# An objective method for combining multiple subjective data sets

M. Pinson and S. Wolf

Institute for Telecommunication Sciences (ITS), National Telecommunications and Information Administration (NTIA), U.S. Department of Commerce

## ABSTRACT

International recommendations for subjective video quality assessment (e.g., ITU-R BT.500-11) include specifications for how to perform many different types of subjective tests. In addition to displaying the video sequences in different ways, subjective tests also have different rating scales, different words associated with these scales, and many other test variables that change from one laboratory to another (e.g., viewer expertise and criticality, cultural differences, physical test environments). Thus, it is very difficult to directly compare or combine results from two or more subjective experiments. The ability to compare and combine results from multiple subjective experiments would greatly benefit developers and users of video technology since standardized subjective data bases could be expanded upon to include new source material and past measurement results could be related to newer measurement results. This paper presents a subjective method and an objective method for combining multiple subjective data sets. The subjective method utilizes a large meta-test with selected video clips from each subjective data set. The objective method utilizes the functional relationships between objective video quality metrics (extracted from the video sequences) and corresponding subjective mean opinion scores (MOSs). The objective mapping algorithm, called the iterated nested least-squares algorithm (INLSA), relates two or more independent data sets that are themselves correlated with some common intermediate variables (i.e, the objective video quality metrics). We demonstrate that the objective method can be used as an effective substitute for the expensive and time consuming subjective meta-test.

Keywords: single stimulus continuous quality evaluation (SSCQE), double stimulus continuous quality scale (DSCQS), comparison, correlation, video quality, image quality, subjective testing, objective testing.

## 1. INTRODUCTION

International recommendations for subjective video quality assessment such as ITU-R BT.500-11 [1] include specifications for how to perform many different types of subjective tests. These specifications include conditions under which subjective video quality testing should be performed: lighting conditions, monitor resolution, monitor contrast, viewing distance, viewing angle, subjective scale, video presentation, the number of viewers to be used, and so forth. Tests performed in accordance with ITU-R BT.500 yield reasonably robust subjective video quality measurements.

However, multiple tests that adhere to the same testing methodology identified in ITU-R BT.500 may still differ from each other in significant ways. Test variables such as time of day, physical location, viewer expectations, and the range of video quality displayed can all impact the subjective mean opinion scores (MOSs). Thus, MOSs from different subjective experiments are not in general directly comparable, even when they are performed in the same laboratory using the same subjective testing methodology.

As an illustrative example, consider the Full Reference Television (FR-TV) Phase 1 subjective test performed by the Video Quality Experts Group (VQEG) [2]. Each laboratory was given an identical copy of the two viewing tapes and performed the identical experiment according to ITU-R BT.500-8. In this case, the Double Stimulus Continuous Quality Scale (DSCQS) was used, where "0" represents quality as good as the reference, and "100" represents quality much worse than the reference.

Results from the four laboratories for the 525-line low quality range are plotted in Figure 1. In Figure 1, the average Hypothetical Reference Circuit (HRC) quality (from all four laboratories taken together) is plotted versus the individual laboratory's HRC quality. Here, each data point represents averaging over both scenes and viewers, so that the confidence of each data point is very high. HRC quality can be thought of as the quality of a given video system or HRC, independent of scene. HRC quality is normally a principal component, or contributor, to the overall variance of the subjective data. The method of plotting HRC quality as shown in Figure 1 provides a clear visual means for examining differences between the various laboratories.

Notice that each laboratory's data is scaled differently with respect to the overall estimate of the underlying truth (i.e., the Average HRC Quality as shown on the y-axis). For example, the viewers from laboratory 7 used more of the quality scale than did the viewers from laboratory 3 (i.e., the laboratory 7 quality scores exhibit a gain factor with respect to the laboratory 3 quality scores). Also notice that the laboratory 2 quality scores are shifted to the right with respect to the laboratory 5 quality scores.

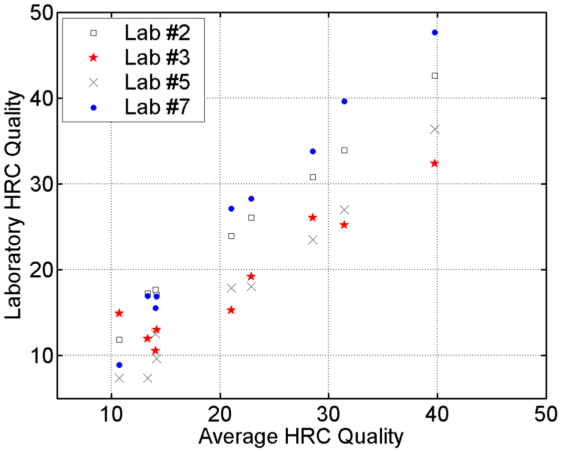

Figure 1. VQEG FR-TV Phase-1 525-line low quality test, plotting the average HRC quality from all four laboratories taken together versus the laboratory HRC quality.

Figure 2 is like Figure 1 except that it depicts laboratory HRC quality linearly mapped to average HRC quality using least-squares fitting. Since the average HRC quality is the best estimate of the underlying truth, the linear scaling factors shown in Figure 2 show how each laboratory's subjective scores are gained and DC-shifted with respect to this underlying truth. Note that substantial gains and offsets may be required to bring the individual laboratory's scores to one common scale (which in this case is defined by the Average HRC Quality as shown on the y-axis).

The above example from the VQEG FR-TV phase-1 data demonstrates that subjective scores are relative rather than absolute. In other words, one can obtain consistent indications from one subjective experiment to another on the relative quality difference between two HRCs (e.g., HRC 1 is lower quality than HRC 2), but absolute quality ratings are not in general repeatable even when the utmost care is taken. Direct comparisons between subjective ratings from multiple subjective tests require an additional step, that of placing all the subjective data onto one common scale. This paper presents a subjective and objective method for performing a common scale mapping. Results from both methods are given for the mapping of six original subjective data sets onto one common scale.

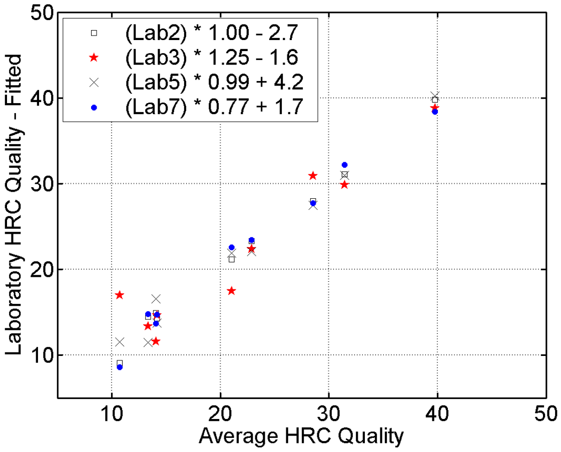

Figure 2. VQEG FR-TV Phase-1 525-line low quality test, plotting the average HRC quality versus the laboratory HRC quality after linear fitting.

The subjective method involves the use of a subjective meta-test. Carefully selected video clips from the original subjective tests are all combined into a large meta-test. The common subjective scale provided by the meta-test is used to estimate linear scaling factors for each original subjective data set. These scaling factors are then applied to the entirety of the original subjective data to map all scores onto the meta-test's common subjective scale. Throughout this paper 'original ratings' will refer to the subjective ratings associated with the original subjective experiment. 'Secondary ratings' will refer to the subjective ratings associated with the SSCQE meta-experiment performed on those same video clips.

The second method uses an objective mapping algorithm, called the iterative nested least squares algorithm (INLSA). This algorithm relates two or more independent data sets that are themselves correlated with some common external variables (i.e., objective video quality metrics). In essence, INLSA utilizes the functional relationships between objective video quality metrics (extracted from the video sequences) and their corresponding subjective mean opinion scores to deduce a mapping of the subjective data onto a single common scale.

## 2. ORIGINAL SUBJECTIVE EXPERIMENTS

The video sequences were selected from six different video quality experiments. These six original experiments examined impairments from television systems, with a particular emphasis on MPEG-2 coding impairments. Some of the video systems included MPEG-1 coding, VHS record/playback, multiple-generation dubbing with 1/2-inch professional tape recorders, and MPEG-2 bit-streams corrupted with digital errors.

Table 1 summarizes each of the six video tests, listing the subjective test method used, the number of viewers, the total number of video clips evaluated in the original experiment, the number of video clips selected for the secondary subjective experiment, and the version of ITU-R Recommendation BT.500 that was used. Data sets one, two, three, four and six are described in detail in [3], where the data set numbers shown in the table correspond to those that were used in [3]. Data set twelve was collected after the publication of [3] and is documented in [4].

Table 1 Original Subjective Tests

| Data Set | Method | Viewers | Total Clips | Clips Used | ITU-R BT.500 |
|---|---|---|---|---|---|
| One | DSCS | 32 | 42 | 20 | -3 |
| Two | DSCS | 32 | 105 | 20 | -3 |
| Three | DSCS | 32 | 112 | 20 | -3 |
| Four | DSCQS | 67 | 90 | 35 | -8 |
| Six | DSCQS | 80 | 90 | 70 | -8 |
| Twelve | SSCQE | 32 | 40 | 20 | -10 |

Data sets one, two and three employ the Double Stimulus Comparison Scale (DSCS) method (section 6.2.4.1 of [1]). Viewers are shown *pairs* of video sequences (the original reference sequence and the impaired sequence) in a randomized order. The viewers are then asked to rate the *difference* between the first and second video sequence on a discrete seven point scale. The viewers indicate whether the video quality of the second clip is much better, better, slightly better, the same, slightly worse, worse, or much worse than the first clip. The nominal range of the mean opinion scores is from 0 (no impairment) to -3 (maximum impairment), with some chance of a positive score occurring (e.g., when the coded output is perceived as being of higher quality than the reference). These three subjective experiments were performed by NTIA/ITS using 9 second long video sequences.

Data sets four and six were generated by the Video Quality Experts Group (VQEG) during the first phase of FR-TV testing [2]. Data set four corresponds to the 525-line high quality test, data set six to the 525-line low quality test. These data sets employ the DSCQS method. Like DSCS, DSCQS viewers are shown *pairs* of video sequences in a randomized order. Unlike DSCS, DSCQS viewers are shown each *pair* twice and then asked to rate the quality of each video sequences using a continuous scale. The *difference* between these two scores is then used to quantify changes in quality. The nominal range of this difference is from 0 (no impairment) to 100 (maximum impairment), with some chance of negative differences occurring (e.g., when the coded output is perceived as being of higher quality than the reference). Data sets four and six have some overlap, with all source (i.e., reference) scenes and two video systems being in common to both data sets. This overlap spans 20 video clips.

Data set twelve employs the Single-Stimulus Continuous Quality Evaluation (SSCQE) method. SSCQE viewers are shown an arbitrarily long video sequence nonstop. While watching the video, viewers indicate their current opinion on a continuous slider mechanism with an associated scale. Data set twelve uses hidden reference removal, a second stage in post-processing of the SSCQE scores that is also being proposed by VQEG for the upcoming Reduced Reference No Reference Television (RRNR-TV) test [5]. With hidden reference removal, the original video sequences are presented, but viewers are not aware that they are evaluating the original video. The viewer's opinion of the original video sequence is subtracted from the viewer's opinion of the distorted video sequence, analogous to the DSCQS method. The quality scores are scaled such that the nominal range of the difference is from 0 (worst quality) to 100 (best quality), with some chance of excursions above 100 (e.g., when the coded output is perceived as being of higher quality than the reference). Data set twelve uses ten 1-minute video sequences run through four video systems. The video sequences were split into two viewing sessions, with two orderings for each session.

Notice that the original subjective scores for data sets one through six contains only one score for each 8-10 second sequence, whereas the original subjective scores for data set twelve consists of subjective ratings produced every half second. Preliminary analysis we performed on an unpublished subjective data set provided by the Communications Research Centre (CRC) showed that the SSCQE rating located at the end of the 8-10 second video sequence is most highly correlated with the DSCQS score of that same sequence.[1] This observation was verified using the subjective meta-test data presented in this paper [4]. Therefore, for all analyses, the SSCQE ratings at the end of the video sequence are used; and all other SSCQE ratings are removed from consideration.

Table 2 summarizes the nominal dynamic range for the MOSs from each of the three methods that were used to evaluate video quality.

Table 2 Dynamic Range of the Subjective Scales

| Method | Best Quality | Worst Quality |
|---|---|---|
| DSCS | 0 | -3 |
| DSCQS | 0 | 100 |
| SSCQE | 100 | 0 |

## 3. SUBJECTIVE META-TEST METHOD FOR COMBINING MULTIPLE DATA SETS ONTO ONE COMMON SCALE

The subjective meta-test method for combining multiple data sets consists of the following five steps. (1) Carefully select video sequences from the six original experiments such that the full range of quality is represented and is uniformly distributed throughout this range. (2) Combine these sequences into a new pool of video sequences. (3) Subjectively rate the new pool of video sequences in a secondary subjective meta-test. (4) Using least-squares fitting, compute the optimal linear scaling factors for mapping the original mean opinion scores of each data subset to the secondary mean opinion scores. (5) Apply these linear scaling factors to the entirety of the original subjective data sets to map them onto the secondary meta-test's scale, which will be defined as the common scale.

The six original video quality experiments contain a total of 479 video sequences. Of these, we chose 185 distorted video sequences and their 30 associated reference video sequences for inclusion into the subjective meta-test. Five scenes and four video systems were chosen for inclusion from each of the original subjective data sets one, two, three and twelve (for data set twelve, five 9-second scene segments were selected from the 1-minute scenes). An indicator for the coding difficulty of each original scene was obtained by averaging the subjective mean opinion scores for that scene across all video systems. Scenes were then chosen to evenly span the full range of available quality in each test. The selection of the original video systems to include in the secondary test was performed in a similar manner.

Data sets four and six are unique, having a high accuracy (i.e., a large number of viewers) and an overlap of 20 clips between the two data sets. Therefore, additional clips were selected from data sets four and six for inclusion into the secondary subjective meta-test. Thirty-five video clips were chosen from data set four, including all the clips that overlapped with data set six. All the video clips from data set six were chosen *except* the overlapping clips with data set four. Since the video content for the overlapping clips is truly identical (i.e., stored on a digital medium), this exclusion is dictated by the data analysis and not the unavailability of secondary subjective MOSs. This exclusion was necessary in order to produce subjective and objective mapping functions for data sets four and six that were independent of each other. Even with the exclusion, data set six had twice as many clips as data set four, which is appropriate since data set six spanned about twice the range of quality as data set four.

Having selected our pool of 185 video sequences, we performed the subjective meta-test using SSCQE with hidden reference removal. To simplify analysis, the five scenes from each data set were treated as a 45-second super-scene. For data sets four and six, which had 8-second video clips, an extra 5 seconds of video was inserted at the beginning of the super-scene to fill out the 45 seconds (the scores from these extra 5 seconds were not intended to be analyzed). Test presentation ordering was randomized over both super-scenes and video systems, with the added constraint that the same super-scene or video system would never appear twice in a row. A panel of 20 viewers was split into two groups, with each group seeing one of the two possible orderings. Each viewing session was approximately one half hour in duration.

After all meta-test data had been collected and processed for hidden reference removal, the subjective SSCQE score at the *end* of each of the 185 video clips was retained and

[1] This data was received by means of a private communication with Philip J. Corriveau at Communications Research Centre (CRC), Ottawa, Ontario, Canada.

all other SSCQE scores were discarded. A first order linear predictor was trained for each of the six data subsets separately, using the original subjective ratings as the predictor variable (x-axis) and the secondary subjective ratings as the response variable (y-axis). The resulting gains and offsets were then used to map all of the original subjective data sets onto the secondary mean opinion scale.

The drawback of the subjective meta-test procedure is the time, expense, specialized hardware, and expertise that is required. The addition of a new data set would also require repeating the entire meta-test process again, with each subsequent meta-test becoming longer and more involved than the one before. Let us now consider a much less expensive alternative, a purely objective means for combining multiple data sets onto one common scale.

## 4. OBJECTIVE METHOD FOR COMBINING MULTIPLE DATA SETS ONTO ONE COMMON SCALE

### 4.1 Iterative least squares algorithm

The iterative least squares algorithm (INLSA) [6] objectively maps multiple subjective data sets onto a single scale by means of some common external variables (i.e., the video quality metrics). INLSA contains two least-squares problems. One attempts to homogenize heterogeneous data sets through the use of a single first-order correction for all of the data points in each data set. The other solves for the approximate linear combination of the parameters, across all data sets. By iterating over these two least-squares problems, the algorithm provides an optimal procedure for the simultaneous computation of the model weights and subjective data set scaling factors.

Before applying INLSA, the original subjective MOSs from each data set should be normalized to lie between zero and one, where zero represents no impairment and one represents maximum impairment. This provides a common scale and serves as a good starting point for proper convergence of INLSA. This normalization is accomplished by the transformation

$$\boldsymbol{s}_i = (\boldsymbol{s}_i^{\mathrm{o}} - best_i) \,/\, (worst_i - best_i), \qquad (1)$$

where $\boldsymbol{s}_i^{\mathrm{o}}$ is the original score vector for the $i^{\mathrm{th}}$ subjective test, $\boldsymbol{s}_i$ is the corresponding normalized score vector on [0, 1], $best_i$ is the no impairment value of the $i^{\mathrm{th}}$ original subjective scale, and $worst_i$ is the maximum impairment value of the $i^{\mathrm{th}}$ original subjective scale. For example, $best_i = 0$ and $worst_i = -3$ for DSCS (see Table 2).

Successive iterations of INLSA perform three steps. Step one holds the objective parameter weights constant and tries to bring all subjective scores onto a common scale. This least-squares treatment provides a single first-order correction for all of the scores in each subjective test,

$$\tilde{\boldsymbol{s}}_i = a_i\boldsymbol{s}_i + b_i\boldsymbol{1} \qquad (2)$$

where $\boldsymbol{s}_i$ is the normalized score vector on [0, 1] for the $i^{\mathrm{th}}$ subjective test from equation (1), $\tilde{\boldsymbol{s}}_i$ is the corresponding corrected score vector, and $\boldsymbol{1}$ is a column vector of ones.

Step two holds the data set weights ($a_i$, $b_i$) constant and searches for the optimal combination of video quality parameters. Given $r$ video quality parameters and $n$ video sequences taken from $m$ different subjective tests, form $r$ column vectors $\boldsymbol{p}_i$, $i$=1 to $r$, where each column vector has length $n$. Thus, $\boldsymbol{p}_i$ contains the values of the $i^{th}$ parameter for the $n$ video sequences. We then build the $n$ by $r$ parameter matrix

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_r]. \qquad (3)$$

Arrange the $n$ corresponding subjective scores from equation (2) into the length $n$ column vector $\tilde{\boldsymbol{s}}$. We can then solve the least-squares problem

$$\tilde{\boldsymbol{s}} \approx \hat{\boldsymbol{s}} = \boldsymbol{Pw} \qquad (4)$$

to find the set of weights $\boldsymbol{w}$ that describe the linear relationship between the parameters and subjective scores. Equation (4) defines the second step used by the iteration process.

As defined so far, INLSA has two excess degrees of freedom (one $a_i$ and one $b_i$) that preclude a unique solution. This situation is easily remedied by constraining $a_j = 1$ and $b_j = 0$, for some value of $1 \le j \le m$ (i.e., for the $j^{th}$ subjective video quality test). The third step enforces this constraint by shifting and scaling $\{a_i\}_{i=1..m}$, $\{b_i\}_{i=1..m}$, and $\boldsymbol{w}$. Thus, when INLSA completes, all subjective scores are scaled to the $j^{th}$ subjective video quality test, which has been normalized to [0, 1] by equation (1).

Figure 3 depicts steps one and two of INLSA. The least-squares problem defined by equation (2), depicted on the left, identifies appropriate values for $\{a_i\}_{i=1..m}$, $\{b_i\}_{i=1..m}$, while the least-squares problem defined by equation (4), depicted on the right, identifies $\boldsymbol{w}$. INLSA iteratively performs these steps to improve the estimates for $\{a_i\}_{i=1..m}$, $\{b_i\}_{i=1..m}$, and $\boldsymbol{w}$ until the convergence criteria have been met.

The above summary omits an important piece of INLSA [6]. The subjective scores $\boldsymbol{s}$ and the combined parameters

$\hat{\boldsymbol{s}}$ are both noisy measurements of the true underlying MOSs. The subjective scores $\boldsymbol{s}$ have errors because they are estimated means based on a finite number of viewers influenced by environmental factors unique to that particular video quality test. The video quality parameters have errors because they are only estimators of the true underlying perceived video quality. So, rather than implementing a standard least-squares solution, we would like to be able to apportion the least squares fitting error between both sources. To this end, INLSA utilizes a Direct Estimation (DE) algorithm [7] to find $\{a_i\}_{i=1..m}$, $\{b_i\}_{i=1..m}$ when given a cost-weighted error power ratio that specifies how to distribute the total fitting error between the subjective scores and the parameters. The DE algorithm provides a low complexity closed form solution by restricting the problem to the utilization of a single piece of side information (cost-weighted error power ratio) and estimating only a scalar gain and bias. For the results in this paper, we choose to distribute the fitting error equally between the subjective scores and video quality parameters (i.e., error power ratio is equal to one in the DE algorithm).

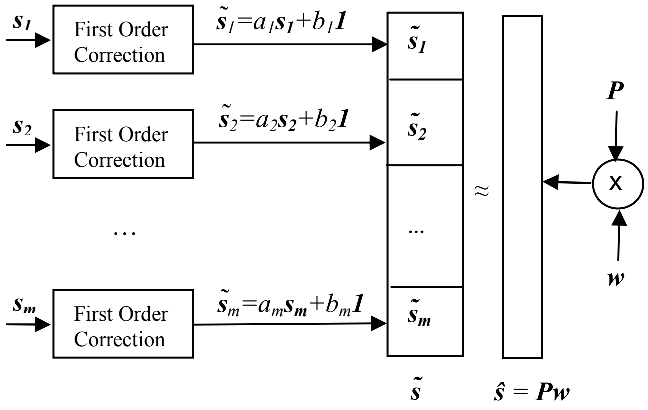

Figure 3. Block diagram depicting one iteration of INLSA.

Without loss of generality and to simplify analysis, the combined data set produced by INLSA was scaled using a first order least-squares fit to the secondary subjective data set. The resulting coefficients were applied to the INLSA gain and offset values as well as the original subjective data. Thus, the gains and offsets produced by INLSA are placed on the same scale as the secondary meta-test data (i.e., the SSCQE zero to 100 point scale shown in Table 2). The Pearson correlation coefficient between the INLSA mapped subjective data for all six data sets and the secondary meta-test data is not impacted by this scaling.

### 4.2 Objective video quality metrics

The objective video quality parameters used to train INLSA were taken from our "General" video quality model. This model was specifically developed by ITS to work over a very wide range of video quality (i.e., low bit rate video conferencing to high end television applications). The General model [3] can be calculated using the Video Quality Metric (VQM) software [8]. This model was trained on 1563 video clips taken from eleven video quality data sets. Original subjective data sets one through six were part of the eleven data sets used to train the General model [3] (data set twelve was collected after the General model was finalized and hence was not used for training).

Table 3 identifies the ability of the General model to predict the original and secondary subjective data. The values listed in Table 3 are calculated as the root mean square error (RMSE) of a first order linear regression, run with the general model as the predictor variable (x-axis) and either the scaled original subjective data (i.e., the SSCQE zero to 100 point scale shown in Table 2) or the subjective meta-test data as the response variable (y-axis). Since all subjective data have been placed onto the same 100-point scale, the RMSE values in this table can be directly compared to RMSE values occurring later in this paper.

Table 3 RMSE of General Model to Scaled Original Data and Secondary Data

| **Data Set** | **General to Scaled Original Data** | **General to Secondary Data** |
|---|---|---|
| One | 6.37 | 5.37 |
| Two | 7.93 | 10.04 |
| Three | 6.67 | 8.62 |
| Four | 3.80 | 5.94 |
| Six | 6.73 | 7.65 |
| Twelve | 12.25 | 11.84 |
| **All** | **8.62** | **9.27** |

Because the total variance of each original data set is different (i.e., each original data set spans a different range of video quality), the Pearson linear correlation coefficient does not tell the whole story. For example, when the variance of a subjective data set is small (e.g., a small fraction of the full range of the subjective scale), a poor-looking correlation coefficient (e.g., 0.7) could produce accurate estimates in a RMSE sense. Thus, RMSE can complete the picture because it provides an estimate to the prediction errors with respect to a common video quality scale, which in our case is the 100-point secondary meta-test scale. For those unfamiliar with RMSE, this is an approximation of the standard deviation of the residuals (or observed errors) when the predictor variable is used to estimate the response variable [9].

## 5. RESULTS

### 5.1 Subjective meta-test results

The calculated meta-test scaling factors are listed in Table 4, along with the RMSE associated with each linear regression. This table includes the 95% confidence interval for each weight (i.e., gain and offset). Notice that the scaling factors for data sets one, two and three (which used the DSCS method) are very similar. The gains for data sets four and six (which used the DSCQS scale) are not as consistent and may be due to original viewers' adjustment of their ratings based on the range of quality in the test. Since the range of quality in original data set six is about twice that of data set four, viewers may have tended to expand their ratings in data set four to fill the scale (i.e., apply a gain factor). Thus, when data set four is placed on the secondary scale, it must have a smaller gain than data set six. Note that even though data set twelve and the secondary meta-test data used the same SSCQE scale, there is a significant gain required to map data set twelve to the meta-test scale.

Table 4 Subjective Meta-test Mapping: Scaling Factors & RMSE

| Data Set | Gain | Offset | RMSE |
|---|---|---|---|
| One | 18.59 ± 2.64 | 98.08 ± 4.69 | 4.55 |
| Two | 18.21 ± 2.53 | 98.08 ± 3.85 | 5.55 |
| Three | 17.82 ± 4.57 | 96.41 ± 7.51 | 9.60 |
| Four | -0.65 ± 0.18 | 94.19 ± 3.07 | 5.05 |
| Six | -0.91 ± 0.10 | 97.79 ± 2.69 | 5.70 |
| Twelve | 0.80 ± 0.16 | 12.93 ± 9.50 | 7.24 |

### 5.2 INLSA results

With the scaling factors produced by the subjective meta-test specified, let us now move to INLSA. In this section, INLSA will be evaluated in terms of its ability to replicate the subjective meta-test's mappings. We initialized INLSA two different ways to obtain greater understanding of the objective mappings this algorithm provides (this same examination would also have been appealing for the meta-test, but was not possible due to time and expense). The two initial conditions we looked at for INLSA were:

1. General model parameters run on the 185 clips present in the secondary meta-test.
2. General model parameters run on the 185 clips from the secondary meta-test plus an additional 274 video clips that were not in the secondary meta-test. This set of 459 video clips includes the entirety of the six original subjective data sets (except for the 20 overlapping clips from data set six, which were excluded for previously mentioned reasons).

The INLSA scaling factors from running INLSA on the 185 video clips selected for the secondary meta-test and the seven parameters of the General model are presented in Table 5. This corresponds to a scenario of choosing between a subjective meta-test and INLSA, where both use the same subsets of original video clips. The gains and offsets that lie within the 95% confidence interval of the subjective meta-test's values (Table 4) are identified by check marks. The RMSE for each data set was calculated by directly differencing the INLSA mapped original subjective data with the secondary subjective data.[2] Keep in mind that the RMSE values in Table 4 serve as a lower bound; the INSLA RMSE values cannot be smaller than those values.

Table 5 INLSA Mapping Using 185 Video Clips : Scaling Factors and RMSE

| Data Set | Gain | Offset | RMSE |
|---|---|---|---|
| One | 15.54 | 99.80 √ | 8.66 |
| Two | 16.25√ | 97.28 √ | 6.12 |
| Three | 21.01√ | 100.76 √ | 10.15 |
| Four | -0.91 | 94.93√ | 6.65 |
| Six | -0.95√ | 96.62√ | 6.40 |
| Twelve | 0.79√ | 18.05√ | 8.74 |

[2] Notice that this RMSE calculation differs from that performed for the subjective meta-test.

Table 6 lists the results when training INLSA on the aforesaid 459 video clips and the seven parameters of the General model. This corresponds to a scenario of choosing between a subjective meta-test spanning less than half of the available video clips (presumably for economical reasons) and INLSA utilizing all available video clips. Again, the gains and offsets that lie within the 95% confidence interval of the subjective meta-test's values are identified by check marks. For RMSE, the resultant INLSA mapping was evaluated using the 185 video clips that were present in the secondary meta-test.

Table 6 INLSA Mapping Using 459 Video Clips: Scaling Factors and RMSE

| Data Set | Gain | Offset | RMSE |
|---|---|---|---|
| One | 15.18 | 99.14 √ | 8.67 |
| Two | 16.27 √ | 95.05 √ | 6.01 |
| Three | 22.52 | 101.05 √ | 10.89 |
| Four | -0.75 √ | 93.98 √ | 5.71 |
| Six | -0.95 √ | 97.18 √ | 5.89 |
| Twelve | 0.80 √ | 17.78 √ | 8.85 |

Table 7 compares the relative correlation and RMSE performance of the meta-test and INLSA methods, where all comparisons are with respect to the secondary meta-test data (considered as one large data set). When examining this table, recall that the subjective meta-test has the highest possible correlation coefficient and the lowest possible RMSE. As all fits are linear, no fit can do better than the meta-test fit.

Table 7 Relative Correlation and RMSE Performance of Each Algorithm When Compared with Secondary Meta-test Data

| Algorithm | Correlation | RMSE |
|---|---|---|
| Subjective meta-test map using 185 video clips | 0.936 | 5.62 |
| INLSA map using 459 video clips | 0.914 | 6.90 |
| INSLA map using 185 video clips | 0.910 | 7.20 |

The examination of classification errors gives us another means of evaluating INLSA. Classification errors result when the mapped original subjective data and the secondary meta-test data lead to different conclusions on a pair of video clips. Classification errors are recommended in [10] as a means of evaluating the effectiveness of a VQM. The method used herein differs from [10] in that two subjective data sets are being compared with one another, rather than a subjective data set being compared with an objective data set. When the subjective quality scores of two video clips (*A* and *B*) are compared, three classifications are possible: *A* can either be greater than, identical to, or less than *B*, given the 95% confidence intervals for each data point. Let $\boldsymbol{m}_{Aj}$ and $\boldsymbol{c}_{Aj}$ be respectively the mean and 95% confidence interval for video clip *A* in subjective data set *j* (original or secondary); and likewise for video clip *B*. When $\boldsymbol{m}_{Aj} + \boldsymbol{c}_{Aj} < \boldsymbol{m}_{Bj} – \boldsymbol{c}_{Bj}$ we declare that $A < B$; when $\boldsymbol{m}_{Aj}$ - $\boldsymbol{c}_{Aj} > \boldsymbol{m}_{Bj} + \boldsymbol{c}_{B}$ we declare that $A > B$; and in all other cases we declare that $A = B$.

Classification errors result when the classification indicated by the mapped original data differs from the classification indicated by the secondary data. Four categories of results are possible:

Correct Decision: A correct decision is made when the secondary data and the mapped original data both indicate the same classification.

False Differentiation: A false differentiation occurs when the secondary data concludes that $A = B$ but the mapped original data concludes that $A < B$ or $A > B$.

False Tie: A false tie occurs when the secondary data concludes that $A < B$ or $A > B$ but the mapped original data concludes that $A = B$.

False Ranking: A false ranking occurs when the secondary data concludes that $A < B$ but the mapped original data concludes that $A > B$, or vice versa.

For each method of combining multiple data sets, Table 8 lists the percentage of pair-wise comparisons, where $A \neq B$, that fall into each classification. Keep in mind that the error classification distribution for the subjective meta-test serves as a lower bound.

Of the four types of classification errors, false ranking is probably the most objectionable. Notice that INLSA had only 0.08% more false rankings than did the subjective meta-test. For all comparisons, the 459-clip INLSA mapping performed only slightly better than the 185-clip INLSA mapping; and both perform respectably with regards to reproducing the subjective meta-test's mapping.

Table 8 Relative Classification Error Performance of Each Algorithm When Compared with Secondary Meta-test Data

| Algorithm | Correct Decision | False Tie | False Differentiation | False Ranking |
|---|---|---|---|---|
| Subjective meta-test map using 185 video clips | 76.6% | 5.1% | 18.2% | 0.08% |
| INLSA map using 459 video clips | 74.0% | 7.0% | 18.8% | 0.16% |
| INLSA map using 185 video clips | 73.8% | 7.2% | 18.8% | 0.16% |

### 5.3 Overlap analysis

The mappings produced by the subjective meta-test have been presumed to define the "true" mappings, within the estimated confidence bounds as determined by the subjective data. In this section, we will evaluate the performance of the subjective meta-test and INLSA mappings using another independent means. This independent method is only available for the 20 overlapping video clips that are in common to data sets four and six.

The original subjective scores for these 20 video clips were fed into a least squares fit, where the ratings from data set four were used to predict the ratings from data set six. Table 9 lists the resulting linear prediction coefficients along with their 95% confidence intervals. The relative gain and offset between data sets four and six can also be computed using the mapping weights in Table 4, Table 5, and Table 6, as follows:

$$\text{\textit{Relative Gain}} = \boldsymbol{g_4} / \boldsymbol{g_6}, \qquad (5)$$

$$\text{\textit{Relative Offset}} = (\boldsymbol{o_4} - \boldsymbol{o_6}) / \boldsymbol{g_6}, \qquad (6)$$

where $\boldsymbol{g_i}$ is the $i^{th}$ data set's gain, and $\boldsymbol{o_i}$ is the $i^{th}$ data set's offset. These relative gains and offsets are also shown in Table 9.

Examining Table 9, the relative gains and offsets between data sets four and six calculated from both INLSA methods and the subjective meta-test method all lie within the 95% confidence intervals produced by the overlap analysis. While INLSA appears to yield results that are closer to the overlap results, we cannot determine whether either method (meta-test or INLSA) is better or worse than the other.

Table 9 Relative Gain and Offset between Original Data Sets Four and Six, Calculated Four Different Ways

| Algorithm | Relative Gain | Relative Offset |
|---|---|---|
| Overlap analysis map using 20 original video clips | 0.919 ± 0.25 | 2.41 ± 3.73 |
| Subjective meta-test map using 185 video clips | 0.714 | 3.96 |
| INLSA map using 459 video clips | 0.789 | 3.37 |
| INLSA map using 185 video clips | 0.958 | 1.78 |

## 6. CONCLUSIONS

We have shown that subjective mean opinion scores from different video quality experiments cannot in general be compared without first mapping all scores to one common subjective scale. This is true even when the subjective experiments follow exactly the same subjective testing procedures. In other words, subjective mean opinion scores do not provide absolute quality ratings, but merely provide relative quality ratings of one video system or scene with respect to another. In order to better utilize results from multiple subjective tests for the development and evaluation of objective video quality metrics, all subjective quality scores must first be mapped to one common scale.

We have presented two methods for performing this mapping. One method uses a secondary subjective meta-test and is very expensive and time consuming. The other method uses objective video quality metrics together with a pure mathematical algorithm called INLSA that is inexpensive and easy to compute. Six subjective data sets were placed onto one common scale using both the subjective meta-test method and the INLSA method. Except for differences in the mapping gains for two of the six data sets, the INLSA gains and offsets agreed with the meta-test gains and offsets within the estimated precision of the meta-test. For one of the subjective data sets where the gain disagreed, a different independent analysis based on overlapping clips that were included in this data set showed that the INLSA gain was reasonable even though it differed from the meta-test gain. In conclusion, we have shown that INLSA can be a very effective and cost saving tool for combining multiple subjective data sets onto one common scale.

## 7. ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Stephen Voran for developing INLSA and the DE algorithm and for designing SSCQE testing devices for our subjective testing laboratory. We also thank Philip J. Correveau for his suggestions regarding SSCQE with hidden reference removal, and Paul Lemmon for constructing the SSCQE testing devices and conducting the secondary subjective meta-test.

## 8. REFERENCES

[1] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” Recommendations of the ITU, Radiocommunication Sector (available at www.itu.org).

[2] ITU-T COM 9-80-E, “Final report from the video quality experts group (VQEG) on the validation of objective models of video quality assessment,” approved for release at VQEG meeting number 4, Ottawa, Canada, Mar. 2000 (available at www.vqeg.org).

[3] S. Wolf and M. Pinson, “Video Quality Measurement Techniques,” NTIA Report 02-392, Jun. 2002 (available at www.its.bldrdoc.gov).

[4] M. Pinson and S. Wolf, “Comparing Subjective Video Quality Testing Methodologies,” SPIE Video Communications and Image Processing Conference, Lugano, Switzerland, Jul. 8-11 2003.

[5] VQEG “RRNR-TV Group Test Plan,” draft version 1.4a, Jun. 2002 (available at www.vqeg.org).

[6] S. D. Voran, “An Iterated Nested Least-Squares Algorithm for Fitting Multiple Data Sets,” NTIA Technical Memorandum TM-03-397, Oct. 2002 (available at www.its.bldrdoc.gov).

[7] S. D. Voran, “Estimation of System Gain and Bias Using Noisy Observations with Known Noise Power Ratio,” NTIA Technical Report 02-395, Sep. 2002 (available at www.its.bldrdoc.gov).

[8] M. Pinson and S. Wolf, “Video Quality Metric Software, Version 2,” NTIA Software/Data Product SD-03-396, Volumes 1-5, Oct. 2002 (software available at www.its.bldrdoc.gov).

[9] J. Neter, M. Kutner, C. Nachtsheim, and W. Wasserman, “Applied Linear Statistical Models, Fourth Edition,” The McGraw-Hill Companies, Inc, 1996.

[10] Technical Report T1.TR.72-2001, “Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics,” sponsored by the Alliance for Telecommunications Industry Solutions (ATIS) and accredited by the American National Standards Institute (ANSI) (available at www.atis.org ).