Comparing subjective video quality testing methodologies

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. Some of these test methods are double stimulus where viewers rate the quality or change in quality between two video streams (reference and impaired). Others are single stimulus where viewers rate the quality of just one video stream (the impaired). Two examples of the former are the double stimulus continuous quality scale (DSCQS) and double stimulus comparison scale (DSCS). An example of the latter is single stimulus continuous quality evaluation (SSCQE). Each subjective test methodology has claimed advantages. For instance, the DSCQS method is claimed to be less sensitive to context (i.e., subjective ratings are less influenced by the severity and ordering of the impairments within the test session). The SSCQE method is claimed to yield more representative quality estimates for quality monitoring applications. This paper considers data from six different subjective video quality experiments, originally performed with SSCQE, DSCQS and DSCS methodologies. A subset of video clips from each of these six experiments were combined and rated in a secondary SSCQE subjective video quality test. We give a method for post-processing the secondary SSCQE data to produce quality scores that are highly correlated to the original DSCQS and DSCS data. We also provide evidence that human memory effects for time-varying quality estimation seem to be limited to about 15 seconds.

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

1. INTRODUCTION

The double stimulus continuous quality scale (DSCQS) method of performing subjective tests is widely accepted as an accurate test method with little sensitivity to context effects (see Appendix 3 to Annex 1 in [1]). Context effects occur when subjective ratings are influenced by the severity and ordering of impairments within the test session. With DSCQS, context effects are minimized since viewers are shown *pairs* of video sequences (the reference sequence and the impaired sequence) in a randomized order. Viewers are shown each *pair* twice. After the second showing, viewers are asked to rate the quality of each sequence in the pair. The *difference* between these two scores is then used to quantify changes in quality. The resulting scores are not significantly impacted by memory-based biases from previously viewed video sequences (see section 6.3.2 of [1]). Since standard double stimulus methods like DSCQS[1] provide only a single quality score for a given video sequence, where a typical video sequence might be 10 seconds long, questions have been raised as to the applicability of these testing methods for evaluating the performance of objective real-time video quality monitoring systems.

By contrast, single stimulus continuous quality evaluation (SSCQE) allows viewers to dynamically rate the quality of an arbitrarily long video sequence using a slider mechanism with an associated quality scale. This relatively new method provides a means for increasing the sampling rate of the subjective quality ratings. Having subjective scores at a higher sampling rate would be useful for tracking rapid changes in quality and thus would be more useful for evaluating real-time quality monitoring systems. Proponents of the SSCQE methodology argue that it can be used to assess widely time-varying quality of long video sequences in a way that DSCQS cannot (see sections 6.3 and 6.4 of [1]). However, questions have been raised regarding the accuracy of SSCQE when compared to DSCQS. Since viewers only see and rate the quality of a single video stream (i.e., single stimulus with no immediate reference picture), contextual effects may be present. Individual viewers' scores might also drift over the course of the test (e.g., viewers might concentrate on

[1] Reference [1] presents a new form of double stimulus testing called the simultaneous double stimulus for continuous evaluation (SDSCE) that utilizes side-by-side presentation of the original and impaired clips rather than randomized time ordering. However, this method has the drawback that the viewer must shift attention between the right and left presentations.

moving the slider in the proper direction to track changes in quality and hence loose track of the absolute slider position on the rating scale), adversely impacting the method's reliability. Differences in viewers' reaction times to quality changes would also reduce SSCQE's accuracy.

This paper presents results from a recent subjective meta-experiment where video clips from six separate subjective tests were combined and evaluated using an SSCQE experiment. 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. Five of the original experiments were some form of double stimulus testing while one of the original experiments was itself an SSCQE experiment. By comparing the secondary mean opinion scores (MOSs) from the SSCQE meta-experiment to the original MOSs from the six original experiments, we have begun to find answers to the questions that have been raised. Do SSCQE and DSCQS yield inherently different subjective scores? Are contextual effects and the lack of an immediate reference picture problematic for SSCQE? To what extent does previously viewed video impact the viewer's current opinion score in SSCQE?

2. SUBJECTIVE TESTING

This section describes the three subjective video testing methodologies that were used, the original subjective experiments, and the meta-experiment that was used to obtain the secondary subjective ratings. Throughout this paper, a 'data set' will refer to one subjective video quality experiment.

2.1 SSCQE, DSCQS, and DSCS

One original data set and the secondary data set utilized SSCQE. Our SSCQE experiments used hidden reference removal, a data post-processing step that is not described in [1].[2] With hidden reference removal, the reference video sequences are presented during the test session, but viewers are not aware that they are evaluating the reference video. The viewer's opinion of the reference video sequence is subtracted from the viewer's opinion of the impaired video sequence. Since each opinion is in the range [0, 100] the difference is in the range of [-100, 100]. One hundred is added to the difference, shifting the range to [0, 200]. Here, 0 is the worst quality, 100 is the same quality as the reference, and values greater than 100 indicate quality better than the reference. Scores greater than 100 seem to be limited to the first several seconds of the video scene (i.e., viewers seem to require about 6 to 8 seconds to move the slider to the proper position after a scene transition from a low quality scene to a reference high quality scene). The hope was that SSCQE with hidden reference removal could provide time varying quality assessments with the added benefits of double stimulus testing. SSCQE with hidden reference removal has also been proposed by the Video Quality Experts Group (VQEG) for future tests of reduced reference - no reference (RRNR) objective video quality monitoring systems [2]. Viewer training instructions for our SSCQE experiments were taken directly from [2].

Two original data sets utilized DSCQS which is described in section 5 of [1]. One study indicates that DSCQS is not influenced by contextual effects (see Appendix 3 to Annex 1, in [1]). However, a problem we discovered with DSCQS testing is that viewers occasionally switch scores (e.g., enter their opinion of the reference video where the impaired video score should be recorded, and vice versa.) This can be demonstrated by an examination of individual viewer scores for one video sequence from subjective laboratory seven for the VQEG 525-line low quality full-reference test [3]. In Figure 1, most individual viewer scores are scattered in a rough distribution around 77. The subjective response of -75 is indicative of viewer switching. Even though we have just plotted results for one video clip, viewer score switching seems to be a fairly common occurrence.

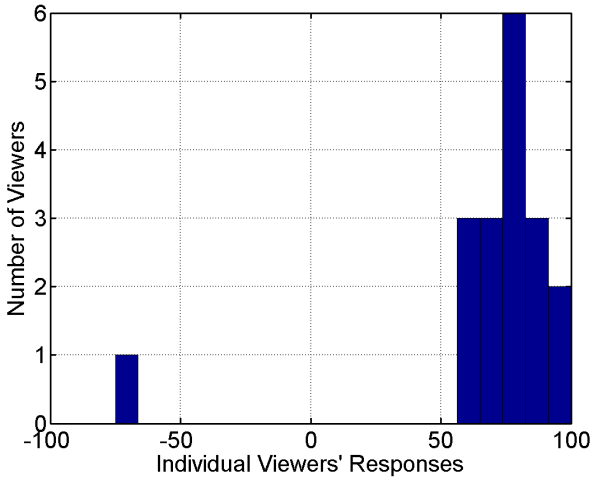

Figure 1. Histogram of individual DSCQS scores for one video sequence.

[2] SSCQE with hidden reference removal was first described to the authors of this paper by Philip J. Corriveau at Communications Research Centre (CRC), Ottawa, Ontario, Canada.

Three of the original subjective data sets employed a third testing methodology known as the double stimulus comparison scale (DSCS). DSCS is described in section 6.2.4.1 ITU-R Recommendation BT.500 [1]. DSCS viewers are presented with a pair of video sequences. Like the DSCQS method, the order within the pair is randomized, but unlike DSCQS, the pair is presented once instead of twice. With DSCS, the viewers directly rate the *difference* between the first and second video sequence on a discrete seven point scale (as opposed to DSCQS where both the first and second video sequences are each rated on a continuous quality scale). The viewers indicate whether the video quality of the second clip was better, worse, or the same as the first clip. The DSCS experiments used in this study were well-balanced with respect to order of presentation and range of video quality.

In summary, DSCQS viewers make two absolute ratings on a continuous scale at discrete times; DSCS viewers make one difference rating on a discrete scale at discrete times; and SSCQE viewers make absolute ratings on a continuous scale continuously over time.

2.2 Original subjective experiments

Video sequences for the SSCQE meta-experiment were taken from six different video quality tests. 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 method used, the number of viewers, the total number of video clips evaluated, and the version of ITU-R Recommendation BT.500 that was used.[3]

Data sets one, two, three, four, and six are described in detail in [4], where the data set numbers shown in the table correspond to those that were used in [4]. As a brief summary, subjective experiments one, two, and three were performed by NTIA/ITS and used scenes that were 9 seconds long. Subjective experiments four and six were performed by VQEG during their first round of full-reference television (FRTV) tests and used scenes that were 8 seconds long [3]. Data sets four and six have some overlap, with all source scenes and two video systems being in common to both data sets. Data set four included only high-quality television systems with a much more limited range of quality than data set six. The VQEG data sets have more viewers than the other data sets since four independent international laboratories were used.

Data set twelve was generated after the publication of [4]. Data set twelve was also performed by NTIA/ITS using SSCQE, modified for hidden reference removal as described earlier. Data set twelve used ten 1-minute video sequences and four video systems. The video sequences were split into two viewing sessions, with two random orderings for each session. Each viewer watched one session only, to minimize viewer fatigue. All viewers rated the reference video sequence as a hidden reference.

Table 1 Original Subjective Tests

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

2.3 Secondary subjective experiment

The secondary subjective test was performed by NTIA/ITS using SSCQE with hidden reference removal. Video clips were chosen from the six original data sets. This subjective meta-experiment was designed to last 30 minutes. A panel of 20 viewers was split into two groups, with each group seeing one of the two possible orderings.

To limit the duration of the secondary test, five scenes and four video systems were chosen from each of the original subjective data sets one, two, three, four, 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 MOSs for that scene across all video systems. Scenes were then chosen to evenly span the full range of available quality in each data set. Video systems were chosen in a similar manner.

The entirety of data set six was included in the secondary subjective test due to its high accuracy (i.e., large number of viewers), wide quality distribution, and because we wished to explicitly compare DSCQS quality scores with SSCQE quality scores. We will thus present a more

[3] We used the numbering established in [4] so that additional information on each data set could be located within that reference easily and accurately.

detailed quality comparison analysis for data set six. Also because of the overlap between data sets six and four, more video clips were included from data set four than from data sets one, two, three, and twelve.

To allow time for SSCQE scores to stabilize for hidden reference removal and to simplify analysis, the five scenes from each data set were treated as a 45-second super-scene. To minimize the magnitude of slider adjustments required within the super-scene, original clips were ordered from easy-to-encode to hard-to-encode. For data sets four and six, which have 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 super-scenes and video systems, with the added constraint that the same super-scene or video system would never appear twice in a row.

Figure 2 plots the average viewer time response for a super-scene from the secondary experiment (i.e., averaged across all viewers, super-scenes, and video systems under test). Two curves are shown on the plot, one including all of the data and the other excluding data set twelve, since the clips within data set twelve's super-scene were mistakenly ordered from most difficult to encode to easiest to encode.

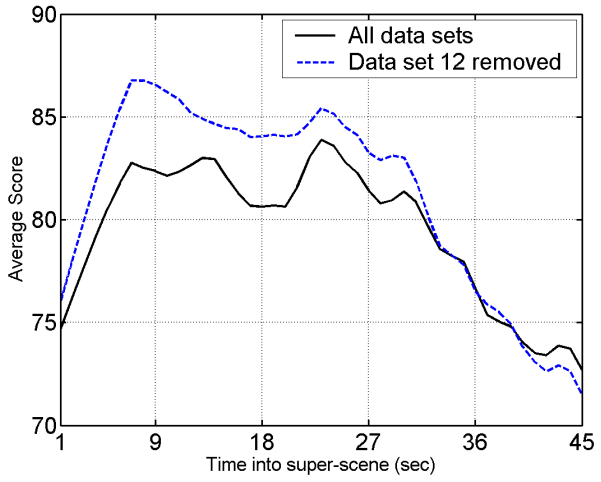

Figure 2. Score within super scene, averaged across viewers, super-scenes, and systems under test.

Notice that the ordering of scenes within the super-scene results (on average) in a sharp increase in viewer score during the first seven seconds and a gradual fall off of scores during the last 20 seconds. The sharp increase represents the sharp transition from the hard-to-code clips at the end of the previous super-scene to easy-to-code clips at the start of the current super-scene. The gradual fall off is due to increasing coding difficulty within the current super-scene. The above effect is more pronounced for the curve that has data set twelve removed.

2.4 Putting scores on a common scale

Although all six original subjective tests conformed to some version of Rec. ITU-R Recommendation BT.500, each test had its own unique set of variables that influenced the subjective scores (e.g., subjective scale, viewer expectations, and range of video quality included in the test). Thus, even when two subjective tests utilize exactly the same subjective testing methodology and scale, there is normally some gain and shift of one set of subjective scores with respect to the other. Having the secondary subjective test ratings allowed us to linearly scale and shift each of the original data sets (using linear regression techniques) to one common scale, namely, the secondary subjective scale. This simplified our analysis and data set comparisons without affecting computation of such things as Pearson correlation coefficients.

For the linear regression (and all other analyses), each original data set was considered separately. In other words, each original data set had its own unique scaling and shifting factors that were used to map the original ratings to the secondary subjective scale. Thus, the 20 original clips that were in common to data sets four and six will appear twice on the secondary scale, once from data set four and once from data set six. Likewise, the analysis to be presented includes 205 video clips even though the viewers rated only 185 video clips in the secondary subjective test.

3. REPRODUCIBILITY OF SUBJECTIVE RATINGS

The SSCQE method produces a subjective rating every half second, whereas the DSCQS and DSCS methods produce one subjective rating for each 8-10 second long video clip. There has been work on methods to convert SSCQE subjective ratings to DSCQS subjective ratings. One study examined both a 10 second averaging process of the SSCQE ratings and a non-linear averaging process, yielding promising results [5]. Another study found that the DSCQS subjective rating of a 30 second video clip is different than the average SSCQE subjective rating for those 30 seconds [6].

Preliminary analysis we performed on subjective ratings provided by the Communications Research Centre (CRC) showed that the SSCQE rating located at the end of each

video clip is highly correlated with DSCQS scores of those same video clips.[4] This observation was verified for the subjective meta-experiment presented in this paper. Thus, unless otherwise specified, the SSCQE ratings at the end of the 8 or 9 second video clip are used for all comparisons of secondary and original ratings. These comparisons include Pearson linear correlation coefficients and root mean square error (RMSE). RMSE values are calculated as the root mean square error of a linear regression, where the scaled original subjective data is the predictor variable (x-axis) and the secondary subjective data is the response variable (y-axis). RMSE is an approximation of the standard deviation of the residuals (or observed errors) when the scaled original subjective data is used to estimate the secondary subjective data [7].

Because the total variance of each original data set is different (i.e., each original data set spans a different range of video quality), correlation coefficients do not tell the whole story. RMSE can complete the picture because it provides an estimate of the prediction errors with respect to a common video quality scale, namely the secondary subjective data scale.

3.1 Repeatability

For each data set, Figure 3 plots the secondary subjective ratings versus the original subjective ratings (after scaling and shifting). A score of 100 corresponds to excellent quality.

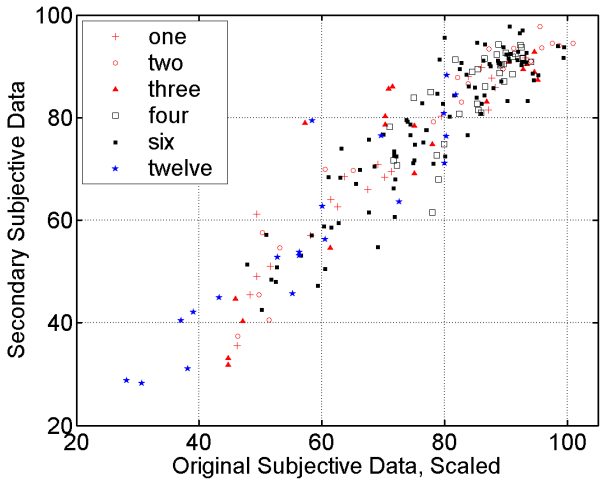

Figure 3. Secondary versus original subjective ratings, plotted by data set.

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

Table 2 compares the original and secondary subjective scores for each data set. The column labeled 'Clips' contains the number of video clips held in common between the two tests. The 'Variance' column is the total variance of those clips within the original data set after scaling to match the secondary data's 100 point scale. The 'Correlation' and 'RMSE' columns provide comparisons of the original and secondary subjective scores. The bottom row labeled 'All' lists these values computed over the set of original video sequences selected for inclusion into the secondary subjective test. The correlation coefficients and RMSE values in Table 2 are impressive considering the different subjective test methodologies that were used.[5]

Data set four has a total variance significantly lower than all the other data sets. Notice that data set four's video clips are clustered close together at the high quality end of the scale in Figure 3, but the spread of points around the 45º line is comparable to that seen in other data sets. Data set four's correlation is significantly lower than the correlations for data sets one and two, but the RMSE values for these three data sets are similar in magnitude. This example demonstrates that simply looking at correlation coefficients can be misleading.

Table 2 Comparison of Individual Data Sets

Data Set	Clips	Variance	ρ	RMSE
One	20	238	0.961	4.55
Two	20	371	0.963	5.55
Three	20	325	0.888	9.60
Four	35	44	0.782	5.34
Six	90	170	0.917	5.70
Twelve	20	314	0.929	7.24
All	205	254	0.936	6.01

Data sets one, two, four, and six have RMSE values that are reasonably similar. Data sets three and twelve have substantially higher RMSE values. To understand these higher RMSE values, we examine the effect of clip presentation ordering of the original and secondary subjective tests. Because the results for data sets three

[5] For the secondary data, we also examined SSCQE without hidden reference removal. Here, the squared correlation coefficient (an estimate of the fraction of the variance explained) was reduced by 0.05 on average.

and twelve depend upon only 20 data points, single outliers in these data sets can have a significant impact on RMSE. Table 3 expands upon Table 2, listing results separately for each of the two secondary clip presentation orderings. Most of the correlations and RMSEs are reassuringly similar for order #1 and order #2.

The largest RMSE overall belongs to data set three. The largest difference between order #1 and order #2 correlations and RMSEs likewise belongs to data set three. Investigations revealed that most of data set three's differences seen in Table 3 are caused by a single outlier clip in the secondary data. This clip occurred very close to the beginning of order #2. The video system from which this clip came was a VHS tape recording. As studio quality reference video had not yet been presented (except during a training session), several viewers rated this VHS video clip as having perfect quality. By contrast, for order #1 this same video clip was presented later in the session after viewers had a chance to see very high quality video. For order #1, the resulting secondary subjective ratings compare more favorably to the original subjective ratings. The other four scenes that were presented as part of that same super-scene exhibit the same ordering effect, although to a lesser extent. Thus, multiple presentation orderings appear to be important for good SSCQE subjective test design. Ideally, each viewer should see a unique randomized order. Then, clip ordering effects are mitigated when averaging across viewers.

Table 3 Impact of SSCQE Clip Presentation Ordering

Data Set	Order 1 RMSE	Order 2 RMSE	Order 1 ρ	Order 2 ρ
One	3.62	6.54	0.976	0.920
Two	6.21	5.42	0.951	0.966
Three	8.02	12.40	0.931	0.802
Four	5.06	6.53	0.749	0.762
Six	6.31	7.25	0.894	0.882
Twelve	8.02	7.30	0.917	0.926
All	6.31	7.54	0.932	0.901

Table 4 shows how the correlation rises and the RMSE falls when the VHS outliers are removed from data set three. Notice that, once all five VHS outliers have been removed, the correlation and RMSE of the data set as a whole ('Overall ρ' and 'Overall RMSE') are better than those seen in Table 2 for the other data sets.

Close analysis of data set twelve identified one outlier. For both orderings of the secondary data, this video clip's subjective ratings are noticeably shifted away from the main cluster of data. This data point is plotted in Figure 4, at an original value of approximately 60 and a secondary value of approximately 80. This particular clip had a sharp drop in quality during the last 2 seconds. For some reason, the original viewers seem to have reacted faster to the rapid drop in quality than the secondary viewers. If this clip is removed from data set twelve, the correlation between the original and secondary data rises from 0.929 to 0.962 and the RMSE drops from 7.24 to 5.32. While the former RMSE (7.24) is somewhat higher than the usual (see Table 2), the latter RMSE (5.32) puts data set twelve into the range of data sets one, two, four, and six. With the six outliers from data sets three and twelve removed, the secondary subjective data set can be replicated from the original data sets with an RMSE of 5.7 or less.

Table 4 Influence of VHS Outliers on Data Set Three

Conditions	Order 2 ρ	Order 2 RMSE	Overall ρ	Overall RMSE
All 20 clips	0.802	12.40	0.888	9.60
Discard 1 VHS outlier	0.872	10.28	0.924	8.17
Discard all 5 VHS outliers	0.960	6.36	0.986	3.92

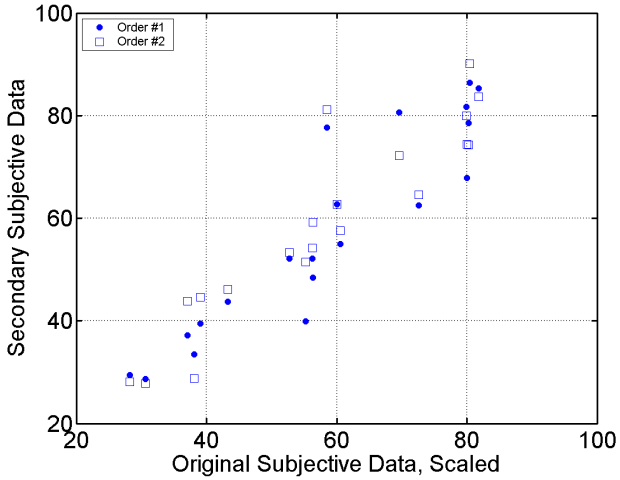

Figure 4. Secondary versus original subjective data set twelve, order #1 and order #2, plotted separately.

3.2 Lab to lab comparisons, data set six

Data set six was collected using four different testing laboratories, referred to in [3] as labs 2, 3, 5, and 7. Table 5 presents the cross correlations between the original labs and the secondary SSCQE data for data set six. The lab-to-lab correlations clearly show that lab 3 was an outlier. The reasons for this have not been documented. Correlations between SSCQE and labs 2, 5, and 7 (i.e., ranging from 0.902 to 0.929) are comparable to the correlations between labs 2, 5, and 7 (ranging from 0.913 to 0.935). These results indicate that the secondary SSCQE subjective ratings are nearly comparable to the original DSCQS subjective ratings from each of these three labs.

Table 5 Cross Correlations Between Original Labs and Secondary SSCQE Ratings for Data Set Six

	Lab2	Lab3	Lab5	Lab7
SSCQE	0.911	0.682	0.902	0.929
Lab7	0.933	0.727	0.935	
Lab5	0.913	0.807		
Lab3	0.747			

Table 6 lists the number of viewers (N) and confidence intervals (CI) for the original labs taken individually and altogether, and the secondary SSCQE data. For roughly the same number of viewers (i.e., 20), the CI for the secondary data appears to be comparable to the original data CI for one lab.

Table 6 Lab-to-Lab Comparison of Average Confidence Interval

	Lab2	Lab3	Lab5	Lab7	DSCQS	SSCQE
N	17	18	27	18	80	20
CI	8.12	10.19	5.27	8.19	4.05	5.94

3.3 Summary

The results presented here indicate that SSCQE may be used to approximate DSCQS and DSCS subjective ratings provided:

- SSCQE data is processed using the hidden reference removal technique.
- At least two randomized viewer orderings (i.e., sessions) are performed in the SSCQE experiment.
- The SSCQE rating at the end of the 8-9 second video clip is used (i.e., an 8-9 second stabilization period before sampling the SSCQE rating).

4. MEMORY EFFECTS

One interesting question to examine is the impact of human memory on SSCQE results. To what extent is the viewer's current impression of the video quality dependent upon past impairments? There is some evidence that viewers have non-symmetrical memory in that they are quick to criticize degradations in video quality but slow to reward improvements [5].

The impact of specific training instructions on the subject's memory (for quality estimation purposes) has not been extensively studied. It is reasonable to assume that these instructions (or lack thereof) could influence results of any investigation into SSCQE memory effects. For both data set twelve and the secondary experiment, our viewers were encouraged (in the training session) to respond to rapid changes in quality (see training instructions in [2]). Therefore, the results presented here might be different from what other researchers have found. We examine three approaches to assess the impact of memory effects on subjective results. These approaches are based on the repeatability of subjective scores and the importance of scene length.[6]

4.1 The nine second approach

The secondary SSCQE experiment used 8 to 9 second scene segments drawn from the original data sets in Table 2. If memory effects extended much beyond 9 seconds, we would expect to see a significant impact on the correlation coefficients and RMSE values that are presented in Table 2. In other words, if memory extended significantly beyond 9 seconds, why would instantaneous SSCQE data sampled at the end of each 9 second scene agree with the original subjective data where viewers only saw that 9 second scene? Thus, Table 2 by itself presents an argument that significant memory effects do not extend much beyond 9 seconds for our tests.

It is possible that some viewers may interpret scene changes as cues to update their opinions, thus artificially truncating their natural memory effect to 9 seconds. Let us now examine an approach that utilizes longer video sequences.

[6] The subjective testing community has debated the validity of using short program segments for SSCQE testing. For instance, section 6.3.1 of [1] recommends using program segments at least 5 minutes long.

4.2 Correlating SSCQE samples from two experiments

Recall that data set twelve's original and secondary subjective data both used the SSCQE method. Figure 5 plots the correlation (y-axis) between the original and secondary subjective ratings for clips in data set twelve, using just a single SSCQE sample taken x seconds into the nine second clips (x-axis). For example, the correlation for x = 1 second corresponds to original and secondary viewers having viewed only one second of common video for each of the 20 sequences in data set twelve.

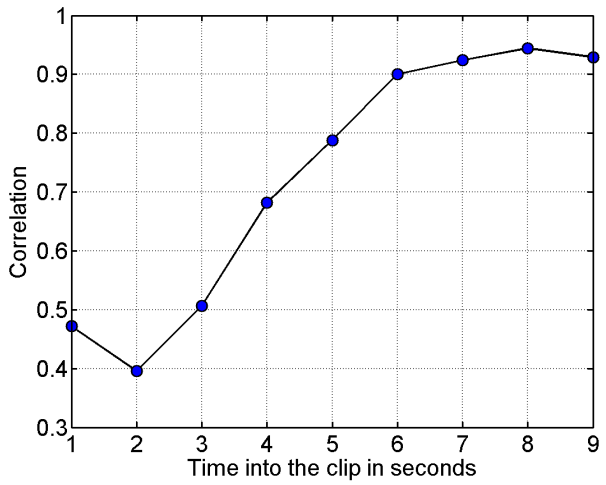

Figure 5. Correlation of original & secondary ratings using a single SSCQE sample extracted from each clip in data set twelve.

The minimum correlation level is influenced by the design of the original and secondary tests. Recall that for both the original and secondary experiments, the 9 second sequences were part of a longer sequence (i.e., a 1 minute clip for the original, and a 45-second super-scene for the secondary). Since these longer sequences were shown to viewers together for each system under test, the rating of the content prior to the beginning of the 9 second sequences was (for three out of five scenes) created using the same video system. The two scenes for which the previous content was uncorrelated were (1) the first clip displayed in the secondary subjective test's super-scene; and (2) a 9 second clip that was taken from the very beginning of one of data set twelve's 1 minute video sequences. For the other three scenes, the prior content was always from the same video system under test.

Between 1 and 3 seconds, the correlation is low. Between 7 and 9 seconds, the original and secondary data correlations are stable and relatively high (between 0.924 and 0.944). This plot indicates that on average, viewers used only the previous 7 or 8 seconds of video for forming their SSCQE quality decision in our tests.

4.3 Predicting order #2 from order #1 plus memory

When considering Figure 5, the correlations were based on just twenty video clips (i.e., five 9-second scenes and four systems under test). Therefore, these results do not constitute sufficient proof as to the duration of memory effects. Let us now consider a different analysis that utilizes many more data points. Recall that our secondary test was designed with two randomized viewing orders (order #1 and order #2), each of 30 minutes duration. The approach is visually depicted in Figure 6 and described as follows:

(1) For each second of video in order #1, find the corresponding second of video in order #2 that contains that same video content.

(2) Looking at the MOS time histories, build a linear regression to predict the current MOS of order #2 (value D) using the matching sample of order #1 (value A), the delta between matching sample A and the rating 'Time Ago' seconds prior (the length of B), and the delta between order #1 and #2 ratings at 'Time Ago' seconds prior (the length of C). Note that if 'Time Ago' = 0, the prediction should be perfect.

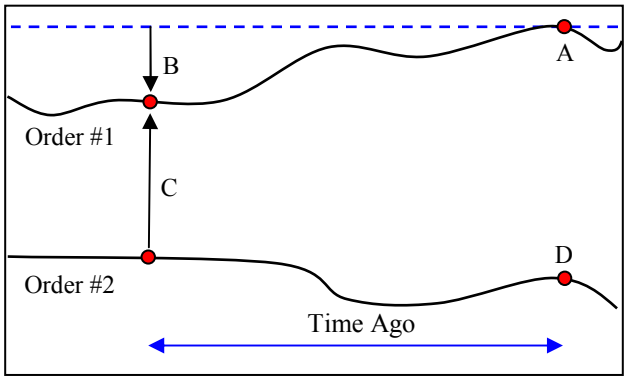

Figure 6. Predicting order #2: $D = w_0 + w_1A + w_2B + w_3C$.

(3) Use the resulting linear prediction model to assess how much of order #2's subjective rating is determined by order #1's subjective rating and the influence of past events. Value A indicates the ability of the order #1 MOS to predict the order #2 MOS. Value B is calculated within order #1 as the current value minus the value at Time Ago seconds. Value B's role is to eliminate order #1's memory effect from the model. Value C is calculated at Time Ago seconds as order #2 minus order #1. Value C's role is to introduce order #2's memory effect into the model.

Linear prediction weights (w_1, w_2, and w_3) for the model described in Figure 6 are plotted in Figure 7 as a function of 'Time Ago'. The constant term (w_0) has been left out for simplicity. For Figure 7, the current samples (A and D) are limited to the last sample in each of the 185 video clips, excluding the reference video sequences. Any term for which a 95% hypothesis test concludes that the term is statistically equivalent to zero has been plotted as being exactly zero. The w_1 term varies around 1.0, indicating that order #2's subjective data is on average quite similar to order #1's subjective data. The w_2 term is generally negative, which has the effect of removing order #1's memory effect. The w_3 term is generally positive, which has the effect of introducing order #2's memory effect. The w_2 and w_3 terms drop out of the equation after a 'Time Ago' of about 14 seconds, leaving only the w_0 and w_1 terms. The w_2 term appears again around 90 seconds later, due to the design of the secondary test which was based on 45 second long super-scenes (i.e., while the same super-scene is never presented twice in a row, it can be presented again after some other super-scene is presented). From this analysis, the viewer's current score does not appear to be influenced by video seen more than 14 seconds ago.

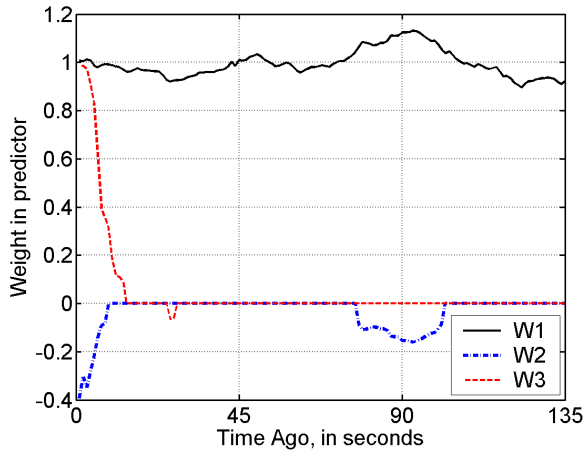

Figure 7. Linear predictor weights as a function of 'Time Ago'.

Figure 8 shows the correlation between the actual and predicted order #2 viewer opinion scores when weights are selected in four different ways. Model 'Secondary, end' is trained over the secondary subjective data, limited to predicting the score at the end of each impaired video clip. This model uses 185 data points and corresponds to Figure 7 above. Model 'Secondary, all' is trained over the entire half hour of continuous secondary subjective data. The third model, labeled 'Original, data set 12', is trained over the entirety of the original subjective data set twelve, comparing its order #1 to its order #2 (this original experiment was conducted in a similar manner to the secondary experiment, using two randomized viewing orders). The fourth model, labeled 'Original, dancers', is trained over the original subjective data set twelve's scene 'dancers'. This video sequence unambiguously contains 1 minute of continuous content. Notice that, for all four models, the correlation drops to a minimum level after approximately 15 seconds, which gives further evidence that video presented earlier than 15 seconds ago has little impact on the current SSCQE score.

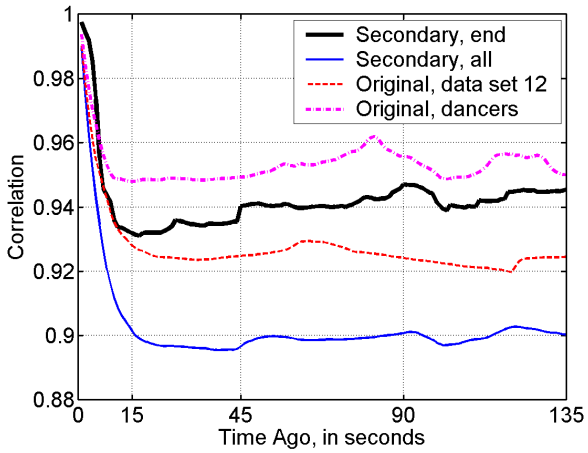

Figure 8. Influence of past events on current viewer opinion scores.

4.4 Summary

The analyses indicate that the viewers who participated in our SSCQE experiments considered at most the last 9 to 15 seconds of video when forming their quality estimate. This is not to say that long sequences are completely without other merits. Nonetheless, when long video sequences are used in SSCQE tests, test designers should not necessarily expect a panel of viewers to rate the video inherently differently than if shorter sequences are used.

5. CONCLUSIONS

SSCQE with hidden reference removal and multiple randomized viewer orderings (at least two) can produce quality estimates comparable to DSCQS and DSCS. The best method for performing this translation is to take the last SSCQE time sample of the scene and compare that with the DSCQS or DSCS value. This reassuring result shows that viewers perform essentially the same error pooling function (i.e., the judgment process where perceived errors distributed in space and time are mapped

to overall estimates of perceived quality) in SSCQE, DSCQS, and DSCS tests.

While the amount of prior video that is used to form the SSCQE quality estimate might be dependent on training instructions, our SSCQE test subjects appeared to utilize at most the previous 9 to 15 seconds of video. This fact, together with the demonstrated ability of SSCQE with hidden reference removal to replicate double stimulus testing results, has ramifications for simplifying subjective test design. Properly designed SSCQE testing (with short 9 to 15 second sequences) may be an effective substitute for more complicated DSCQS testing. The advantages of using SSCQE as a substitute would include faster testing (or more clips rated for the same amount of viewing time spent) and less viewer fatigue.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of Philip J. Corriveau for his suggestions regarding SSCQE with hidden reference removal and comparisons of this form of testing to DSCQS, Stephen Voran for designing SSCQE testing devices for our subjective testing laboratory, and Paul Lemmon for constructing SSCQE testing devices and conducting the secondary subjective experiment.

7. REFERENCES

[1] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Geneva, 2002 (available at www.itu.org).

[2] Video Quality Experts Group (VQEG), "RRNR-TV Test Plan" (available at www.vqeg.org).

[3] 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 Mar. 2000, Ottawa, Canada (available at www.vqeg.org).

[4] S. Wolf and M. Pinson, "Video Quality Measurement Techniques," NTIA Report 02-392, Jun. 2002 (available at www.its.bldrdoc.gov/n3/video/documents.htm).

[5] R. Hamberg and H. Ridder, "Time-varying Image Quality: Modeling the Relation between Instantaneous and Overall Quality," *SMPTE Journal*, Nov. 1999, p. 802-811.

[6] R. Aldridge, D. Hands, D. Pearson and N. Lodge, "Continuous quality assessment of digitally-coded television pictures," IEEE Proceedings online no. 19981843, Oct. 1997.

[7] J. Neter, M. Kutner, C. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models, Fourth Edition*, New York: The McGraw-Hill Companies, Inc, 1996.