference removal. ACR and SAMVIQ can be implemented with reference removal using a Hidden Reference (HR). The HR variants were included in the NTT study (i.e., ACR-HR 5-level, ACR-HR 11-level, and SAMVIQ-HR). Note that reference removal occasionally yields values outside the range of the rating scale.

In the NTT study, 48 subjects evaluated the same 42 videos with these eight different methods. The subjective methods were compared in terms of the range of MOSs (MOS_{range}), correlation of MOSs, total assessment time, ease of evaluation (assessed via a questionnaire), and mean CI (MCI). The absolute value of the Pearson and Spearman correlations between each pair of methods was always 0.97 or higher. We recommend reading this landmark study to understand their other analyses. This report only considers their MMOS range and MCI results.

This report evaluates subjective CIs (ΔS_{CI}) based on 95% of stimulus pairs, while the NTT study reports the average (MCI). This difference is unimportant as we are only concerned with the relative impact of rating method on ΔS_{CI}. The two pieces of information we can garner from the NTT study are (a) how much of the available ratings scale was used by each method, and (b) how the CIs of other methods compare to the CIs of the ACR 5-level method.

Table 4 presents the NTT study's range and MCI data, formatted as follows:

- "Scale used" is the percentage of the rating scale used by the actual subjects, compared to the usable rating scale
- "MCI_{norm}" is MCI normalized by the scale's range (i.e., $MCI_{norm} = MCI/MOS_{range}$)
- "Relative MCI_{norm}" is MCI_{norm} divided by ACR 5-level's MCI_{norm}

MCI_{norm} is copied directly from [39], where it is provided with the low precision shown in Table 4. We calculated "Relative MCI_{norm}" from the MCI and MOS_{range} values provided in [39], which are provided with higher precision. If "Relative MCI_{norm}" is calculated as a ratio of MCI_{norm} values in Table 4, the results will differ due to rounding error. All percentages are rounded to 5%; we do not believe added precision is justified, due to rounding in [39] and the limited size of this dataset.

Table 4. Impact of Subjective Test Method on CI

Method	**Scale Used**	**MCI_{norm}**	**Relative MCI_{norm}**
ACR 5-level	75%	0.07	100%
ACR-HR 5-level	75%	0.09	120%
ACR 11-level	60%	0.08	110%
ACR-HR 11-level	55%	0.10	135%
DSCQS	55%	0.09	115%
DSIS	90%	0.07	95%
SAMVIQ	60%	0.07	95%
SAMVIQ-HR	60%	0.08	110%

Table 4 provides critical context, as we examine ΔS_{CI} for datasets that do not use the ACR 5-level method. Without this data, we would naively expect subjects to use these other scales similarly to the ACR 5-level scale. Instead, we must acknowledge that the differences in how subjects use each rating scale impact the range of MOS values and therefore the relative meaning of the resulting CIs. Any comparisons between CIs from different methods will require a bridge, similar to a Rosetta stone.

The CI normalization technique described above does not solve this problem for the general case. Increasing or decreasing the quality range in an experiment, without changing the method, does not influence the precision of subject ratings.

In the present context, the most important conclusion in the NTT report is that all subjective methods produce CIs that are similar to or slightly worse than those of ACR 5-level ratings. Small improvements may be possible, for example when using SAMVIQ or DSIS. However, the variations in Relative MCI_{norm} in Table 4 are within the distribution we observe in Figure 3. Therefore, unknown factors may be involved. Theoretically, if the rating method suits the experiment particularly well, the resulting CI should decrease. This would explain the diversity of opinions and experiences shared anecdotally among experts.

4.7 Conclusions

Our analysis of subjective tests gives the following relationships between the number of subjects and our newly-defined MOS CI, ΔS_{CI}, for well-designed and carefully executed subjective tests using the 5-level ACR scale:

- 24 subjects yield a MOS CI of $\Delta S_{CI} = 0.5$
- 15 subjects yield a MOS CI of $\Delta S_{CI} = 0.7$
- 9 subjects yield a MOS CI of $\Delta S_{CI} = 1.1$
- 6 subjects yield a MOS CI of $\Delta S_{CI} = 1.5$

Here the MOS CI is the difference in MOS values at which 95% of the pairs will be statistically different (according to the Student's *t*-test using a 95% confidence level).

These values indicate the expected performance of a well-designed and carefully conducted subjective test. Deviations from this ideal yield larger ΔS_{CI} or require more subjects to obtain the specified ΔS_{CI}. Unexplained factors in the experiment design and implementation may produce CIs up to the next category of subjects (e.g., 24 subject tests typically have ΔS_{CI} between 0.5 and 0.7). Other methods and scales will yield CIs that are similar or slightly larger, when adjusted for differences in rating scales and subject rating behaviors (e.g., how much of the scale is used). Reference removal, when conducted as post-processing, yields slightly larger CIs (ΔS_{CI}).

5. DECISIONS REACHED BY DIFFERENT LABS

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 at which different conclusions are reached by different labs. See [33] and [37] for a more in depth analyses of lab-to-lab comparisons. The intermediate data for this section appears in Appendix B.

Given a subjective test, 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 MOS values and 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 then tally the frequency of the four possible classification types, defined below. The confusion matrix is presented in Table 5.

- *Agree Ranking* Both labs conclude that quality of **A** is better than the quality of **B**, or both labs conclude that the quality of **A** is worse than the quality of **B**
- *Agree Tie* Both labs conclude that **A** and **B** have statistically equivalent quality
- *Unconfirmed* One lab can rank order the quality of **A** and **B** but the other lab concludes that **A** and **B** have statistically equivalent quality
- *Disagree* The labs reach opposing conclusion on the quality ranking of **A** and **B**

Table 5. Confusion Matrix For Different Subjective Test Labs

		Subjective Test 1		
		Better	**Equivalent**	**Worse**
Subjective Test 2	**Better**	Agree Ranking	Unconfirmed	Disagree
	Equivalent	Unconfirmed	Agree Tie	Unconfirmed
	Worse	Disagree	Unconfirmed	Agree Ranking

5.1 Decisions Reached by Multiple Labs Conducting the Same Experiment

Table 1 provides us with several datasets that were conducted identically at multiple labs: AGH/NTIA/Dolby, ITS4S, CCRIQ, Private Speech Dataset #3, VQEG MM2, VQEG FRTV Phase I, VQEG FRTV Phase II 525-line, RRNR-TV, and the VQEG Hybrid HDTV Common Sets. These datasets allow us to compute the frequency of each outcome. Data from the Hybrid HDTV common sets are aggregated into a single measurement, because the small number of stimuli (24) yields unstable measurements (i.e., extreme results are more likely to be observed due to the sample size).

VQEG FRTV Phase I has two severe disadvantages. First, each dataset spans a narrow range of quality. VQEG uses these datasets as an illustrative example of a flawed experiment design. Despite being conducted by an SDO, we do not consider VQEG FRTV Phase I to be a well-designed experiment. Second, the ratings include scoring inversion errors. Subjects marked ratings for both the original and impaired sequence at the same time on paper scoring sheets. Accidents occurred where subjects switched their ratings.[5] These errors can only be reliably detected and removed from the data when the quality of the PVS is much lower than the quality of the SRC. We will retain all ratings, including such obvious errors.[6]

The VQEG FRTV Phase I dataset also has several advantageous characteristics. VQEG FRTV Phase I contains four subjective tests: 525-line low, 525-line high, 625-line low, and 625-line high. "525-line" means the test contains NTSC format standard definition video, and "625-line" means the test contains PAL format standard definition video. "Low" means the test contains low quality videos, and "high" means the test contains high quality videos. For each, we have ratings from 67 or 70 subjects divided among 4 labs. The rating method was DSCQS, so we have easy access to two ratings for each stimulus[7] (PVS and DOS). The SRC ratings will not be used, because the range of quality is abnormally narrow.

Therefore, we will divide our datasets into two subsets. First is the VQEG FRTV Phase I datasets: 24 lab-to-lab comparisons that will be treated separately. Second is the remainder, 66 lab-to-lab comparisons that will be referred to as well-designed experiments. This subset is heavily influenced by the VQEG MM2 dataset, which contributes 45 of the 66 lab-to-lab comparisons.

Figure 7 shows distribution of incident rates for the confusion matrix identified in Table 5. The blue histogram depicts the results from well-designed experiments (66 lab-to-lab comparisons). The tan overlay depicts the results from the VQEG FRTV Phase I dataset, which has a narrower quality range (24 lab-to-lab comparisons). Dark brown indicates areas where the tan and blue histograms overlap. From the *agree ranking* (upper left), *agree tie* (upper right), and *unconfirmed* (lower left) distributions, we observe that a narrow range of quality causes the *agree ranking* rate to fall, while the *agree tie* and *unconfirmed* rates rise. The *disagree* incidents are not impacted by the narrow range of quality (lower right). One of the 24 lab-to-lab comparisons is an outlier, with a 1.84% *disagree* rate.[8]

[5] In the DSCQS method, subjects view the SRC and PVS (in random order) and then rate them. Scoring inversions were found in the pretest data. Upon reviewing the ratings and videos, we found a very low quality PVS that was rated "high quality," while the associated SRC was rated "low quality." The subject confirmed that this was a scoring error.

[6] When obvious inversions are removed, using various thresholds, the results in this section do not substantially change.

[7] SRC ratings are available but will not be analyzed. Their range of quality is atypical.

[8] This outlier may be caused either by subtle rating inversions (e.g., for high quality PVSs) or differences in how the labs implemented the DSCQS method. Discussions among the labs identified differences in instructions, training, and test implementation.

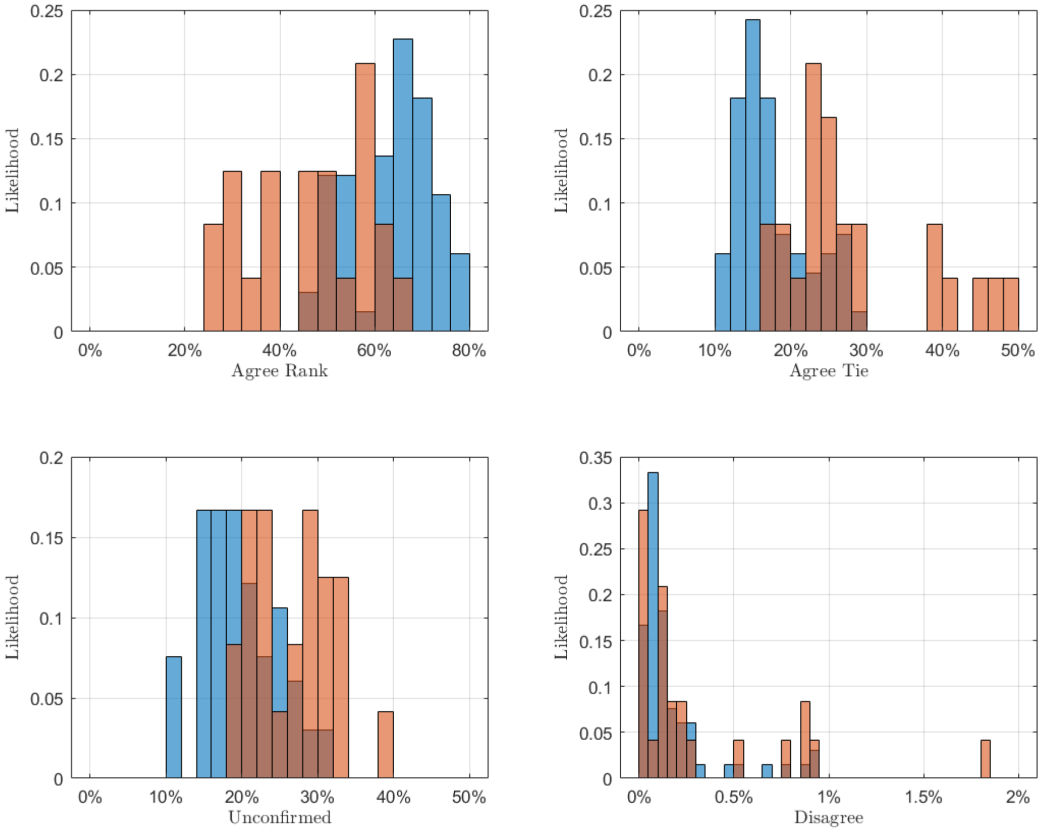

Figure 7. These histograms show the likelihood of the conclusions reached when two labs perform the same subjective tests.

Table 6 contains the overall trend of well-designed experiments. Notice that the *disagree* incidence is always less than 1%. Table 7 contains the overall trend when the subjective test contains a narrow range of quality, based on VQEG FRTV Phase I. Experiments with a narrow range of quality may exceed the *agree ranking*, *agree tie*, and *unconfirmed* rates of a well-designed experiment. Specifically, *agree ranking* decreases, while *agree tie* and *unconfirmed* increases. Except for *disagree*, all of these incidence rates are strongly influenced by the range of quality in the experiment, as we can clearly see from Figure 7.

Table 6. Lab-to-Lab Classification Incident Rates For Well Designed Tests

Outcome	**Mean**	**Mode**	**Min**	**Max**
Agree Rank	63%	67%	47%	77%
Agree Tie	17%	14%	10%	29%
Unconfirmed	19%	18%	10%	31%
Disagree	0.17%	0.06%	0%	0.95%

Table 7. Lab-to-Lab Classification Incident Rates For Narrow Range of Quality

Outcome	**Mean**	**Mode**	**Min**	**Max**
Agree Rank	45%	37%	24%	65%
Agree Tie	28%	25%	17%	48%
Unconfirmed	26%	21%	19%	38%
Disagree	0.32%	0.02	0%	1.82%

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 fairly strong and reliable trade-off between *agree ranking* and *agree tie.* This trade-off is shown in Figure 8 and described by:

$$\sqrt{agree\ ranking} \approx\ -1.2 \times agree\ tie + 1.0 \tag{2}$$

where *agree ranking* and *agree tie* are expressed as fractions. The square root is needed to remove a non-linearity.

Motivated by (2) we now define a new statistic *concur* as:

$$concur = \sqrt{agree\ ranking} + 1.2\ \times agree\ tie \tag{3}$$

Note that *concur* takes a value of 1.0 when the approximation in (2) 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 Figure 8. 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 Figure 8.

Concur is a single figure-of-merit for comparing the results of two tests (subjective-to-subjective, or subjective-to-objective). 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.

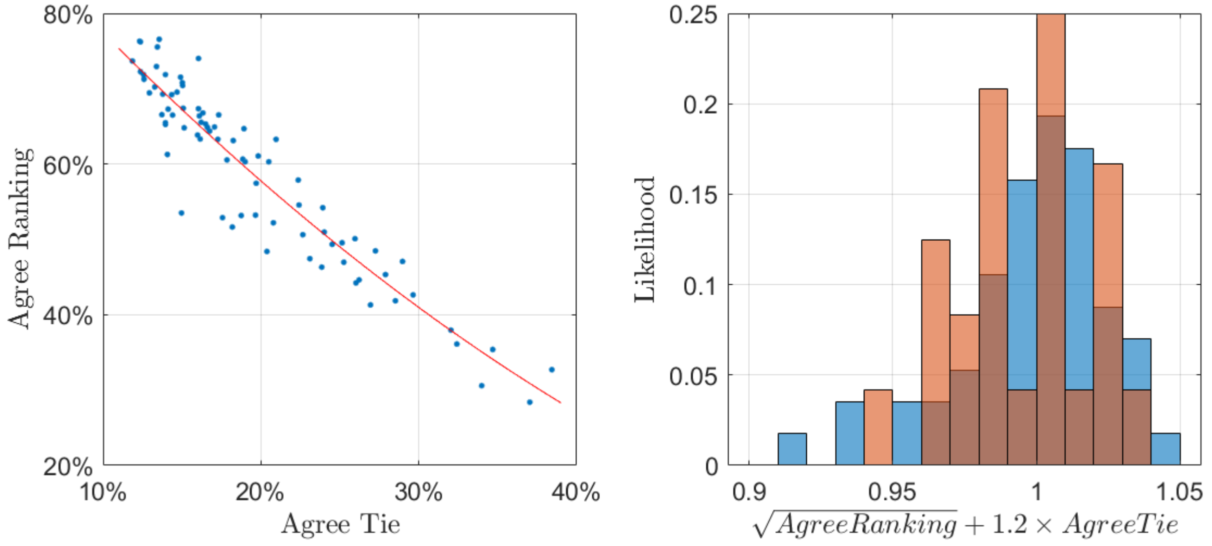

Figure 8. The scatter plot (left) shows the relationship between *agree ranking* and *agree tie* incidence rates, with (1) plotted in red. The histogram (right) shows the distribution of (2) for well-designed tests in blue and tests with a narrow range of quality in tan.

5.2 Decisions Reached When Method Changes

The UPM-Acreo dataset has three sessions that rate the same stimuli using different rating methods: (a) video only ACR, (b) video only CIETI, and (c) audiovisual CIETI, which we will refer to as CIETI(AV). Audio was not impaired in this experiment. This data provides us with insights on how the incidence rates change when the subjective testing method is not held constant.

The first three lines of Table 8 list the incidence rates for the UPM-Acreo datasets. The CIETI(AV) vs CIETI and CIET vs ACR incidence rates match our expectations from Section 4. However, CIETI(AV) vs ACR has an unusually high *disagree* incidence rate of 1.43%. This likely indicates genuine disagreement caused by the combined impact of the different rating methods, confounded by the influence of audio quality on CIETI(AV) MOSs. The range of quality present in the original video impacts the overall MOS, as shown in [14].

The next three lines of Table 8 list the incidence rates for Private Video Dataset #3, Netflix Quality Variation 2017. The videos were split into overlapping pools of 180 videos and rated by different subjects. We will not evaluate the overlapping stimuli separately, due to the small number of videos (40). Each subject pool can be sub-divided by monitor type and rating method, for a total of four subsets. While some subjects used the standard ACR method, other subjects continuously rated each video with the single stimulus continuous quality evaluation (SSCQE) method, and then provided an overall score on the ACR 5-level scale. Thus the SSCQE ratings (not analyzed) could have impacted the ACR ratings. However, Table 8 indicates that none of these variables impacted the *disagree* incidence rate. In particular the *disagree* incidence rates fall within our expectations for a well-designed and carefully conducted subjective test on the 5-level ACR scale.

The Public Safety #1 and #2 datasets had each subject rate each video on two different scales. The first scale was 5-level ACR. The second scale was a Boolean scale, measuring whether the system was acceptable for first responder applications. Both questions were asked at the same time. As with the prior two experiments, the *disagree* incidence rate in Table 8 is well above 1%.

Unfortunately, we were not able to obtain datasets where subjects rated the same stimuli using two different standard methods (e.g., ACR 5-level vs 11-level). Such experiments are extremely rare. Generally, *disagree* incidence rates above 1% should be investigated, as these may indicate a major difference in the subjective test method or an error in data processing. For example, these incidence rates can only be calculated when all stimuli at one lab are rated by the same subjects. When this rule is broken, then *disagree* incidence rates skyrocket to 45%. This problem is caused by different subject offsets, as explained in [37].

Table 8. Lab-to-Lab Classification Incident Rates Confounded by Different Rating Methods

Comparison	**Stimuli**	**Subjects**		**Agree Ranking**	**Agree Tie**	**Unconfirmed**	**Disagree**
UPM-Acreo CIETI(AV) vs CIETI	132	21	22	61%	18%	21%	0.12%
UPM-Acreo CIETI(AV) vs ACR	132	21	20	53%	16%	30%	1.43%
UPM-Acreo CIETI vs ACR	132	22	20	56%	20%	24%	0.16%
Netflix Quality Variation 2017, Different Monitor, ACR	180	25	25	74%	13%	13%	0.04%
	180	25	25	76%	14%	11%	0.01%
Netflix Quality Variation 2017, Different Monitor, ACR/SSCQE	180	26	26	74%	13%	13%	0.03%
	180	23	25	74%	13%	12%	0.01%
Netflix Quality Variation 2017, ACR vs ACR/SSCQE	180	25	26	73%	14%	13%	0.08%
	180	25	23	74%	14%	11%	0.04%
	180	25	26	76%	12%	12%	0.04%
	180	25	25	76%	13%	11%	0.02%
PS1, ACR vs Boolean	400	16	16	47%	19%	26%	7.62%
PS2, ACR vs Boolean	576	19	19	45%	18%	29%	7.78%

5.3 Conclusions

Our analysis of subjective tests reaches the following conclusions for agreement or disagreement between two well-designed and carefully executed subjective tests:

- *disagree* is ≤ 1% and typically ≈0.17%
- *unconfirmed* is ≤ 31% and typically ≈20%
- *concur* ranges from 0.91 to 1.05

- *disagree and concur* are minimally influenced by the range of quality in the dataset

Outliers may exceed the above limits.

6. AD-HOC EVALUATION AND PILOT TESTS

Pragmatically, metrics and subjective tests will never fully replace ad-hoc quality assessments. For example, two engineers encode their favorite test sequence at several different bit-rates, watch the resulting videos, and decide which version looks best. The management team receives hardware video systems on loan from competing vendors, transmits their favorite test sequences through them, and makes a purchase decision. Ad-hoc assessments dispense with the scientific method in favor of speed.

A bit of subjective testing knowledge lets us add a minimum of structure and formality to an ad-hoc evaluation. Two or three researchers watch and/or listen to each stimulus in a proposed experiment and write down their ratings. This produces MOSs with low precision that are compared deterministically (>, <, =, ≈). These MOSs can answer simple questions like whether a subjective test has too many high quality stimuli or whether the source video pool has enough variability to produce an interesting variety of quality responses from a video codec.

From a statistical analysis standpoint, pilot studies are similar to ad-hoc assessments. ITU-T Rec. P.913 recommends 8 to 12 subjects for pilot studies to indicate trending. As we observed in Section 4, this few subjects makes a Student's *t*-test undesirable, because most of the interesting comparisons will produce no significant difference in quality. Thus, pilot study MOSs are often evaluated deterministically.

We would like to characterize the accuracy of these conclusions, to the extent possible, to serve as another reference point.

6.1 Simulating Conclusions Reached by Ad-Hoc Tests

While we do not have data for ad-hoc assessments, we have suitable data to simulate likely behaviors: the VQEG FRTV Phase I dataset. All of the characteristics that we previously identified as disadvantages will help us simulate ad-hoc assessments. The narrow range of quality, continuous scale, and occasional rating error seem appropriate and realistic for ad-hoc decisions. 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 most interesting questions often involve a narrow range of quality. Section 4.6 indicates that DSCQS produces CIs only slightly worse than ACR 5-level, so the VQEG FRTV Phase I data is also suitable to simulate pilot tests.

We want to establish the relationship between the decisions reached by an ad-hoc test and the decision reached by a formal subjective test. We have three interesting outcomes, defined below. We will ignore ties, since our ad-hoc or pilot test data will seldom produce identical MOSs.

- *Correct ranking* Both conclude that quality of **A** is better than the quality of **B**

- *False distinction* The ad-hoc test can rank order the quality of **A** and **B** but the subjective test cannot

- *False ranking* The ad-hoc test and subjective test reach opposing conclusion on the quality ranking of **A** and **B**

To frame discussions, we will establish six performance levels: ad-hoc assessments 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.

As noted in Table 1, the VQEG FRTV Phase I dataset contains four subjective tests. Each test has 67 or 70 subjects total, split among four labs. We will simulate the one person ad-hoc test as follows. One of the labs is chosen to provide ad-hoc data. The "ground truth data," a formal subjective test, will be simulated by drawing 24 subjects at random from the other three labs. We will compare the conclusions reached by that subject deterministically (**A** > **B**, **A** = **B**, and **A** < **B**) with conclusions reached by the "ground truth data" using a Student's *t*-test. We will repeat this procedure for each subject in the ad-hoc lab. Then, this process will be repeated with each of the other three labs providing ad-hoc subjects.

The 2, 3, 6, 9, and 12 person ad-hoc and pilot tests will be simulated with the same procedure, but by drawing that number of subjects at random from the ad-hoc lab and averaging their votes. This random choice and evaluation will be repeated 25 times for each ad-hoc lab.

This entire process will be repeated for all four of the VQEG FRTV Phase I subjective tests (i.e., 525-line low quality, 525-line high quality, 625-line low quality, and 625-line high quality).

6.2 Ad-Hoc and Pilot Test Analyses

Figures 9 and 10 show the *correct ranking* and *false distinction* rates, respectively. The data is divided into three categories, based on the range of MOSs spanned by the VQEG FRTV Phase II. **Wide** includes the low quality 525-line test, which spans 75% of the [0..100] scale. To put this into perspective, the NTT experiment spans 90% of the DSCQS scale (see Table 4). **Medium** includes both the high quality 525-line and low quality 625-line tests, which span similar ranges (51% and 56% respectively). **Narrow** includes the high quality 625-line test, which spans 39% of the [0..100] scale. In ACR 5-level scale language, **narrow** covers from "excellent" to part way between "good" and "fair." The colors used to indicate these three categories are shown in Figure 10. On these plots, the x-axis shows the likelihood in percent of one of the three quality comparison outcomes listed above. The y-axis shows the likelihood (over simulation runs) of a particular incidence rate.

Figure 9 shows the estimated distribution of *correct ranking* incident rates for ad-hoc assessments and pilot tests. As expected the likelihood of *correct ranking* positively correlates with both the number of subjects and the range of stimulus quality. The worst case is a single person evaluating stimuli with similar quality levels, where we expect ≈36% of stimulus pairs will be correctly ranked. We do not have data for experts (e.g., golden eyes or golden ears), who may have higher success rates. The best case is a 6 to 12 subject pilot study with a wide range of quality, where we expect that ≈72% of stimulus pairs will be correctly ranked.

Figure 10 shows the estimated distribution of *false distinction* incidence rates for ad-hoc assessments and pilot tests. The average responses as follows: **wide** ≈ 25% *false distinction*, **medium** ≈ 35% *false distinction*, and **narrow** ≈ 51% *false distinction*. The levels for 1, 2, 3, 6, 9, and 12 subjects are not shown, because the *false distinction* incidence rate does not depend on the number of subjects (i.e., very similar histograms).

Figure 11 shows the estimated distribution of *false ranking* incidence rates for ad-hoc assessments and pilot tests. The average *false ranking* incidence rates are also given in Table 9. Maximum observed rates are three to four times higher than the average rates, with a long tail that indicates higher values may occur. The sub-categories of the VQEG FRTV Phase I tests are not shown, because the range of quality is not influential. These *false ranking* rates are much higher than those observed in lab-to-lab comparisons between subjective tests, which are typically ≈0.06%.

Subjects' scoring includes a random component. This is expected behavior that must be accepted, not a flaw or fault that can be eliminated [37]. All the complexities of probability theory have been set aside to perform a deterministic analysis that does not accommodate the complexities of random processes. The natural consequence is that small numbers of subjects produce results that often differ from the results of full subjective tests.

Table 9. Estimated False Ranking Rates for Ad-hoc Assessment and Pilot Tests

	1 Person	**2 People**	**3 People**	**6 Subjects**	**9 Subjects**	**12 Subjects**
Average	11.4%	8.5%	6.8%	4.4%	3.5%	3.0%
Range	3% to 30%	2% to 26%	1% to 21%	1% to 17%	1% to 13%	0% to 10%

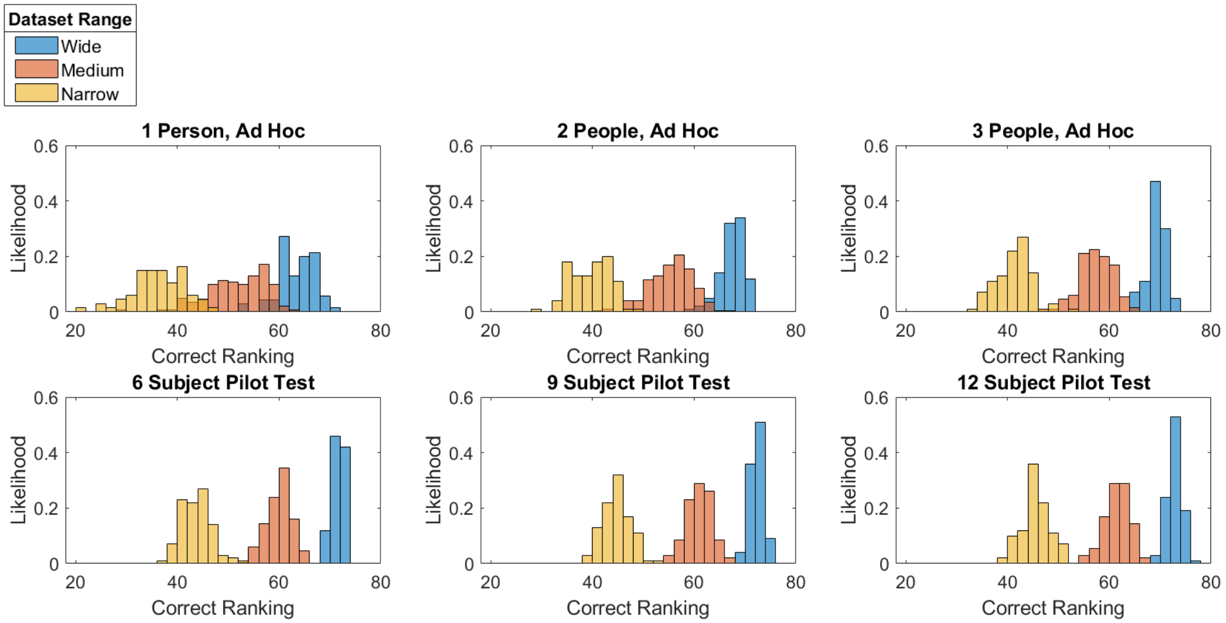

Figure 9. Correct ranking incidence rates for ad-hoc and pilot tests.

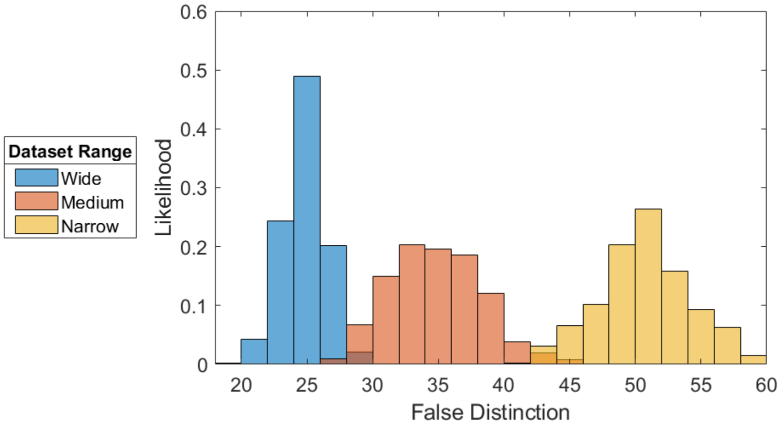

Figure 10. False distinction ranking incidence rates for ad-hoc and pilot tests.

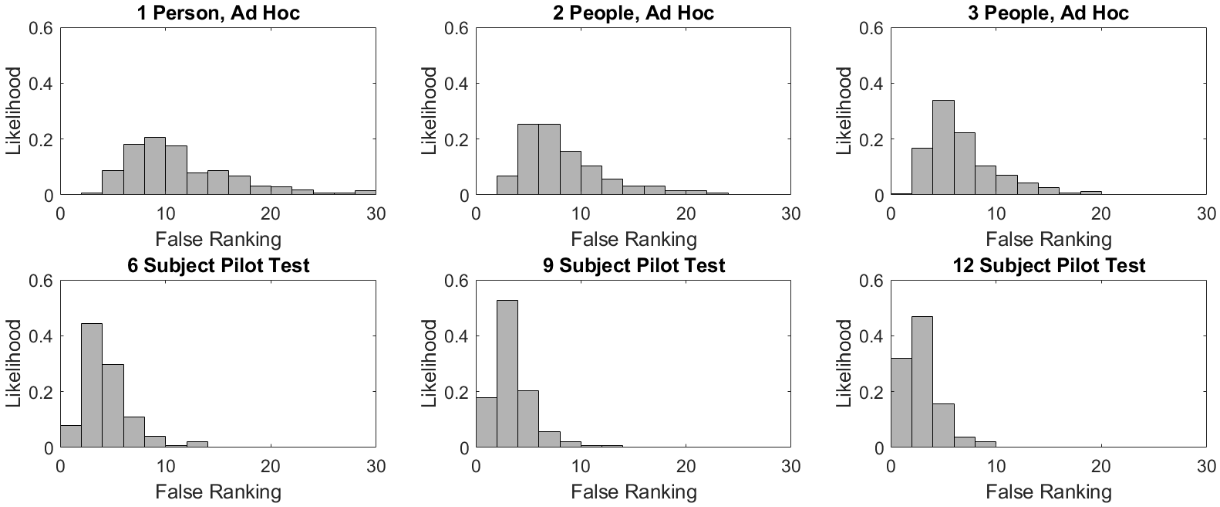

Figure 11. False ranking incidence rates for ad-hoc and pilot tests.

7. METRIC CONFIDENCE INTERVALS AND AD-HOC TEST COMPARISON

7.1 Confidence Interval Theory

Objective quality metrics can be considered as substitutes—or proxies—for subjective quality ratings. For this reason, we typically denote the metric value for a certain stimulus, **A,** as ($\widehat{MOS}_A$).

Let us define ΔM as the distance between the metric value for stimulus **A** ($\widehat{MOS}_A$) and metric value for **B** ($\widehat{MOS}_B$). We want to establish the relationship between the decisions reached by the metric and the decision reached by a subjective test, as a function of ΔM. We have five possible outcomes, defined below and in Table 10. For simplicity, we will assume that ($\widehat{MOS}_A \geq \widehat{MOS}_B$).

- *Correct ranking* Both subjective testing and objective metric conclude that quality of **A** is better than the quality of **B**
- *Correct tie* Both conclude that **A** and **B** have statistically equivalent quality
- *False tie* The subjective test can rank order the quality of **A** and **B** but the metric cannot
- *False distinction* The metric rank orders the quality of **A** and **B** but the subjective test does not
- *False ranking* The metric and subjective test reach opposing conclusions on the quality ranking of **A** and **B**

Table 10. Confusion Matrix Between Subjective Test Results and Metric Results

		Subjective Test		
		Better	**Equivalent**	**Worse**
Metric	**Better**	Correct ranking	False distinction	False ranking
	Equivalent	False tie	Correct tie	False tie
	Worse	False ranking	False distinction	Correct ranking

Our first goal is to calculate the *ideal CI*, which we will define as the value of ΔM that yields the same error rates as a well-designed subjective test with 24 subjects that is conducted in a controlled environment and adheres to ITU-R BT.500, ITU-T Rec. P.913, or ITU-T Rec. P.910. Based on results presented in Section 3, we will limit the *false ranking* rate to 1%. The *unconfirmed* incidence from Table 5 includes both *false tie* and *false distinction* categories, so the 20% *unconfirmed* criteria for a typical subjective test must be divided by two. Thus, we will limit the *false distinction* rate to 10%

The *ideal CI* has two problems. First, metrics have a higher relative rate of *false ranking* compared to *false distinction*, due to imperfect modeling of human perception. In practice, the *ideal CI* will usually only depend on the *false ranking* rate. Second, the *ideal CI* will be too large for some applications, where users are willing to tolerate more *false ranking* incidents to improve the *correct ranking* incidence rate.

Our second goal is to calculate the *practical CI*, which will use less stringent ΔM selection criteria. We will merge *false ranking* and *false distinction* into a single category. Additionally, we will loosen our *unconfirmed* criteria from the typical subjective test (20%) to the maximum *unconfirmed* rate observed for a 15 subject test (31%). *False ranking* will remain at 1%, as we have no evidence for higher rates. Thus, we will limit the sum of *false ranking* plus *false distinction* to 16.5%, which is half of the maximum *unconfirmed* rate of 31% plus 1% *false ranking* rate. These criteria are justified by observed errors between actual subjective tests. Overall, we will define *practical CI* as the smallest value for ΔM that yields error rates no greater than those found when comparing two subjective tests with 15 subjects each.

Other CIs can be chosen from examining the overall relationship between ΔM and the five incidence rates. For example, the *false ranking* rate could be limited to 2% or 4%.

We need to avoid predisposing our measurement sensitivity in favor of subjective testing. The problem, as explained in the introduction to Section 4, is that we have confounding factors:

- Source of data (subjective ratings versus objective metric)
- Type of data (collection of ratings versus $\widehat{MOS}$)
- Method of comparison (statistical test, CI, or deterministic math)

The metric's stimulus pair comparisons must use deterministic math, because we only have $\widehat{MOS}$ (i.e., the metric will not predict the distribution of ratings for a subject panel).

Therefore, the subjective data's stimulus pair comparisons must use deterministic math and a CI, instead of the Student's *t*-test or another statistical test. We will use the empirical CI of a well-designed and carefully conducted subjective test, from Section 4. If we allowed the subjective test CI to vary (e.g., calculated for each test individually), then our baseline for comparison would be unstable. Additionally, this design choice simplifies the resulting algorithm.

7.2 Algorithm to Calculate Ideal CI and Practical CI

First, let us examine the subjective data. 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. For each pair of stimuli, we will measure (MOS_A – MOS_B) and reach decisions as follows:

- If MOS_A – MOS_B > ΔS, conclude **A** is "better" than **B**
- If MOS_A – MOS_B < -ΔS, conclude **A** is "worse" than **B**
- Otherwise, conclude **A** is "equivalent" to **B**

where ΔS = 0.5 (i.e., the expected precision of a well-designed and carefully conducted subjective test with 24 subjects and the ACR method, from Section 4). A lower value for ΔS could negatively impact the metric's measured precision. If ΔS is too large, then a high performing metric will be more precise than this algorithm can measure. This problem can also be caused by using data from poorly-designed or carelessly conducted subjective tests.

Second, we will examine the metric data. Given the range of metric values, we will choose possible values for the metric CI, ΔM, based on the range of metric values divided by 100 and rounded to two significant digits. Smaller distances would claim a level of precision that cannot be justified. We will choose the same pairs of stimuli, **A** and **B**, defined above. For each pair of stimuli and for each value of ΔM, we will calculate $\widehat{MOS}_A - \widehat{MOS}_B$ and the corresponding decisions reached as follows:

- If $\widehat{MOS}_A - \widehat{MOS}_B$ > ΔM, conclude **A** is "better" than **B**
- If $\widehat{MOS}_A - \widehat{MOS}_B$ < -ΔM, conclude **A** is "worse" than **B**
- Otherwise, conclude **A** is "equivalent" to **B**

Third, we will compare the conclusions reached by the subjective data and the metric. We will compute the frequency of each outcome in Table 10, as a function of ΔM. Note that the odds of a *correct tie* and *false tie* increase as ΔM increases, while the odds of *correct ranking*, *false distinction*, and *false ranking* decrease.

Fourth, we will choose an appropriate level of ΔM based on our criterion from Section 7.1:

- *Ideal CI* is the minimum ΔM where *false ranking* ≤ 1% and *false distinction* ≤ 10%
- *Practical CI* is the minimum ΔM where *false ranking* plus *false distinction* ≤ 16.5%

Special considerations apply when using data from multiple subjective tests. First, we recommend that each dataset be weighted equally. Second, the raw MOSs must be used. The accepted practice for most statistical analyses is to remove lab-to-lab differences (e.g., as per the VQEG HDTV Superset [30]). However, our calculations of ΔS in Section 5.3 depend on the expected behaviors of subjects using a 5-level ACR scale. Dataset rescaling could impact these assumptions and therefore ΔS. Therefore, lab-to-lab differences cannot be removed from the MOSs. Third, extreme differences between datasets may cause problems. For example, a subjective test that explores professional photographs may produce the same distribution of MOSs as a subjective test that explores low quality consumer videos, but the relationship between a metric and these very different datasets would be very different.

7.3 Equivalence to a Subjective Test

We would like to determine if a certain metric has the same precision as a subjective test, when decisions are reached using *ideal CI* or *practical CI*. We can re-use the *false ranking* and *false distinction* thresholds from Section 7.2. However, we will need to establish new thresholds for the other incidence rates.

It would be natural to use *agree ranking* and *agree tie* rates to define equivalence, but they are confounded by the range of quality in the test. Using *agree ranking* and *agree tie* to define the figure-of-merit *concur* in (2) removes this confounding factor. Thus we define equivalence using *concur*. All observed values of *concur* (between subjective tests) are no less than 0.91. Thus we define a metric to be equivalent to a subjective test when it produces *concur* ≥ 0.91.

We will not place limits on the final category, *false tie*, for two reasons. First, all of our performance statistics from subjective testing include the confounding impact of the range of quality in the dataset. Pragmatically, we lack defensible limits. Second, the *false tie* rate is arguably the least offensive type of error a metric can make, and its rates are inherently limited by the other four factors.

7.4 Equating a Metric to a Number of People

Finally, we want to liken the metric to a number of people in an ad-hoc assessment or pilot test. This will let us make simple statements like, "This metric is as accurate as an ad-hoc test with two subjects." The analogy assumes the use of deterministic math to compare $\widehat{MOS}$s. The incidence rates from Section 6 give us appropriate data.

First, let us examine the subjective data. 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. For each pair of stimuli, we will measure (MOS_A – MOS_B) and reach the same decisions as described in Section 7.2:

- If $MOS_A - MOS_B > \Delta S$, conclude **A** is "better" than **B**
- If $MOS_A - MOS_B < -\Delta S$, conclude **A** is "worse" than **B**
- Otherwise, conclude **A** is "equivalent" to **B**

where $\Delta S = 0.5$, which is the expected precision of a well-designed and carefully conducted subjective test with 24 subjects and the ACR method, from Section 4.

Second, we will examine the metric data. We will choose the same pairs of stimuli, **A** and **B**, defined above. For each pair of stimuli, we will measure $\widehat{MOS}_A - \widehat{MOS}_B$ and calculate the decisions reached deterministically (better, worse, or equivalent).

- If $\widehat{MOS}_A > \widehat{MOS}_B$, conclude **A** is "better" than **B**
- If $\widehat{MOS}_A < \widehat{MOS}_B$, conclude **A** is "worse" than **B**
- Otherwise, conclude **A** is "equivalent" to **B**

This is like Section 7.2 but with ΔM set to zero. We use $\Delta M = 0$ because we want to equate the performance of deterministic metric comparisons (as shown above) to deterministic comparisons of ad-hoc MOSs (as shown in Section 6.2).

Third, we will compare the conclusions reached by the subjective data and the metric. We will compute the frequency of *false ranking*. All other outcomes will be ignored, as they are highly influenced by the range of quality examined. In addition, *false ranking* is arguably the most egregious of the three errors that can be made.

Fourth, we will equate the metric to a number of subjects using our observations from Section 6. We will use average *false ranking* incidence (e.g., *false ranking* rates of 11.4%, 8.5%, 6.8%, 4.4%, 3.5% and 3.0% for 1, 2, 3, 6, 9, and 12 subjects respectively). Our decision thresholds are half way between observed mean values. We will limit the high end of a 1 person ad-hoc assessment to 12.85% based on the distance to the 2 person ad-hoc assessment. This is significantly more conservative than the alternative value described below. We will limit the lower range of the 12 person pilot test at 0%, as this was the lowest observed value. The decisions are as follows:

- 1 person ad-hoc assessment: 12.85% ≥ *false ranking* > 9.95%
- 2 person ad-hoc assessment: 9.95% ≥ *false ranking* > 7.65%
- 3 person ad-hoc assessment: 7.65% ≥ *false ranking* > 5.60%
- 6 subject pilot test: 5.60% ≥ *false ranking* > 3.95%
- 9 subject pilot test: 3.95% ≥ *false ranking* > 3.25%
- 12 subject pilot test: 3.25% ≥ *false ranking*

The same special considerations apply when using data from multiple subjective tests. First, we recommend that each dataset be weighted equally. Second, the raw MOSs must be used.

The ranges of *false ranking* rates we observed in Section 6.2 overlap. An argument could be made for defining the 1 person ad-hoc assessment to include *false ranking* rates up to the maximum observed value of 30% *false ranking*. We could also limit the *false distinction* incidence rate, but it would be difficult to justify a threshold other than 51% or 60%. With such a high threshold, the added value would be negligible.

8. MEASURED CONFIDENCE INTERVALS FOR VQEG VALIDATED METRICS

8.1 Confidence Intervals and Subjective Test Equivalence of VQEG HDTV and Multimedia Metrics

Let us begin with MOSs and metrics evaluated by VQEG during the HDTV [30] and Multimedia [40] validation tests. These datasets and metrics provide a robust set of well-designed subjective datasets to calculate *ideal CI* and *practical CI*. Here, all metrics except PSNR are referred to by a randomly assigned letter: A to E for the five HDTV metrics and A to H for the eight Multimedia metrics. Note that these are different metrics. All metrics are linearly mapped onto a [1..5] scale.

The VQEG HDTV validation test includes 828 videos from six subjective tests. The MOSs are mapped onto a single scale using a common set of sequences that were included in all tests. The resulting merged dataset is referred to as the Superset. We will treat the six VQEG HDTV datasets as a single dataset and compute *ideal CI* and *practical CI* using the Superset MOSs.

The Multimedia validation test evaluated metrics for three different video resolutions, but we will limit our analyses to the VGA resolution metrics. More analyses of this sort would have minimal added value. The VGA metrics were analyzed against 13 datasets: twelve with 166 videos and one with 142 videos. Each dataset will be treated separately and weighted equally in our calculation of *ideal CI* and *practical CI*.

Warning: Analyses in this report will not match the VQEG independent lab group analyses of these metrics, due to major differences in processing. For example, we will compare all metrics to the MOSs, where VQEG compared full reference and reduced reference metrics to Differential Mean Opinion Scores (DMOS); and VQEG performed a non-linear mapping of metrics to MOS or DMOS (see [30]).

The VQEG HDTV metrics' *ideal CI* and *practical CI* are presented in Tables 11 and 12, respectively. The VQEG Multimedia metrics' *ideal CI* and *practical CI* are presented in Tables 13 and 14, respectively. The column ρ contains the Pearson correlation between all MOSs and each metric. For Tables 13 and 14, ρ pools data from all 13 datasets (i.e., without rescaling to remove lab-to-lab differences), while the column $\overline{\rho}$ computes Pearson correlation separately for each of the 13 Multimedia datasets and then takes the average. All metrics are sorted by the *correct ranking* incidence, which produces a different order for each table.

The column "Equivalent" indicates whether the metric is equivalent to another subjective test lab, based on either *ideal CI* or *practical CI*. This determination is made using the figure-of-merit *concur*, as described in Section 7.3. A hard threshold is used to determine equivalence, which means that some metrics will barely exceed that threshold. Metric **F** in Table 14 provides an example. If we lowered the *concur* threshold from 0.91 to 0.90, based on the tail of values in the Figure 8 histogram, metric **F** would be marked as equivalent to a subjective test, when using the *practical CI*. Alternatively, we could find outliers and omit these stimuli from the dataset *ex post facto* (e.g., a stimulus that is problematic for the metric, or a subject whose ratings shift the MOSs). Therefore, these statistics should be calculated and reported for both the entire dataset and the screened dataset (e.g., before and after subject screening).

These tables show that the best full-reference (FR) metrics produce decisions equivalent to those expected from a hypothetical subjective test lab using 24 (*ideal CI*) or 15 (*practical CI*) subjects.

Table 11. Ideal CI for VQEG HDTV Validation Test Metrics

Metric	**ρ**	**Equivalent**	**Ideal CI**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
E	0.86	Yes	1.00	50%	1%	6%	24%	19%
C	0.82	No	1.40	40%	1%	4%	34%	21%
PSNR	0.77	No	0.88	34%	1%	5%	40%	20%
D	0.74	No	1.04	30%	1%	4%	44%	21%
A	0.75	No	1.12	31%	1%	5%	43%	20%
B	0.63	No	1.60	19%	1%	3%	55%	22%

Table 12. Practical CI for VQEG HDTV Validation Test Metrics

Metric	**ρ**	**Equivalent**	**Practical CI**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
E	0.86	Yes	0.48	60%	3%	13%	12%	12%
C	0.82	No	0.64	56%	4%	12%	16%	13%
PSNR	0.77	No	0.48	50%	3%	12%	22%	13%
D	0.74	No	0.56	47%	4%	11%	25%	14%
A	0.75	No	0.60	48%	4%	12%	24%	13%
B	0.63	No	0.84	38%	5%	10%	32%	15%

Table 13. Ideal CI for VQEG Multimedia Validation Test Metrics

Metric	**ρ**	$\bar{\rho}$	**Equivalent**	**Ideal CI**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
H	0.85	0.87	Yes	0.80	51%	1%	7%	21%	19%
F	0.79	0.80	Yes	0.56	48%	1%	8%	25%	18%
A	0.84	0.85	Yes	0.92	47%	1%	6%	26%	20%
B	0.81	0.83	Yes	0.68	46%	1%	7%	26%	19%
G	0.81	0.81	No	0.68	42%	1%	6%	30%	21%
PSNR	0.75	0.77	No	0.88	37%	1%	6%	35%	20%
D	0.74	0.75	No	1.56	30%	1%	4%	42%	23%
E	0.50	0.54	No	1.44	11%	1%	2%	62%	25%
C	0.38	0.39	No	2.88	6%	1%	2%	66%	25%

Table 14. Practical CI for VQEG Multimedia Validation Test Metrics

Metric	ρ	$\bar{\rho}$	Equivalent	Practical CI	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie
H	0.85	0.87	Yes	0.48	59%	2%	14%	12%	13%
A	0.84	0.85	Yes	0.52	57%	3%	13%	14%	14%
B	0.81	0.83	Yes	0.40	56%	2%	13%	15%	14%
F	0.79	0.80	No[9]	0.36	56%	2%	13%	16%	14%
G	0.81	0.81	No	0.36	55%	2%	12%	16%	14%
D	0.74	0.75	No	0.72	49%	4%	12%	20%	15%
PSNR	0.75	0.77	No	0.52	48%	3%	12%	23%	15%
E	0.50	0.54	No	0.64	32%	6%	10%	36%	17%
C	0.38	0.39	No	1.44	24%	7%	9%	43%	18%

8.2 Ad-Hoc Test Equivalence for VQEG Validated Metrics

Let us now evaluate metrics' precision when CIs are not used. We will equate their performance to a number of subjects in an ad-hoc subjective test. For the ad-hoc subjective test, decisions are made using simple comparisons of average ratings, not statistical tests or CIs. For the metric decisions are made using simple comparisons of metric output values; again no statistical tests are used.

Table 15 shows the results for the HDTV metrics, and Table 16 shows the results for the Multimedia metrics. The best metrics are equivalent to a three subject ad-hoc test. PSNR is equivalent to a one person subject ad-hoc test. These determinations are made using the figure-of-merit *false ranking*, in accordance with the thresholds given in Section 7.4.

These tables show that the best full-reference (FR) metrics are equivalent to a three-person ad-hoc test. None of these metrics can replace a 6 subject pilot test.

[9] *Concur* = 0.9088, which is immediately below the 0.91 threshold for (2). The rounded incidence rates in this table ease comprehension but lessen the accuracy of this equivalence calculation.

Table 15. Ad-Hoc Test Equivalence for VQEG HDTV Validation Test Metrics

Metric	ρ	Number of Subjects	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie
E	0.86	3	68%	7%	25%	0%	0%
C	0.82	2	66%	9%	25%	0%	0%
D	0.74	1	63%	12%	25%	0%	0%
PSNR	0.77	1	64%	11%	25%	0%	0%
A	0.75	1	63%	12%	25%	0%	0%
B	0.63	NA	58%	17%	25%	0%	0%

Table 16. Ad-Hoc Test Equivalence for VQEG Multimedia Validation Test Metrics

Metric	ρ	$\bar{\rho}$	Number of Subjects	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie
H	0.85	0.87	3	67%	6%	27%	0%	0%
A	0.84	0.85	3	66%	7%	27%	0%	0%
B	0.81	0.83	3	66%	7%	27%	0%	0%
G	0.81	0.81	3	66%	7%	27%	0%	0%
F	0.79	0.80	3	67%	7%	27%	0%	0%
PSNR	0.75	0.77	1	63%	11%	27%	0%	0%
D	0.74	0.75	1	62%	11%	27%	0%	0%
E	0.50	0.54	NA	55%	19%	27%	0%	0%
C	0.38	0.39	NA	51%	22%	27%	0%	0%

8.3 Plots for Several Metrics

Let us now more closely examine the following VQEG Multimedia metrics:

- **H**, the most accurate metric
- **A**, another accurate metric
- **PSNR**, the de facto minimum performance baseline
- **E**, an inaccurate metric

Figure 12 shows the relationship between ΔM and the probability of each classification outcome. The x-axis plots increasing values for ΔM, and the y-axis plots the probability of each classification outcome. The *ideal CI* and *practical CI* are marked as vertical lines. *Correct tie* and *false tie* increase as ΔM increases, while *correct ranking*, *false ranking*, and *false distinction* decrease as ΔM increases. The ratio $(practical\ CI/ideal\ CI)$ differs for each metric; the relationship between these CIs is not a simple.

Generally speaking, worse metrics have larger CIs and a higher probability of errors. The relative probability error outcomes may also change. Our expectations from subjective tests in Sections 5 and 6 are that *false ranking* should be much lower than *disagree*, and thus by extension *false ranking* should be much lower than *false distinction*. For the more accurate metrics (**A**, **H**, and **PSNR**), *false ranking* is much lower than *false distinction*. However, for the inaccurate metric, **E**, *false ranking* is only slightly less than *false distinction*.

Figure 13 shows the accuracy of these four metrics by means of (metric versus MOS) scatter plots, using data from all 13 VQEG Multimedia VGA datasets without removing lab-to-lab differences. This matches our recommendation above to leave all MOSs on their native scale.

Figure 14 displays (PSNR versus MOS) scatter plots for each dataset individually. This shows that the relationship between MOS and PSNR is similar for all 13 datasets, with only minor differences. For example, see the VGA13 dataset, for which PSNR under-predicts MOS more often than what is typical for the other twelve datasets.

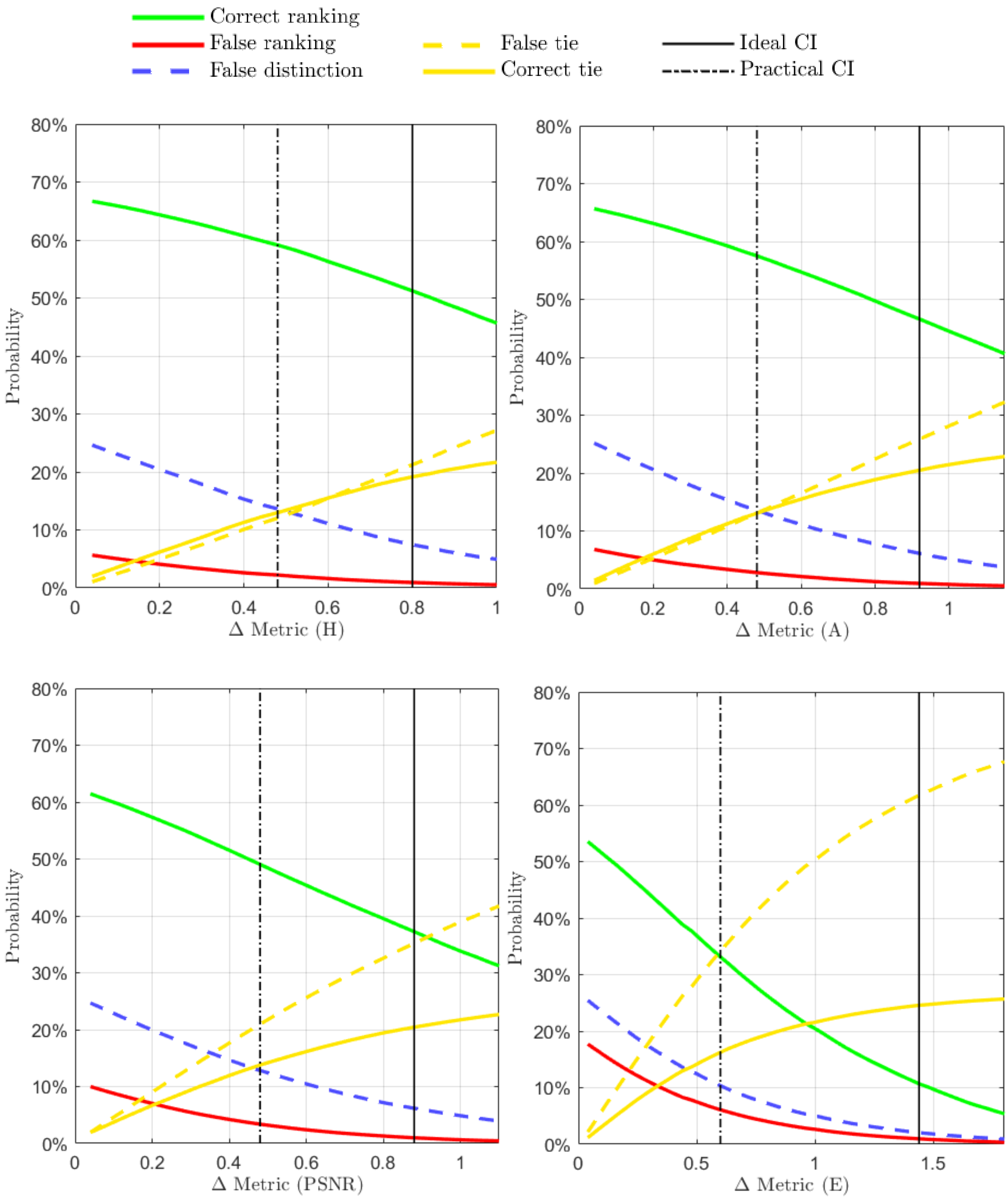

Figure 12. The probability of each classification type is plotted for VQEG Multimedia metrics **H** (upper left), **A** (upper right), **PSNR** (lower left), and **E** (lower right).

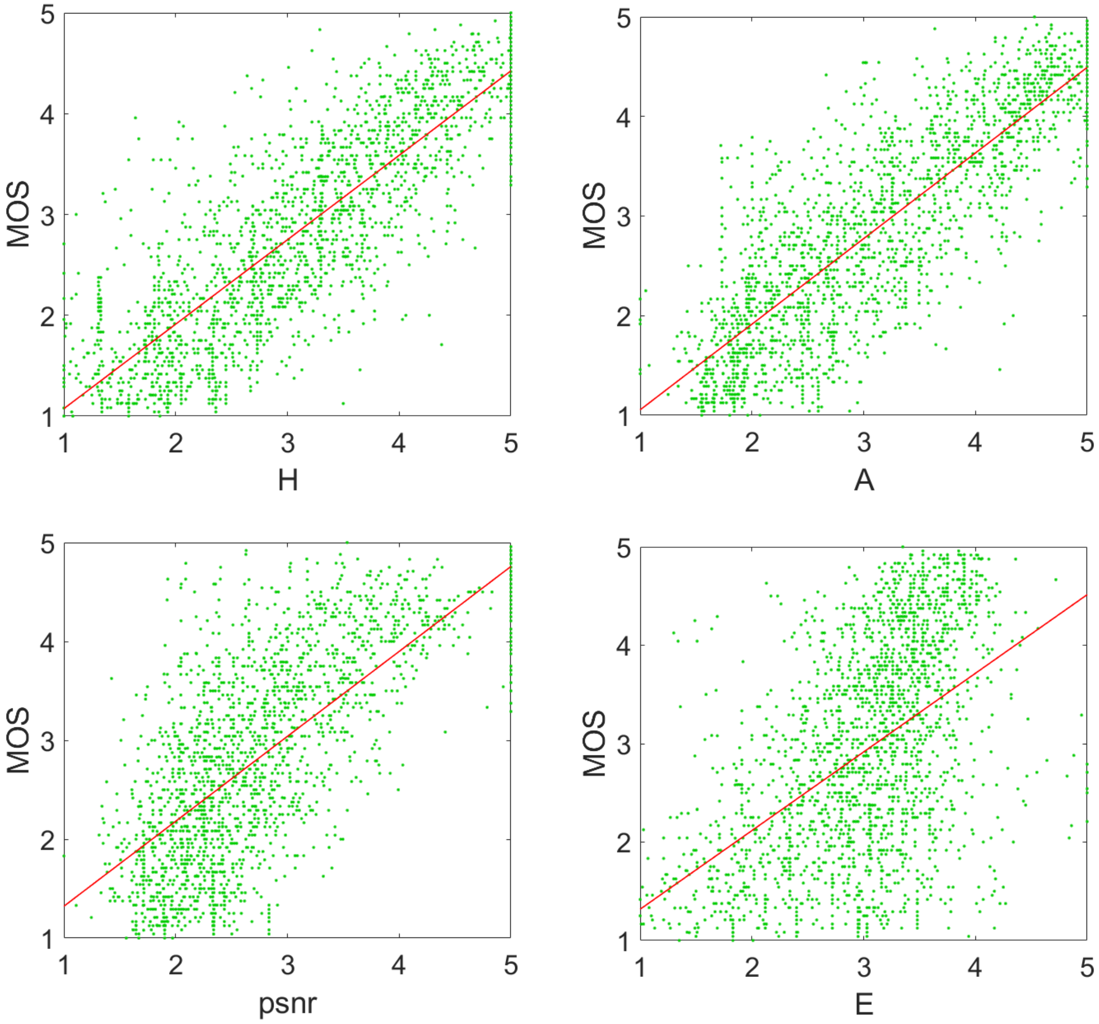

Figure 13. These scatter plots depict the accuracy of VQEG Multimedia metrics **H** (upper left), **A** (upper right), **PSNR** (lower left), and **E** (lower right). Data from all 13 datasets are pooled without removal of lab-to-lab differences. The red line depicts a linear fit.

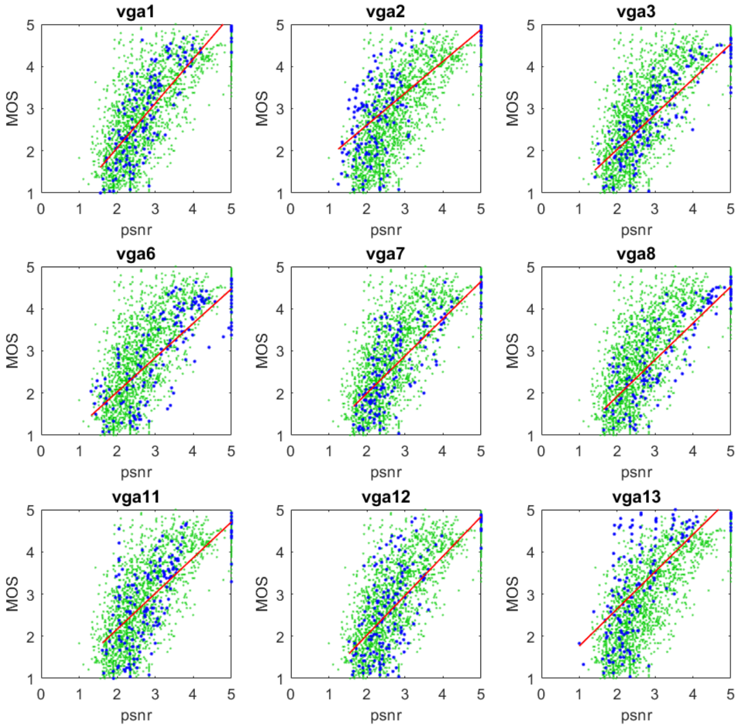

Figure 14. These scatter plots depict the accuracy of **PSNR** for each of the 13 VQEG Multimedia VGA datasets. The current dataset is plotted in blue. The light green points show the spread of (**PSNR** vs MOS) for all 13 datasets. The red line shows the linear fit between **PSNR** and MOS for the current dataset.

8.4 Confidence Intervals and Subjective Test Equivalence of VQEG AVHD-AS / P.NATS Phase 2 Metrics

So far, our metrics are anonymized and date from before 2010. This section provides measurements for a few contemporary metrics that can be evaluated by the reader. VQEG and the ITU conducted a joint validation test, referred to as AVHD-AS / P.NATS Phase 2. These are private datasets, whose use and distribution is limited by a multiple party non-disclosure agreement (NDA). The test plan is not publicly available, and the accuracy of the metrics is not publicly available. These datasets were designed according to rigorous test plans, similar to the other subjective tests conducted by VQEG and ITU as part of the standards development process.

Participants in AVHD-AS / P.NATS Phase 2 were invited to evaluate their metrics on these datasets for this report. The inclusion of these metrics should not be construed as any endorsement, approval, recommendation, prediction of success, or that they are in any way superior to or more noteworthy than the AVHD-AS / P.NATS Phase 2 metrics that are not mentioned.

OPTICOM provided analyses for three metrics:

- **BSM0** is a parametric model that takes as a inputs the video codec, the encoded video bitrate (averaged over ~8 sec), the encoded video resolution, and the encoded video framerate;

- **BSM1** is a parametric model that takes as an input the **BSM0** inputs plus the encoded video frame types (I or non-I frames) and the frame sizes in bytes; and

- **P.1204.5** is described in ITU-T Rec. P.1204.5 and takes as an input the **BSM0** inputs plus the pixels of the degraded video

These metrics were evaluated using 1,051 video stimuli from six databases: four that the working group considered "open databases" plus OPTICOM's two proprietary databases. These datasets evaluated 4K videos (3840 × 2160) at 60 fps with approximately 8 second duration. More information about these metrics can be found at OPTICOM GmbH, at https://www.opticom.de/.

Yonsei University provided analyses for one metric:

- **BSM0** is a parametric model that takes as inputs the video codec, the encoded video bitrate (averaged over ~8 sec), the encoded video resolution, and the encoded video framerate

This metric was evaluated using Yonsei's two proprietary databases. These datasets evaluated 4K videos (3840 × 2160) at 60 fps with approximately 8 second duration. This metric has not yet been published.

Table 17 lists the simplified conclusions that would be presented to a user:

- When decisions are reached *without* CIs, the equivalent number of people in an ad-hoc test
- *Practical CI*, and whether the metric is equivalent to a 15 subject test, when *practical CI* is used to make decisions
- *Ideal CI*, and whether the metric is equivalent to a 24 subject test, when *ideal CI* is used to make decisions

Table 17 showcases the admirable level of performance that can be achieved by the best video quality metrics.

When examining Table 17, three constraints must be kept in mind. First, these metrics are analyzed on different subjective tests. Second, we cannot compare the accuracy of these metrics using the information in Table 17. Other statistics must be used to evaluate the relative accuracy of one metric over another (e.g., Pearson correlation or a significance test based on Root Mean Squared Error). Third, the "number of subjects equivalence in an ad-hoc test" provides limited granularity (i.e., 1, 2, 3, 6, 9, or 12 subjects). We cannot conclude that a metric is equivalent to 4 or 5 subjects, for example. The coarse granularity of these measurements reflects the coarse granularity of our lab-to-lab comparisons.

Table 17. AVHD-AS / P.NATS Phase 2 Metric CI and Ad-Hoc Test Equivalence

Metric	**Range**	**Number of Subject Equivalence in Ad-Hoc Test**	**Practical CI**	**Equivalent to 15 Subject Test**	**Ideal CI**	**Equivalent to 24 Subject Test**
OPTICOM BSM0	1.0 to 5.0	3	0.40	Yes	0.71	Yes
OPTICOM BSM1	1.0 to 5.0	9	0.36	Yes	0.52	Yes
P.1204.5	1.0 to 5.0	9	0.32	Yes	0.48	Yes
Yonsei BSM0	1.55 to 4.63	6	0.27	Yes	0.45	Yes

9. IMPACT OF MULTIPLE DATASETS ON METRIC CONFIDENCE INTERVALS

9.1.1 Datasets

The VQEG datasets used in the prior section were carefully designed according to detailed test plans. Consequently, the Multimedia datasets respond similarly to each metric, as do the HDTV datasets. While there are certainly differences, these datasets cannot demonstrate the impact that major differences in experiment design and implementation have on *ideal CI* and *practical CI*.

We need a variety of subjective datasets that explore similar topics but have very different experiment designs. The *NRMetricFramework* GitHub repository [36] identifies a suitable set of datasets. These datasets are particularly suited for NR metric development for the following reasons:

- All media can be downloaded royalty free for research and development purposes
- Media adhere to consumer expectations around the quality of modern cameras
- Simulated impairments are avoided
- The media are either images or short video sequences (e.g., 4 seconds)
- Temporal changes in quality are avoided
- The dataset implements an unrepeated scene experiment design [8]
- Media were rated on the ACR scale

Broadly speaking, each dataset depicts a commercial camera or broadcast video application that would benefit from an NR metric. Temporal quality variations are avoided, because this can be studied separately and applied as post-processing. Conventional experiment designs re-use the same image or video multiple times (e.g., compressed at different bit-rates), while unrepeated experiment designs avoid re-using scenes. Unrepeated scene designs are preferred, because NR metrics will encounter heterogeneous scenes (subject matter) when deployed.

Table 18 lists the datasets identified in the NR Metric Framework repository. The information presented here differs from Table 1, because our dataset selection criteria differ. Table 18 contains the following columns:

• Dataset	Name of the dataset
• Ref.	Reference containing details of the experiment
• Media	Whether the dataset contains images or videos
• Impairments	Overall type of impairments: “Camera” for impairments created by the camera itself

"Compression" if the media was compressed with software after camera capture
"Pans" for dataset ITS4S4; see the discussion below

- Stimuli — Number of media viewed and rated by subjects
- Key Characteristics — Brief insights into the nature of the experiment.
- Suitability — "Optimal" if the dataset adheres closely to the above criteria
"Suboptimal" if it adheres to most but not all of the above criteria

Table 18. Freely Available Datasets Suitable for NR Metric Research

Dataset	Ref.	Media	Impairments	Stimuli	Key Characteristics	Suitability
AGH-NTIA-Dolby	[8]	Video	Compression	230	Exploratory experiment design, diverse scenes	Suboptimal
BID	[9]	Image	Camera	582	Diverse scenes	Optimal
CCRIQ	[10]	Image	Camera	221	Rigorous experiment design, balanced scenes	Optimal
CCRIQ2 & VIME1	[11]	Image	Camera	189	Limited scenes, exploratory experiment design	Suboptimal
CID2013	[12]	Image	Camera	474	Rigorous design, limited scenes, unusual subjective method	Optimal
DIQA	[13]	Image	Camera	175	Rigorous experiment design, black text on white paper	Suboptimal
ITS4S	[17]	Video	Compression, Camera	813	Exploratory experiment design, diverse scenes	Optimal
ITS4S2	[18]	Image	Camera	1,429	Diverse scenes, inexact experiment design	Optimal
ITS4S3	[19]	Video	Camera	594	Exploratory experiment design, diverse scenes	Optimal
ITS4S4	[20]	Video	Pans	196	Exploratory experiment design, diverse subject matter	Optimal
KoNVid1k	[21], [22]	Video	Camera	1,200	Rigorous experiment design, diverse scenes	Suboptimal
LIVE-Wild	[23]	Image	Camera	1,153	Inexact experiment design, diverse subject matter	Optimal
vqegHDCuts	[41]	Video	Compression	2,145	Conventional experiment design, diverse scenes	Suboptimal

Some datasets are not fully compliant with these specifications and are thus marked "suboptimal." KonVId1k and AGH/NTIA/Dolby contain longer scenes with temporal quality variations (e.g., due to scene cuts). KoNVid1k contains videos filmed between 2004 and 2014, many of which contain a lower quality than we would expect from modern camera systems. The

VIME1 portion of the CCRIQ2 & VIME1 experiment uses a loose experiment design that yielded unreliable data. AGH-NTIA-Dolby and vqegHDcuts contain MPEG2, AVC, and HEVC compression artifacts. Unlike all other datasets in Table 18, camera impairments are excluded.

The VQEG HDTV Superset videos contain changes in scene content and coding complexity that are outside the scope of the dataset selection criteria. To address this problem, we created a faux dataset, vqegHDcuts [41]. Transmission error impairments were eliminated. Each SRC was cut whenever the content or camera motion changed (e.g., at a scene cut, before and after a fade, before and after a camera pan). The MOS of the entire sequence was assigned to each segment, which adds error to the MOSs. This is an unprecedented technique, so the magnitude of this error is not known. Like the AGH-NTIA-Dolby dataset, the vqegHDcuts dataset avoids camera impairments.

The DIQA dataset [13] contains objective ratings from optical character recognition (OCR) algorithm success/failure rates instead of MOSs. Three OCR algorithms (Omni, Tesseract, and Fine Reader) were used with the same set of document photographs, to create three sets of faux MOSs. Instead of "suboptimal," the DIQA dataset could be more fairly labeled as an alternate strategy for creating NR metric training data.

The following datasets have interesting characteristics. The BID dataset [9] explores blur impairments of different types. The ITS4S4 dataset [20] explores quality impairments associated with camera pans at different speeds. This dataset contains a mixture of actual camera pans and simulated camera pans. Other impairments are avoided where possible. The LIVE public-domain subjective in the wild image quality challenge database (LIVE-Wild) [23] contains images that have been cropped to (500 × 500) pixels.

9.1.2 NR Parameter

We will examine a sample metric provided in version 1.0 of the NR Metric Framework [36]. We will refer to it as an NR parameter, because it focuses on one aspect of quality, instead of estimating MOS. Note that this repository uses 90% of stimuli for training and reserves 10% of stimuli for metric verification. Our analyses use the 90% of stimuli intended for training.

NR parameter *Viqet-Sharpness* calculates the sharpness/blurriness of an image or video using a Laplacian filter. *Viqet-Sharpness* is calculated as the average of the 1% of Laplacian filtered pixels with the highest magnitude, divided by the square root of the standard deviation of the Sobel filtered image. If that standard deviation is less than one, than this correction term is dropped. The Sobel filter adjusts *Viqet-Sharpness* for the overall magnitude of edge energy in the scene (e.g., whether the scene contains high contrast edges or mostly similar shades). There is also a correction term for 4K images, which are re-scaled to High Definition Television (HDTV) resolution. This is a simplification of the relationship observed in the CCRIQ dataset [10].

Viqet-Sharpness is an improved version of a blurring detection parameter developed on the CCRIQ dataset and distributed in the VIQET software [42]. Code for *Viqet-Sharpness* is available in version 1.0 of the NR Metric Framework function nrff_blur.m. *Viqet-Sharpness* was trained on CCRIQ, ITS4S, ITS4S2, ITS4S3, and LIVE-Wild. These datasets contain camera impairments. AGH-NTIA-Dolby and vqegHDcuts were available but not used for training. These datasets do not contain camera impairments and respond differently to the Laplacian filter.

Viqet-Sharpness adheres to the philosophy of the NR Metric Framework that temporal integration should be studied as a separate topic of research and applied as post-processing (see Section 2 of [17]). When applied to videos, *Viqet-Sharpness* examines each frame separately and takes the average (arithmetic mean) over time. Essentially, *Viqet-Sharpness* assumes that video quality changes over time are minimal and can be safely ignored.

9.1.3 Practical CI Dataset Comparisons

Figure 15 depicts the accuracy of *Viqet-Sharpness* for the Table 18 datasets. Using these scatter plots, we will visually categorize these 15 datasets into three subsets for further analysis: unfavorable, favorable, and moderate. Table 19 presents the *practical CI* for *Viqet-Sharpness* for each dataset. Table 20 presents the *Viqet-Sharpness* metric's equivalence, as a number of subjects in an ad-hoc test. In both tables, datasets are sorted by Pearson correlation (ρ).

The unfavorable subset contains AGH-NTIA-Dolby, vqegHDcuts, and KoNVid1k. For these three datasets, *Viquet-Sharpness* shows little relation to subjective scores. Unlike the other datasets in Table 18, AGH-NTIA-Dolby and vqegHDcuts contain compression impairments and no camera artifacts. The KoNVid1k dataset has very few stimuli with MOS above 4.0 (good) and many outliers where *Viqet-Sharpness* over-predicts quality. The KoNVid1k videos were filmed between 2004 and 2014, occasionally using cameras that pre-date 2004. Many of the videos do not adhere to consumer expectations, and the MOSs from subjects in 2016 reflect this. KoNVid1k and AGH-NTIA-Dolby contain longer sequences with temporal changes that *Viqet-Sharpness* ignores. The vqegHDcuts dataset videos were cut to remove temporal changes, but the faux MOSs retain the impact of temporal integration. Overall, there are numerous reasons why *Viqet-Sharpness* respond poorly to the unfavorable subset.

The favorable subset contains DIQA Finereader, DIQA Omni, DIQA Tesseract, CID2013, and CCRIQ2 & VIME1. These five datasets show high accuracy and a narrow scattering of data around the fit line. All three datasets contain a narrow scope of similar scene content, where the same composition is photographed by multiple cameras. VIME1, for example, focuses on buildings and a statue in Glasgow, Scotland, photographed at dusk. We hypothesized in Section 2 of [17] that such a lack of scene variety is problematic for NR metric research, because deployed systems will encounter heterogeneous subject matter. The response of *Viqet-Sharpness* to these datasets supports this theory. The DIQA photographs contain a variety of impairments, including blur, noise, and imperfect white balance. However, the three OCR algorithms seem to only be hindered by blur, so *Viqet-Sharpness* is a fortuitous metric for these datasets.

The moderate subset contains BID, CCRIQ, ITS4S, ITS4S2, ITS4S3, ITS4S4, and LIVE-Wild. The scatter plots and fits (red line) indicate that these datasets perform similarly (e.g., the blue dots for the current dataset have a similar distribution as the green dots for all datasets combined). The moderate datasets implement unrepeated scene experiment designs that robustly sample the large diversity of possible scenes. The image datasets (BID, CCRIQ, ITS4S2, and LIVE-Wild) have similarly shaped scatter plots as video datasets (ITS4S, ITS4S3, and ITS4S4). In contrast to the unfavorable datasets, these video datasets contain short videos (4 seconds duration) whose quality remains stable over time. For example, these datasets omit videos that are fairly static at the beginning and then pan quickly.

Overall, *practical CI* seems less sensitive to differences among multiple datasets than the above discussion predicts. When we calculate *practical CI* on all datasets or only the moderate datasets, the results are very similar (1.26 vs 1.28). *Viqet-Sharpness* appears to behave poorly on the AGH-NTIA-Dolby dataset based on the scatter plot and ρ, however *practical CI* = 1.12, which is lower than most datasets. We expected *practical CI* to be higher for the ITS4S4 dataset, because it avoids blur where possible. Instead *practical CI* = 1.2, which is similar to the value calculated using all datasets. *Viqet-Sharpness* appears to be nearly worthless for the vqegHDcuts dataset based on the scatter plot and ρ, however the *practical CI* is 1.36, which is on the higher side but not worse than ITS4S and ITS4S3, which yield *practical CI* of 1.38 and 1.30 respectively.

However, *practical CI* is sensitive to the favorable datasets, yielding relatively small values: *practical CI* = 0.70 for the favorable subset. None of the favorable datasets were used to train *Viqet-Sharpness* so, counter to expectations, the training data actually yields a more realistic *practical CI* than the test data. This probably has more to do with the importance of experiment design on NR metric training data than any characteristic of *practical CI*. Still, when estimating *practical CI*, it is important to obtain and use a sufficiently large number of datasets.

Table 20 indicates that, overall, *Viqet-Sharpness* does not meet the standards of an ad-hoc test with a single person. This is also true for the moderate subset, the unfavorable subset, and 10 of 15 datasets. For the favorable subset, *Viqet-Sharpness* is equivalent to a two person ad-hoc test, based on observed *false ranking* rates and the thresholds given in Section 7.4. The best result is for the DIQA Fine Reader dataset, where *Viqet-Sharpness* is equivalent to a 9 person subjective test.

This overly optimistic analysis highlights the need for sufficient data. For *Ideal CI* and *Practical CI* to be reliable, they must be computed using a robust sampling of the scenes, impairments, and quality levels that are of interest to the metric's users. This typically requires multiple subjective tests that were not used to train the metric. The same is true when estimating a metric's accuracy.

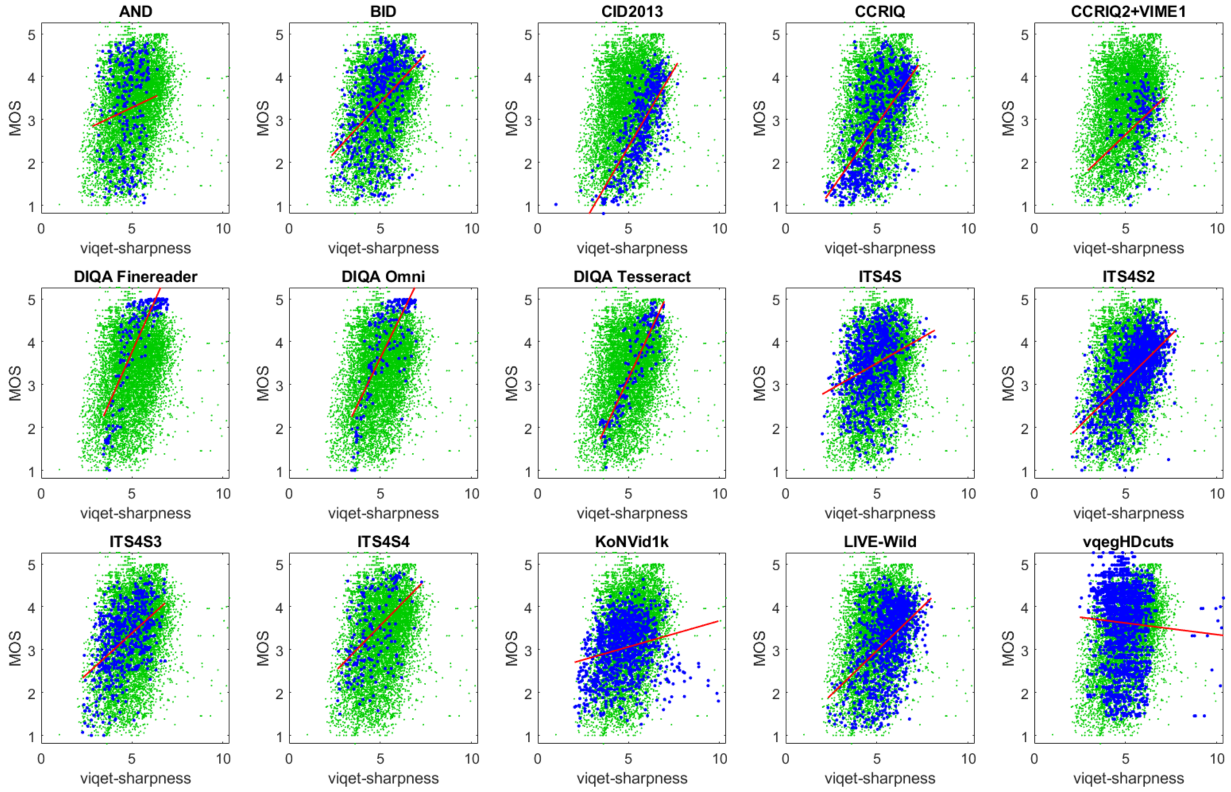

Figure 15. These scatter plots depict the accuracy of *Viqet-Sharpness* NR metric.

Table 19. Viqet-Sharpness Practical CI for Multiple Datasets

Dataset	**Subset**	**ρ**	**Practical CI**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
DIQA Tesseract	Favorable	0.90	0.44	61%	1%	14%	9%	15%
DIQA Fine Reader	Favorable	0.84	0.80	44%	0%	15%	14%	27%
DIQA Omni	Favorable	0.78	0.80	43%	3%	13%	21%	20%
CID2013	Favorable	0.74	0.70	42%	3%	11%	25%	18%
CCRIQ	Moderate	0.67	0.95	41%	4%	12%	27%	16%
ITS4S2	Moderate	0.61	1.14	29%	3%	13%	29%	26%
ITS4S3	Moderate	0.5	1.30	22%	4%	12%	37%	25%
ITS4S4	Moderate	0.5	1.20	27%	5%	11%	37%	20%
LIVE-Wild	Moderate	0.49	1.14	24%	4%	11%	37%	24%
BID	Moderate	0.47	1.15	27%	6%	10%	38%	19%
CCRIQ2 & VIME1	Favorable	0.46	1.00	20%	4%	12%	35%	29%
ITS4S	Moderate	0.3	1.38	16%	5%	10%	43%	26%
KoNVid1k	Unfavorable	0.22	1.76	13%	6%	10%	39%	32%
AGH-NTIA-Dolby	Unfavorable	0.13	1.12	13%	8%	8%	52%	19%
vqegHDCuts	Unfavorable	0.05	1.36	8%	7%	8%	50%	26%

Dataset	Subset	ρ	Practical CI	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie
Dataset		$\overline{\rho}$	**Practical CI**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
All Datasets	—	0.51	1.26	21%	5%	10%	39%	25%
Favorable Subset	Favorable	0.75	0.70	42%	3%	13%	21%	20%
Moderate Subset	Moderate	0.41	1.28	19%	5%	10%	41%	25%
Unfavorable Subset	Unfavorable	0.14	1.52	9%	6%	8%	48%	28%

Table 20. Ad-Hoc Test Equivalence for Viqet-Sharpness

Dataset	Subset	ρ	Number of Subj.	Correct Ranking	False Ranking	False Distinction	False Tie	Correct Tie
DIQA Tesseract	Favorable	0.90	6	67%	4%	29%	0%	0%
DIQA Fine Reader	Favorable	0.84	9	54%	4%	44%	0%	0%
DIQA Omni	Favorable	0.78	2	58%	9%	33%	0%	0%
CID2013	Favorable	0.74	1	59%	11%	30%	0%	0%
CCRIQ	Moderate	0.67	NA	58%	14%	28%	0%	0%
ITS4S2	Moderate	0.61	1	48%	12%	40%	0%	0%
ITS4S3	Moderate	0.5	NA	46%	17%	37%	0%	0%
ITS4S4	Moderate	0.50	NA	50%	19%	31%	0%	0%
LIVE-Wild	Moderate	0.49	NA	46%	19%	35%	0%	0%
BID	Moderate	0.47	NA	51%	21%	29%	0%	0%
CCRIQ2 & VIME1	Favorable	0.46	NA	41%	18%	40%	0%	0%
ITS4S	Moderate	0.30	NA	41%	22%	36%	0%	0%
KoNVid1k	Unfavorable	0.22	NA	36%	21%	42%	0%	0%
AGH-NTIA-Dolby	Unfavorable	0.13	NA	40%	33%	27%	0%	0%
vqegHDCuts	Unfavorable	0.05	NA	34%	31%	35%	0%	0%
Dataset		$\overline{\rho}$	**Number of Subj.**	**Correct Ranking**	**False Ranking**	**False Distinction**	**False Tie**	**Correct Tie**
All Datasets	—	0.51	NA	44%	21%	35%	0%	0%
Favorable Subset	Favorable	0.75	2	57 %	10%	34%	0%	0%
Moderate Subset	Moderate	0.41	NA	48%	17%	35%	0%	0%
Unfavorable Subset	Unfavorable	0.14	NA	33%	30%	36%	0%	0%

10. SUMMARY

This report analyzes the conclusions reached by subjective tests and objective metrics. We assess the precision of 60 subjective tests, expressed as CIs. We defined the MOS CI (ΔS_{CI}) as the difference in MOS values at which 95% of the pairs will be statistically different (according to the Student's *t*-test using a 95% confidence level). When the ACR method is used in a well-designed subjective test, 24 subjects will produce $\Delta S_{CI} = 0.5$, and 15 subjects will produce $\Delta S_{CI} = 0.7$. Unknown factors may produce larger values of ΔS_{CI}.

We used 90 lab-to-lab comparisons to assess the repeatability of subjective tests. We expressed repeatability as the likelihood that two well-designed tests will reach the same conclusions or different conclusions. Our analysis of subjective tests indicates that, when two labs conduct the same subjective test, there is a ≤ 1% chance that the labs will agree that the stimuli have significantly different quality but disagree on which stimulus has higher quality. We also defined a figure-of-merit, *concur*, that compares the results of two tests (subjective-to-subjective, or subjective-to-objective) and allows us to analyze the extent to which two subjective tests reach the same conclusion.

Based on these analyses, we observe these same relationships from well-designed subjective tests that assess speech quality, image quality, and video quality.

We also consider the relationships between objective metrics and subjective tests. We use a confusion matrix to classify the conclusions reached by subjective test and a metric. Of most concern are the errors that result when comparing the quality of two stimuli, **A** and **B**. The first type of error is *false distinction*, which means the metric rank orders the quality of **A** and **B**, but a well-designed subjective test cannot (i.e., **A** and **B** have statistically equivalent quality). The second type of error is *false ranking*, which means that the metric concludes **A** is better than **B**, but a well-designed subjective test concludes that **B** is better than **A**.

We proposes a method for explaining the precision of an objective metric to naive users. The method calculates the following:

- *Ideal CI*, calculated with strict criteria
- Whether the metric is equivalent to a 24 person subjective test, when using *ideal CI*
- *Practical CI*, calculated with less stringent criteria
- Whether the metric is equivalent to a 15 person subjective test, when using *practical CI*
- *N*, the number of subjects in an ad-hoc assessment or pilot test that is equivalent to the metric

When using *ideal CI* or *practical CI*, metric values indicate a preference only when the difference is greater than the CI. For example,

$$\mathbf{A} > \mathbf{B} + practical\ CI$$

or

$$\mathbf{B} > \mathbf{A} + practical\ CI$$

Both *ideal CI* and *practical CI* yield *false distinction* and *false ranking* rates equivalent to a well-designed test. We propose that the *practical CI* be used in most circumstances.

The final option equates the metric to an ad-hoc study of *N* subjects. This comparison assumes that any difference in metric values indicates a preference.

$$\mathbf{A} > \mathbf{B}$$

The metric's performance is then equated to an ad-hoc study of 12, 9, 6, 3, 2, or 1 subjects. The metric's performance can also be worse than one person's assessment.

11. REFERENCES

[1] National Telecommunications and Information Administration, Institute for Telecommunication Sciences, "NR Metric Framework," https://github.com/NTIA/NRMetricFramework, accessed 7/27/2020.

[2] ATIS T1.TR.72-2003 "Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics," Alliance for Telecommunications Industry Solutions, https://www.atis.org/docstore/product.aspx?id=10518.

[3] Recommendation ITU-T J.149 (03/04), *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)*, International Telecommunication Union, Geneva, Switzerland, available at http://www.itu.int/rec/T-REC-J.149/en.

[4] M.H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, Vol. 19, Feb. 2004, pp. 101-107.

[5] M. Pinson and S. Wolf, "Techniques for Evaluating Objective Video Quality Models Using Overlapping Subjective Data Sets," NTIA Technical Report TR-09-457, Nov. 2008. https://www.its.bldrdoc.gov/publications/2494.aspx.

[6] S. Voran, "Techniques for Comparing Objective and Subjective Speech Quality Tests," *Proceedings of the Speech Quality Assessment Workshop at Ruhr-Universtät Bochum, Germany*, Nov. 1994, pp. 59-64. https://www.its.bldrdoc.gov/publications/2652.aspx

[7] Babak Naderi, Tobias Hossfeld, Matthias Hirth, Florian Metzger, Sebastian Möller, and Rafael Zequeira Jiménez, "Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach," *Cornell University*, 2020, https://arxiv.org/abs/2003.11300v1.

[8] Lucjan Janowski, Ludo Malfait, and Margaret Pinson, "Evaluating experiment design with unrepeated scenes for video quality subjective assessment." *Quality and User Experience*, vol. 4, no. 2, Jun. 2019. https://doi.org/10.1007/s41233-019-0026-4.

[9] A. Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," in *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64-75, Jan. 2011.

[10] Michele A. Saad; Margaret H. Pinson; David G Nicholas; Niels Van Kets; Glenn Van Wallendael; Ramesh V Jaladi; Philip J. Corriveau, "Image quality of experience: a subjective test targeting the consumer's experience," *Proceedings of the International Symposium on Electronic Imaging 2016, Human Vision and Electronic Imaging 2016*, February 14, 2016, https://www.its.bldrdoc.gov/publications/3172.aspx.

[11] Jakub Nawala; Margaret H. Pinson; Mikolaj Leszczuk; Lucjan Janowski, “Study of Subjective Data Integrity for Image Quality Data Sets with Consumer Camera Content,” *J. Imaging*, 6, no. 3: 7, https://www.its.bldrdoc.gov/publications/3239.aspx.

[12] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen and J. Häkkinen, “CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms,” *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 390-402, Jan. 2015.

[13] Jayant Kumar, Peng Ye, and David Doermann, “A Dataset for Quality Assessment of Camera Captured Document Images.” *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, pp. 39-44, August 2013.

[14] Margaret H. Pinson, William J. Ingram, and Arthur A. Webster, “Audiovisual quality components: an analysis,” *IEEE Signal Processing Magazine*, vol.28, no.6, pp.60-67, Nov. 2011, https://www.its.bldrdoc.gov/publications/2565.aspx.

[15] Margaret H. Pinson; Arthur A. Webster; William J. Ingram, “Preliminary Investigation into the Impact of Audiovisual Synchronization of Impaired Audiovisual Sequences,” NTIA Technical Memo TM-11-474, Mar. 2011, https://www.its.bldrdoc.gov/publications/2549.aspx.

[16] Recommendation ITU-T P.Sup23, *ITU-T coded-speech database*, International Telecommunication Union, Geneva, Switzerland, Feb. 27, 1998. https://www.itu.int/rec/T-REC-P.Sup23-199802-I

[17] Margaret H. Pinson, “ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes,” NTIA Technical Memo TM-18-532, Feb. 2018, https://www.its.bldrdoc.gov/publications/3194.aspx.

[18] Margaret H. Pinson, “ITS4S2: An Image Quality Dataset With Unrepeated Images From Consumer Cameras,” NTIA Technical Memo TM-19-537, Apr. 2019, https://www.its.bldrdoc.gov/publications/3219.aspx.

[19] Margaret H. Pinson, “ITS4S3: A Video Quality Dataset With Unrepeated Videos, Camera Impairments, and Public Safety Scenarios,” NTIA Technical Memo TM-19-538, Apr. 2019, https://www.its.bldrdoc.gov/publications/3220.aspx.

[20] Margaret H. Pinson; Samuel Elting, “ITS4S4: A Video Quality Study of Camera Pans,” NTIA Technical Memo TM-20-545, Dec. 2019. https://www.its.bldrdoc.gov/publications/3233.aspx

[21] Vlad Hosu et al., “The Konstanz Natural Video Database,” 2017. http://database.mmsp-kn.de

[22] V. Hosu et al., “The Konstanz natural video database (KoNViD-1k),” *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, 2017, pp. 1-6.

[23] D. Ghadiyaram and A. C. Bovik, “Massive Online Crowdsourced Study of Subjective and Objective Picture Quality,” in *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372-387, Jan. 2016.

[24] Margaret H. Pinson; Stephen Wolf; Robert B. Stafford, “Video Performance Requirements for Tactical Video Applications,” *2007 IEEE Conference on Technologies for Homeland Security*, pp.85-90, 16-17 May 2007, https://www.its.bldrdoc.gov/publications/2573.aspx.

[25] Samira Tavakoli, Kjell Brunnström, Jesús Gutiérrez, and Narciso García, “Quality of experience of adaptive video streaming: investigation in service parameters and subjective quality assessment methodology,” *Signal Processing: Image Communication*, vol 39, Part B, Nov. 2015, p. 432-443. https://doi.org/10.1016/j.image.2015.05.001

[26] “Full Reference Television Phase I Testplan,” Video Quality Experts Group (VQEG), Feb. 15, 1999. https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx

[27] Ann Marie 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, https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx.

[28] “FR-TV: Full Reference Television Phase II Subjective Test Plan,” Video Quality Experts Group (VQEG), Sep. 2002. https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx

[29] Philip Corriveau and Arthur 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, https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx.

[30] Arthur Webster and Filippo Speranza, “Report of the validation of video quality models for high definition video content,” *Video Quality Experts Group (VQEG)*, Jun. 2010, https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.

[31] Chulhee Lee, Silvio Borer, and Jens Berger, “Hybrid Perceptual / Bitstream validation test final report,” *Video Quality Experts Group (VQEG)*, Jul. 2014. https://www.its.bldrdoc.gov/vqeg/projects/hybrid-perceptual-bitstream/hybrid-perceptual-bitstream.aspx

[32] Arthur Webster et al., “Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I,” Video Quality Experts Group (VQEG), Sep. 12, 2008, https://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx

[33] Margaret H. Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Philip J. Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram, “The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study,” *IEEE Journal of Selected Topics in Signal*

Processing, Vol. 6, No. 6, Oct. 2012, pp. 640–651, https://www.its.bldrdoc.gov/publications/2682.aspx.

[34] Arthur Webster and Filippo Speranza, "Validation of reduced-reference and no-reference objective models for standard definition television, phase I," *Video Quality Experts Group (VQEG)*, 2009, https://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/rrnr-tv.aspx.

[35] Margaret H. Pinson; Lucjan Janowski; Zdzislaw Papir, "Video Quality Assessment: Subjective testing of entertainment scenes," *IEEE Signal Processing Magazine*, vol. 32, no. 1, pp. 101-114, Jan. 2015, https://www.its.bldrdoc.gov/publications/2821.aspx.

[36] National Telecommunications and Information Administration, Institute for Telecommunication Sciences, "NR Metric Framework," https://github.com/NTIA/NRMetricFramework/releases, accessed 03/26/2020.

[37] Lucjan Janowski; Margaret H. Pinson, "The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, December 2015, pp 2210-2224, https://www.its.bldrdoc.gov/publications/2814.aspx.

[38] Lucjan Janowski; Margaret H. Pinson, "Subject Bias: Introducing a Theoretical User Model," *Fifth International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore, 18-20 September 2014, https://www.its.bldrdoc.gov/publications/2774.aspx.

[39] T. Tominaga, T. Hayashi, J. Okamoto and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, Trondheim, 2010, pp. 82-87, https://ieeexplore.ieee.org/document/5517948

[40] Arthur Webster and Filippo 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, https://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx

[41] Margaret H. Pinson, "Analysis of No-Reference Metrics for Image and Video Quality of Consumer Applications," NTIA Technical Memo TM-20-547, January 2020, https://www.its.bldrdoc.gov/publications/3235.aspx.

[42] Video Quality Experts Group (VQEG), "VIQET – VQEG Image Quality Evaluation Tool," https://github.com/VIQET, accessed 4/10/2020.

ACKNOWLEDGEMENTS

The author would like to recognize Stephen Voran (NTIA/ITS) for proposing the use of confusion matrices and classification plots as they appear in ATIS T1.TR.72.

APPENDIX A. SUBJECTIVE TEST CONFIDENCE INTERVAL MEASUREMENTS

This appendix reports ΔS_{CI} for the datasets in Table 1. The "subjects" column contains (condition) if this is a speech dataset where the media are analyzed per condition. ΔS_{CI} is highlighted in bold to draw attention to differences within the dataset. For example, the method changes from ACR to a non-standard method, the subjects change from naive to experts, the media change from speech files to speech conditions, or data from multiple labs presented separately and pooled.

Dataset	**Media**	**Study**	**Subjects**	**Method**	**Scale**	**ΔS_{CI}**
401 **501** **701**	Speech	Crowd	227 (condition) 115 (condition) 144 (condition)	ACR	[1..5]	0.3 0.3 0.4
AGH/NTIA/Dolby	Video	Field	71 31 22 18	ACR	[1..5]	0.3 0.4 0.5 0.7
CCRIQ	Image	Field	26 9 8 9 27 9 9 9	ACR	[1..5]	0.6 1.1 1.2 1.2 0.6 1.1 1.1 1.2
CCRIQ2	Image	Field	19	ACR	[1..5]	0.8
Hybrid HDTV	Video	Lab	24 24 24 24 24	ACR	[1..5]	0.5 0.5 0.5 0.5 0.6
Hybrid VGA	Video	Lab	24	ACR	[1..5]	0.6 0.5 0.6
Hybrid WVGA	Video	Lab	24	ACR	[1..5]	0.6 0.5
ITS4S	Video	Field	27 24	ACR	[1..5]	0.5 0.6
ITS4S2	Image	Field	16	ACR	[1..5]	0.7
ITS4S3	Video	Field	14 17 14 15 13 19	ACR	[1..5]	0.9 0.9 1.0 0.9 1.0 0.7
ITS4S4	Video	Field	26	ACR	[1..5]	0.7
ITS 2010	**Audiovisual**	Lab	≈26	ACR	[1..5]	1.3
ITS AV-Sync 2010	**Audiovisual**	Lab	12 16	ACR	[1..5]	1.0 1.3

Dataset	Media	Study	Subjects	Method	Scale	ΔS_{CI}
ITU-T Rec. P.Sup23, EXP 1	Speech	Lab	24 24 24	ACR	[1..5]	0.5 0.5 0.6
ITU-T Rec. P.Sup23, EXP 2	Speech	Lab	48 48 48	CCR	[-3..3]	0.8 0.8 0.8
ITU-T Rec. P.Sup23, EXP 3	Speech	Lab	24 24 24 24	ACR	[1..5]	0.5 0.5 0.5 0.6
Private Speech Dataset #1	Speech	Lab	≈11 ≈22 43 **344 to 440 (condition)**	ACR	[1..5]	0.9 0.5 0.4 **0.3**
Private Speech Dataset #2	Speech	Lab	8 **512 (condition)**	ACR	[1..5]	1.4 **0.3**
Private Speech Dataset #3	Speech	Lab	10 8 8 8 8 8 8 8	ACR	[1..5]	1.3 1.1 1.2 1.2 1.2 1.2 1.3 1.0
Private Video Dataset #1	Video	Lab Crowd	**15 experts** 61	ACR	[0..100]	**11** 10
Private Video Dataset #2	Video	Lab	30	ACR	[1..5]	0.5
Public Safety #1	Video	Lab	16 16	ACR **Boolean**	[1..5] [0,1]	0.6 **0.4**
Public Safety #2	Video	Lab	19 19	ACR **Boolean**	[1..5] [0,1]	0.6 **0.4**

UPM-Acreo	Video	Lab	20
			22
			21
				ACR
				CIETI(V)
				CIETI(AV)
					[1..5]	0.6
						0.6
						0.6

VIME1	Image	Field	21	ACR	[1..5]	0.8
VQEG FRTV Phase I	Video	Lab	79 or 90	DSCQS, SRC DSCQS, PVS **DSCQS, DOS**	[0..100] [0..100] [-100..100]	5.0 6.0 **7.0**
VQEG FRTV Phase II, 625-line	Video	Lab	27	DSCQS, SRC DSCQS, PVS **DSCQS, DOS**	[0..100] [0..100] [-100..100]	6.0 8.0 **7.0**
VQEG FRTV Phase II, 525-line	Video	Lab	64	DSCQS, SRC DSCQS, PVS **DSCQS, DOS**	[0..100] [0..100] [-100..100]	4.0 5.0 **5.0**
VQEG HDTV	Video	Lab	24 24 24 24	ACR	[1..5]	0.5 0.5 0.5 0.5

Dataset	**Media**	**Study**	**Subjects**	**Method**	**Scale**	**ΔS_{CI}**
			24			0.5
			24			0.5
VQEG HDTV and VQEG Hybrid, Extrapolation	Video	Lab	15	ACR	[1..5]	0.7
			9			1.1
			6			1.5
VQEGMM2	Video	—	213 (all labs pooled)	ACR	[1..5]	**0.2**
		Lab	28			0.6
		Public	9			1.6
		Lab	34			0.5
		Lab	25			0.6
		Public	25			0.6
		Lab	24			0.7
		Public	24			0.6
		Lab	14			0.9
		Public	15			0.9
		Lab	15			0.9
VQEG RRNR-TV	Video	Lab	32	ACR	[1..5]	0.4
			31			0.5
VQEG RRNR-TV	Video	Lab	32	**ACR-HR**	[1..5]	**0.5**
			31			**0.6**

APPENDIX B. LAB-TO-LAB CONCLUSION CLASSIFICATION DATA

This Appendix reports incidence of conclusions reached when two or more labs conduct the same experiment, using the 5-level ACR method.

Dataset	**Subset**	**Stimuli**	**Subjects**		**Agree Ranking**	**Agree Tie**	**Unconfirmed**	**Disagree**
AGH/NTIA/Dolby		230	31	22	76.30%	12.32%	11.36%	0.02%
AGH/NTIA/Dolby		230	31	18	71.87%	12.57%	15.49%	0.08%
AGH/NTIA/Dolby		230	22	18	70.42%	15.04%	14.48%	0.05%
ITS4S	Partial	212	27	24	60.54%	17.85%	21.34%	0.26%
CCRIQ	Red	171	18	17	54.56%	22.41%	22.16%	0.86%
CCRIQ	Red	171	18	18	50.09%	25.96%	23.41%	0.54%
CCRIQ	Red	171	17	18	48.48%	27.25%	24.06%	0.21%
CCRIQ	Blue	212	18	17	47.44%	23.10%	28.69%	0.77%
CCRIQ	Blue	212	18	18	49.34%	24.52%	25.86%	0.28%
CCRIQ	Blue	212	17	18	47.07%	28.97%	23.79%	0.17%
Private Speech Dataset #3	A	288	10	8	50%	26%	24%	0.21%
Private Speech Dataset #3	B	288	8	8	51%	26%	22%	0.23%
Private Speech Dataset #3	C	288	8	8	52%	27%	20%	0.31%
Private Speech Dataset #3	D	288	8	8	52%	26%	22%	0.30%
VQEGMM2		60	28	9	53.22%	19.66%	27.06%	0.06%
VQEGMM2		60	28	34	73.67%	11.86%	14.18%	0.28%
VQEGMM2		60	28	25	71.30%	12.60%	15.99%	0.11%
VQEGMM2		60	28	25	69.21%	14.35%	16.38%	0.06%
VQEGMM2		60	28	24	70.23%	13.28%	16.38%	0.11%
VQEGMM2		60	28	24	71.86%	13.95%	14.12%	0.06%
VQEGMM2		60	28	14	66.38%	16.10%	17.34%	0.17%
VQEGMM2		60	28	15	63.84%	15.99%	20.11%	0.06%
VQEGMM2		60	28	15	65.25%	13.95%	20.56%	0.23%
VQEGMM2		60	9	34	53.50%	14.97%	31.41%	0.11%
VQEGMM2		60	9	25	52.88%	17.57%	29.49%	0.06%
VQEGMM2		60	9	25	52.20%	20.79%	26.95%	0.06%
VQEGMM2		60	9	24	51.64%	18.19%	30.00%	0.17%
VQEGMM2		60	9	24	53.16%	18.76%	27.97%	0.11%
VQEGMM2		60	9	14	51%	24%	25%	0.06%
VQEGMM2		60	9	15	60%	25%	25%	0.06%
VQEGMM2		60	9	15	51%	23%	27%	0.11%
VQEGMM2		60	34	25	76%	12%	11%	0.00%
VQEGMM2		60	34	25	72%	12%	15%	0.06%

Dataset	**Subset**	**Stimuli**	**Subjects**		**Agree Ranking**	**Agree Tie**	**Unconfirmed**	**Disagree**
VQEGMM2		60	34	24	76%	13%	11%	0.00%
VQEGMM2		60	34	24	77%	14%	10%	0.00%
VQEGMM2		60	34	14	69%	14%	17%	0.00%
VQEGMM2		60	34	15	67%	14%	20%	0.11%
VQEGMM2		60	34	15	69%	13%	18%	0.06%
VQEGMM2		60	25	25	72%	15%	14%	0.06%
VQEGMM2		60	25	24	71%	13%	16%	0.17%
VQEGMM2		60	25	24	73%	13%	14%	0.17%
VQEGMM2		60	25	14	67%	14%	19%	0.11%
VQEGMM2		60	25	15	65%	15%	20%	0.00%
VQEGMM2		60	25	15	67%	14%	18%	0.11%
VQEGMM2		60	25	24	70%	15%	16%	0.06%
VQEGMM2		60	25	24	71%	15%	14%	0.06%
VQEGMM2		60	25	14	65%	17%	18%	0.06%
VQEGMM2		60	25	15	65%	19%	16%	0.00%
VQEGMM2		60	25	15	66%	16%	18%	0.06%
VQEGMM2		60	24	24	74%	16%	10%	0.00%
VQEGMM2		60	24	14	67%	16%	17%	0.00%
VQEGMM2		60	24	15	65%	17%	18%	0.06%
VQEGMM2		60	24	15	66%	14%	21%	0.00%
VQEGMM2		60	24	14	67%	16%	17%	0.06%
VQEGMM2		60	24	15	65%	17%	18%	0.00%
VQEGMM2		60	24	15	67%	15%	17%	0.11%
VQEGMM2		60	14	15	61%	20%	19%	0.06%
VQEGMM2		60	14	15	63%	18%	19%	0.06%
VQEGMM2		60	15	15	63%	21%	16%	0.06%
VQEG FRTV Phase I (narrow range)	Low Quality, 525-line	90	18 18 18 18 18 18	18 16 18 16 18 16	60% 60% 57% 65% 59% 59%	18% 17% 22% 17% 20% 19%	22% 23% 21% 19% 21% 22%	0.20% 0.10% 0.00% 0.22% 0.02% 0.02%
VQEG FRTV Phase I (very narrow range)	High Quality, 525-line	90	18 18 18 18 18 18	16 16 16 18 18 18	46% 49% 46% 48% 45% 48%	25% 23% 26% 22% 25% 23%	29% 28% 27% 29% 30% 28%	0.17% 0.12% 0.02% 0.87% 0.77% 0.50%
VQEG FRTV Phase I (narrow range)	Low Quality, 625-line	79	18 17 17 18 18 18	17 16 18 16 18 16	37% 52% 56% 37% 37% 56%	28% 26% 23% 29% 24% 25%	33% 21% 21% 33% 38% 19%	**1.82%** 0.88% 0.00% 0.13% 0.91% 0.10%

Dataset	**Subset**	**Stimuli**	**Subjects**		**Agree Ranking**	**Agree Tie**	**Unconfirmed**	**Disagree**
VQEG FRTV Phase I (very narrow range)	High Quality, 625-line	90	17	16	24%	45%	31%	0.30%
			18	17	29%	48%	23%	0.00%
			17	16	30%	39%	30%	0.15%
			18	16	26%	46%	27%	0.17%
			16	16	29%	39%	32%	0.02%
			18	16	33%	41%	25%	0.07%
VQEG FRTV Phase II, 525-line	PVS	64	32	32	75%	10%	15%	0.15%
	DOS		32	32	75%	10%	15%	0.10%
VQEG RRNR-TV	525-line	168	16	16	61%	14%	24%	0.67%
	625-line		16	15	63%	17%	19%	0.14%
RRNR-TV ACR-HR	525-line	168	16	16	59%	16%	24%	0.95%
	625-line		16	15	61%	21%	18%	0.48%
VQEG Hybrid HD, 5 tests	Common set	24	24	24	68%	11%	20%	0.94%

NTIA FORM 29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.
TR-21-550

2. Government Accession No.

3. Recipient's Accession No.

4. TITLE AND SUBTITLE
Confidence Intervals for Subjective Tests and Objective Metrics That Assess Image, Video, Speech, or Audiovisual Quality

5. Publication Date
October 2020

6. Performing Organization Code
NTIA/ITS.P

7. AUTHOR(S)
Margaret H. Pinson

9. Project/Task/Work Unit No.
6895000-300

8. PERFORMING ORGANIZATION NAME AND ADDRESS
Institute for Telecommunication Sciences
National Telecommunications & Information Administration
U.S. Department of Commerce
325 Broadway
Boulder, CO 80305

10. Contract/Grant Number.

11. Sponsoring Organization Name and Address

National Institute of Standards and Technology
325 Broadway
Boulder, CO 80305

12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

This report describes a methodology that measures the precision of objective metrics. We assess the confidence intervals of subjective tests and the likelihood that two subjective test labs will reach the same or different conclusions when two stimuli are compared. This allows us to compute the metric's confidence interval and, when confidence intervals are used to make decisions, to prove whether the metric performs similarly to a subjective test with 15 or 24 subjects. When confidence intervals are not used, the metric's precision is likened to a certain number of people in an ad-hoc quality assessment. The methods in this report are developed and evaluated using speech quality, video quality, image quality, and audiovisual quality datasets.

16. Key Words (Alphabetical order, separated by semicolons)

audiovisual quality, confidence interval, image quality, metric, speech quality, subjective test, video quality

17. AVAILABILITY STATEMENT

☒ UNLIMITED.

☐ FOR OFFICIAL DISTRIBUTION.

18. Security Class. (This report)

Unclassified

20. Number of pages

83

19. Security Class. (This page)

Unclassified

21. Price:

N/A

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at
325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.gov.