Proposal: Measurement of a JND Scale for Video Quality

Andrew B. Watson
NASA Ames Research Center
Moffett Field, CA, 94035-1000
abwatson@mail.arc.nasa.gov
http://vision.arc.nasa.gov/

This document describes a research proposal to the IEEE G-2.1.6 Subcommittee on Video Compression Measurements.

1. *Research Background*

1.1 IEEE JND Project

In early 1998, the IEEE G-2.1.6 Subcommittee on Video Compression Measurements initiated a "Task Force to define a unit of measure and means of calibration for video quality analysis." This effort was lead by Leon Stanger. In the interim the Task Force has discussed various methods that might be used to derive an absolute scale of video quality. This document describes a specific proposal that might satisfy the needs of the Task Force.

1.2 VQEG Project

The Video Quality Experts Group (VQEG) has recently completed a large study comparing subjective data and predictions from a set of models(VQEG, 2000). The data consisted of observers rating 20 source videos (SRCs) as processed by 16 hypothetical reference circuits (HRCs). An HRC is a particular set of processing operations, such as compression at a particular bit-rate. About 300 observers took part in the VQEG study. The ratings were obtained using the Double Stimulus Continuous Quality Scale (DSCQS) method of ITU-R BT.500-8(ITU-R, 1998).

There are several problems with rating data of this sort. First, they are quite variable. Second, they are subject to criterion and context effects. For example, the ratings given will depend upon the range of quality used in the experiment. In addition, the scale on which they are rated has no inherent meaning, since different experiments use different scales and different ranges of quality. A final problem with these data, is that they used only a single viewing distance. Quality is known to vary markedly with viewing distance(Nakasu, Aoki, Yajima, Kanatsugu & Kubota, 1996), and it would be useful to test this property of the models. This proposal describes a program of research that addresses all of these problems.

2. *New Approach*

In the approach described here , rather than asking the observer to rate a given video, we ask the observer which of two videos is more impaired. This is called “pair comparison”, and also “two-alternative forced-choice.” From the responses to that simple question, we hope to measure the observer’s internal “perceptual scale” for visual impairment. The idea is that each video gives rise to a mental estimate of impairment. This perceptual impairment, as a function of increasing physical impairment, is what we mean by the perceptual scale. This scale would be measured in units of JND (just-noticeable-differences).

2.1 Thurstone Scaling

We derive the scale from the pair comparisons by means of Thurstone's "Law of Comparative judgement" (Thurstone, 1959). Thurstone proposed that physical sensory stimuli (such a sound) might give rise to sensory magnitudes arranged along a one-dimensional internal sensory scale (such as loudness), as pictured in Figure 1. However, the sensory magnitude varies from presentation to presentation, due to the unavoidable variability of neural systems. In one particular case (Thurstone's "Case Five"), the distributions are assumed to be Normal, with a standard deviation of 1 (as depicted by the yellow triangles). In that case, the probability of a correct judgement in a pair comparison is a function only of the distance between the sensory magnitudes induced by the two intensities of the pair. We can therefore estimate these distances by finding which values would most likely give rise to the data in hand. Mathematical details, and an example, will be given below in the Pilot Experiments.

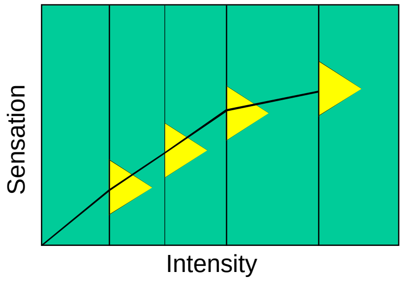

Figure 1. Thurstone Model.

In the research described below, we will use this general method to obtain quality estimates for a wide range of video materials (a subset of the VQEG materials).

2.2 Methods of estimation

A secondary goal of this project will be to develop an efficient method of scale estimation. In the preceding discussion, we did not address the question of which pairs to use. Clearly, some are more informative than others. We have begun development of an efficient method of pair selection, based on Bayesian statistical methods(Watson & Pelli, 1983). This idea is discussed in more detail later in this proposal.

3. General Methods

All of the experiments described in this proposal employ a set of common methods, which we describe here.

3.1 Source materials

Sequences will be selected from the collection created as part of the VQEG experiments (VQEG, 2000). This collection consists of 340 eight-second sequences. Each sequence is either an original source sequence (SRC) or that source sequence modified by a hypothetical reference circuit (HRC). The collection contained 20 SRCs, ten in 525/60 Hz format and ten in 625/50Hz format. There were 16 HRCs, consisting of various analog and digital processing stages. In the remainder of this document, we will identify a particular video by the syntax SRC-HRC, with the original source video considered HRC=0. For example, video 3-9 is the source 3 modified by HRC 9. Single frames from the 525 videos are shown in the following figure.

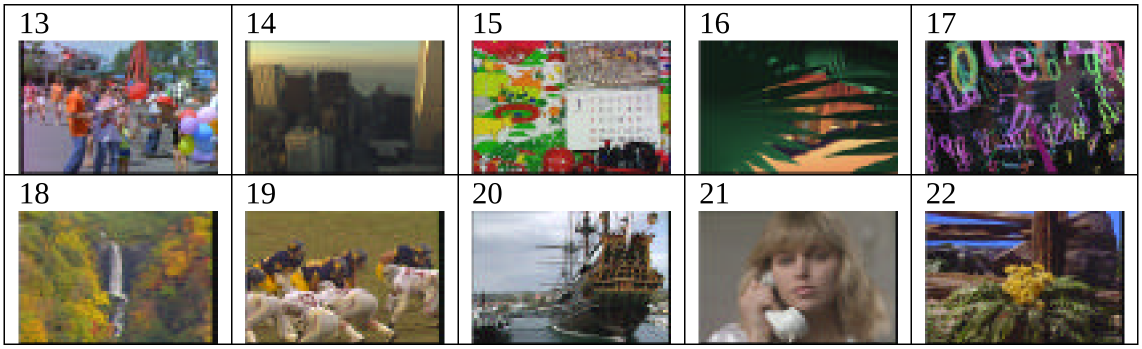

Figure 2. Single frames from 525/60Hz videos used in the VQEG study.

3.2 Blends

In some of the experiments, we will make use of "blends." A blend is a video that is the linear combination between two videos, typically the source video and that same video modified by a particular HRC. A blend is defined by the "source" video, the "sink" video, and the weight *w* (0<*w*<1) used to combine them. Thus

$$\text{blend}\left(\text{source},\text{sink}w\right) = \left(1-w\right)\text{source}+w\,\text{sink} \qquad (\ 1$$

The arithmetic above should be understood as being applied to the raw numbers within the ITU-601 file that specify the values of Y and downsampled Cb and Cr. Blends are a simple way of controlling, in a quantitative way, the amount of a particular artifact that is added to a source video(Brunnstroem, Eriksson & Ahumada, 2000).

In our pilot experiments we have used a series of 21 weights spaced logarithmically from 0.1 to 1. We often express these weights in units of 1/100ths of a log unit (base 10), which we call centiLogs, or cL. Thus the set of 21 weights correspond to 0, -5, -10, …-95, -100 cL. These are best thought of as attenuations of the sink. It is convenient to identify a particular blend by the syntax SRC.HRC.cL. For example, a video created from videos 2.0 and 2.10 with a weight of 0.63 = 10^(-20/100) = 20 cL would be identified as 2.10.20.

3.3 Viewing Conditions

Observers will view the sequences under Recommendation 500 viewing conditions. The viewing distance will be specified for each experiment in picture heights (usually 3 or 5).

3.4 Observers

For the pilot studies, the authors of this study will serve as observers. For subsequent studies, observers will be non-experts who will typically be paid for their services and will participate for a single 1-2 hour session. Observers will be checked for normal color vision and corrected-to-normal spatial acuity using standard eye charts.

3.5 Video Presentation

Each video will be presented under computer control and displayed on a studio quality television monitor capable of displaying ITU-601 digital video streams. In our laboratory, the display apparatus consists of an SGI Octane computer with SDI serial digital video input/output board, a Ciprico FibreChannel Disk Array, and a SONY BVM 20E1U monitor.

3.6 Psychophysical Procedures

In general, this research program will make use of 2-alternative, forced-choice (2AFC) methods. On each trial, the observer will be presented with a pair of videos, separated by a pause of 1 second. At the end of the second presentation, the observer will press a button to indicate whether the first or second video appeared more degraded. Audio feedback will tell the observer whether they were correct or incorrect. A QUEST adaptive staircase(Watson & Pelli, 1983; Watson & Solomon, 1997) will then be used to select the next pair of videos to be presented. Quest operates by estimating, after each trial, the most likely location of threshold. Figure 3 illustrates trials 1, 2, 4, and 32 from a QUEST procedure. The horizontal axis indicates stimulus strength, which in the present case would be the value of the blend weight. The vertical axis shows on the left, probability of a correct response, and on the right, trials. The points show each presentation, green for correct, red for incorrect. The blue histogram is the distribution of trials over strengths. The s-shaped curve that appears on every panel beyond the first is the best-fitting version of a Weibull psychometric function that relates probability to strength. Threshold is defined as the point at which this curve equals a probability of 0.82. The text describes this estimate as well as other parameters yielded by the fit. The gray shape in the background is the posterior probability density for the location of threshold. As the number of trials progresses, this shape narrows, indicating greater certainty regarding threshold.

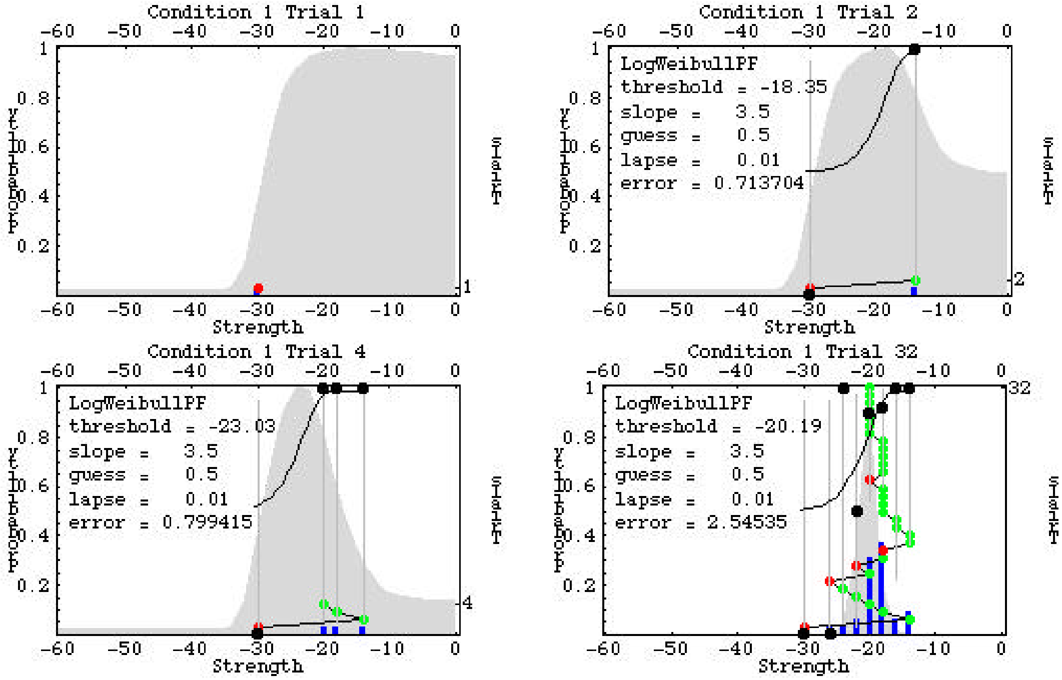

Figure 3. Illustration of the progress of the QUEST procedure at trials # 1, 2, 4, and 32.

3.7 Data Analysis

Beyond the fitting of Weibull functions to estimate thresholds, data analysis will consist of examination of variability of the data, estimation of visual quality scales, and comparison of data to the predictions of quality models. Some of these methods are described in greater detail in the pilot experiments described below.

4. Pilot Experiments

To illustrate and validate the methods and analyses, as well as possible outcomes of the proposed experiments, we have conducted two brief pilot experiments that will be described here. Two observers took part in these experiments: CVR and LCK. Both are research associates participating in this research program.

4.1 Pilot Experiment 1: Thresholds for Individual SRC/HRC Conditions

Each SRC/HRC condition may be regarded as the sum of the original SRC video and an error video. By making use of blends, as described above, it is possible to measure the fraction of the error video that is just detectable. Since many video quality models are based on threshold measurements, these data may provide an interesting comparison to predictions of quality models.

In a threshold experiment, on each trial two videos are presented, separated by 1 second. One of the two is a source video, the other is a blend. For example, one might be video 2.0, and the other 2.10. In Figure 4 I show the data of observer CVR for blends of videos 2.0 and 2.10. As noted above, videos are identified by the format SRC.HRC, with the original source video considered HRC=0. We used the standard 21 blends from –100 to 0 cL in steps of 5 cL. The units of strength shown on the abscissa are steps along this 21 point scale, and thus map to cL by the rule cL = -105 + 5 strength. It is a simple matter to transform the estimated threshold of 14.3, in units of strength, to cL where it equals 33.5 cL, which in turn corresponds to a weight of 0.462.

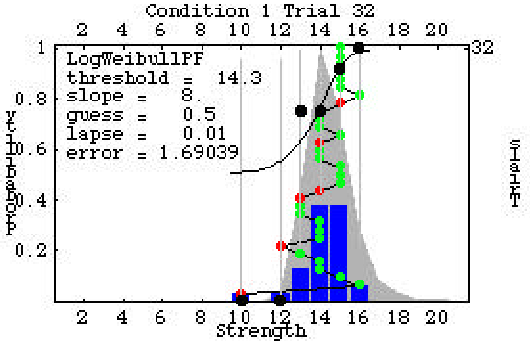

Figure 4. QUEST data for blends from condition 2-10.

Using blends and the QUEST procedure, we have measured thresholds for two source videos: SRCs 2 and 6, and nine HRCs (8-16). These are all of the 625/50 Hz HRCs. The results are shown in Figure 5. It should be emphasized that these thresholds result from only 32 trials, and are thus subject to some variability. Nevertheless, despite some variations, the two observers agree reasonably well. This provides some confidence in the methods and in the stability of these thresholds. The results also show considerable variation in threshold with SRC and HRC, as expected. SRCs and HRCs that yield large impairment scores, in general, yield lower thresholds. If the artifact is highly visible, we will need less of it to reach threshold.

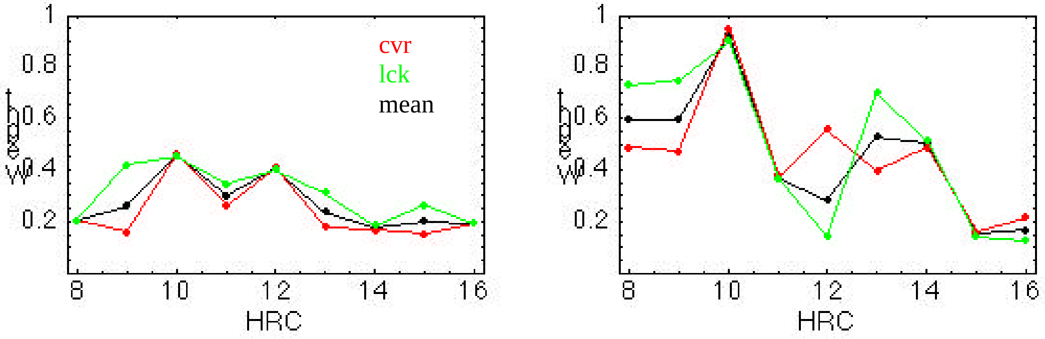

Figure 5. Detection thresholds for SRCs 2 and 6.

The thresholds for video 6.12 warrant comment. There is a large discrepancy between the two observers. This is due to the fact that this HRC was a "transmission impairment," an artifact that appeared briefly at one location in the video. Once observer LCK noticed this artifact, it was easily seen, and a low threshold resulted. This raises the difficult issue of how to deal with "non-stationary" artifacts which may be highly localized in space and time. More generally, it raises the issue of how cognitive factors (such as knowledge of the artifact location) should be manipulated in these experiments. However, we do not anticipate using transmission errors in these experiments, so these issues may be somewhat less pressing.

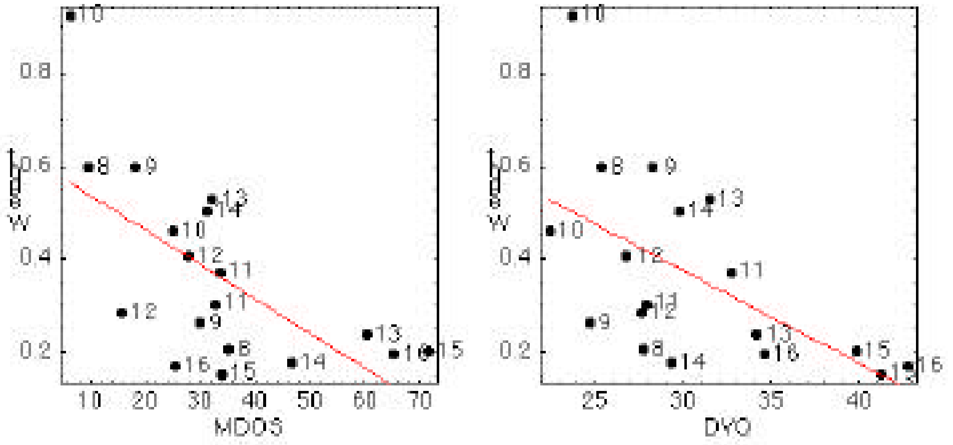

Figure 6. Threshold weight versus MDOS or DVQ.

In Figure 6 we examine more closely the issue of the relation between thresholds and other measures of impairment or quality. The two panels show the relation between threshold weight and mean differential opinion score (MDOS) or DVQ prediction. The figures combine the results for the two SRCs (2 and 6), and show the mean thresholds for the two observers. The figures confirm the general relation suggested above: the lower the threshold, the greater the measured or predicted impairment. The relationship is not perfect, however; the Spearman rank correlations for the two figures are -0.673891 and -0.630547. It should be pointed out, however, that we lack one piece of information that is needed to directly and quantitatively relate the thresholds to the measures for the full HRC, namely, knowledge of how the perceived artifact grows as the blend weight increases. Figure 6 shows clear evidence of a saturation in this function, since data are generally concave uowards, and tend to flatten out at high impairments and low values of *w*. The precise nature of this relation between *w* and perceived impairment is the subject of the next phase of the investigation.

4.2 Pilot Experiment 2: Estimation of Quality Scale

While the measurement of thresholds, as described in the previous section, tells us how much of a given artifact is detectable, it does not tell us how "intense" the full measure of the artifact will be. Consider the example of hearing. We may measure the threshold for each of a set of tones of different frequencies, but this will not tell us how loud each tone will be when all are set to a common sound pressure level. Nor does it tell us the relative loudness of two sound pressures of the same tone. To know these two things, we must measure the growth of loudness as sound pressure increases. Likewise, we must measure the growth of perceived impairment as the magnitude of the artifact is increased.

In the introduction, the general logic of Thurstonian scaling was introduced, and methods for measuring perceptual scales using pair-comparison were described. Here we show a concrete application of those methods.

4.2.1 Threshold vs reference weight

In this experiment we measure the impairment scale by means of a method we call "concatenated thresholds." In essence, we measure the first threshold, then use that threshold weight as the reference from which to measure a second threshold, and so on.

In the previous experiment, each 2AFC trial contained a reference and a test video. The reference video was always the original source, which by definition has a blend weight of 0. In this experiment, we again measure detection thresholds for impairment blends, but in this case the reference video may have a weight greater than 0.

In Figure 7 I plot the results of this experiment for SRC 6. The first threshold (with the original source as a reference) is at about 0.15 for both observers. Using approximately this value as a new reference, the second threshold is at about 0.21 for one observer, about 0.32 for the other. Using the approximate mean as the new reference, we then measure the third threshold, and so on. The complete set of thresholds (t) are plotted versus the corresponding reference weight (w). Figure 7 also plots the difference between the threshold weight and the reference (t-w) as a function of w. Considering the small number of trials per threshold (32), there is good agreement between observers.

These data do not go all the way to a reference weight of 1. In future experiments, we will measure a final threshold by using a reference weight of 1 and test weights of less than 1.

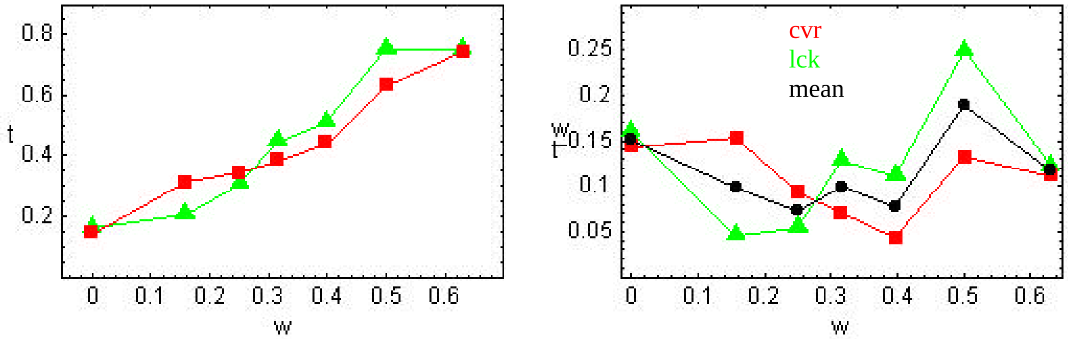

Figure 7. Threshold weight (t) versus reference weight (w), and weight increment (t-w) versus weight (w).

4.2.2 Direct estimation of quality scale

The data in Figure 7 allow us to make a direct estimate of the subjective impairment scale. The first step in this process is to construct an interpolation function *t*(*w*), based on the data in , that returns a threshold weight for any given reference weight. The purpose of this function is simply to span the gaps between the specific points at which we have measured *t* as a function of *w*. An example function, for data of observer CVR, is shown in Figure 8A. We used second-order interpolation, but the method is not very sensitive to the order.

We now want to construct a function ψ that returns a value of *d'* (the unit of the perceptual impairment scale, also known as a JND) for a given *w*. We can obtain samples of this function as follows. When *w*=0, by definition the perceptual scale is zero (ψ(0)=0), so the first sample of the function is {0,0}. We know that the function has grown to ψ=1 when w=t(0), so the second point on the scale function is {t(0), 1}. Now if t(0) were used as a reference weight, the interpolating function tells us that the next threshold (ψ=2) would be reached when *w*=t(t(0)), so the third point is at {t(t(0)), 3}. Thus the complete set of w values that yield increasing integer values of ψ are {t(0),t(t(0)), t(t(t(0))),…}. The function ψ obtained by joining up these sample values is shown in Figure 8B.

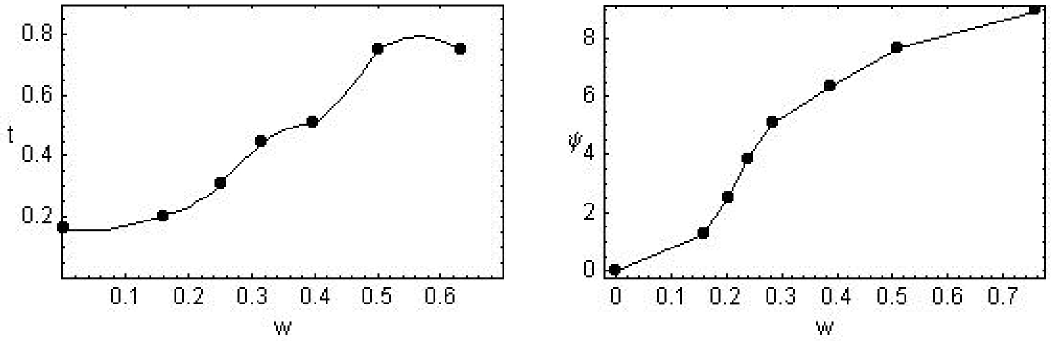

Figure 8. Direct construction of impairment scale.

4.2.3 Estimation of quality scale via function fitting

An alternative method for estimation of the perceptual scale is the use of curve fitting. In essence, we assume a particular form for the scale function, and then we estimate its parameters by means of maximum likelihood estimation. We will illustrate this method here, again using the data of observer CVR.

First we discuss the method in general terms. Consider a scale function of the form ψ(*w*, ρ), where *w* is the blend weight and ρ is a list of parameters. The data consist of a set of k trials, each yielding a record of the form $\{r_j, t_j, d_j\}$, where *j* is a trial index, *r* and t are the weights of reference and test, and d_j is a trial outcome, 1 or 0, depending on whether the observer was correct or not. From the Thurstone model, we know that the probability of a correct decision is

$$P(r,t,\rho) = C\left(\frac{\psi(t,\rho)-\psi(r,\rho)}{\sqrt{2}}\right) \qquad (\ 2$$

where C is the cumulative distribution of a standard Normal density. The likelihood of a complete data set is thus given by

$$L = \prod_{j|d_j=1}^{J} P(r_j,t_j,\rho) \prod_{j|d_j=0}^{J} 1-P(r_j,t_j,\rho) \qquad (\ 3$$

Using standard optimization techniques, we can then find the parameters that maximize the likelihood function L. In practice, it is often easier to optimize the log of L, which yields the same set of parameters.

Figure 9 shows the result of this method applied to the data of CVR. In this example, the function fit to the data was of the form

$$\psi(w) = \text{maximum Max}(0,\ w\ \text{ - threshold})^{\text{power}} \qquad (\ 4$$

with parameters {maximum, power, threshold} = {11.61, 0.5037, 0.1398}. For comparison, I reproduce the data from Figure 8B, the direct estimates of ψ. It is clear that the two methods give gratifyingly close results.

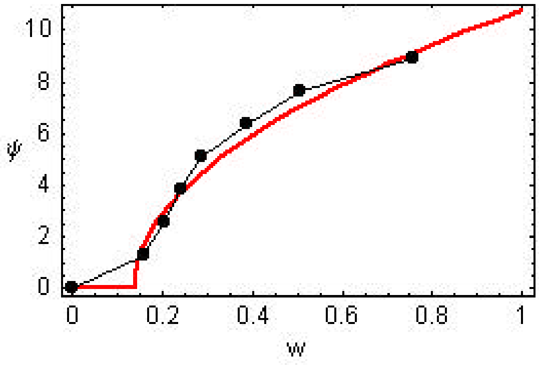

Figure 9. Quality scales estimated directly (black) and through curve fitting (red).

Under the assumptions of this analysis, Figure 9 is a picture of the growth of the subjective artifact, in units of d' (or JND, or standard deviation). Although we did not measure it here (see note above), this function can be estimated all the way up to *w*=1, that is, to the full HRC artifact. In this case, we would project a value of somewhere around 10 JND.

5. *Experiment 1*

5.1 General

Experiment 1 will use the methods of Pilot Experiment 1 to measure impairment scales for a subset of the 340 VQEG conditions. For each condition, we will use the method of concatenated thresholds to measure the complete scaling function for $0 < w < 1$.

5.2 Source materials

Each threshold based on 32 QUEST trials takes about 10 minutes to complete. The number of thresholds that must be measured for each HRC is approximately equal to the JND value for that HRC. In the pilot experiment, we found that video 6.15 measured about 10 JND, and its DMOS was about 34. Adopting the crude assumption of a linear relation between DMOS and JND, and noting that the mean DMOS over the entire VQEG data set was about 18.7, we estimate that the mean JND would be about 5.5. Thus to measure scale functions for the complete VQEG source materials would require
(20 SRC) * (16 HRC) * (5.5 thresholds) * (10 minutes) = 17,600 minutes = 293.3 hours. This is comparable to the approximately 23,000 minutes of observation time required by the VQEG experiments. However, we do not plan at this time to include all these conditions.

5.2.1 HRCs

We have selected five HRCs, as shown in Table 1. The HRCs are sorted in order of the mean DMOS score obtained in the VQEG experiments. We exclude HRCs 1, 11, and 12, (shown in gray) because they proved problematic in the VQEG experiments. The five selected (shown in yellow) are about evenly spaced in terms of mean DMOS, and exclude analog artifacts, which may be less important in future systems.

HRC	Mean DMOS	Mbps	CODEC	Details
2	5.79	19-19-12	422p@ml	3rd generation
7	5.82	6	mp@ml	
10	8.86	4.5	mp@ml	
3	11.06	50-50-…50	422p@ml	7th generation with shift / I frame
5	13.83	8 & 4.5	mp@ml	Two codecs concatenated
4	16.55	19-19/-12	422p@ml	PAL or NTSC 3 generations
6	17.85	8	mp@ml	Composite NTSC and/or PAL
8	19.38	4.5	mp@ml	Composite NTSC and/or PAL
12	20.25	4.5	mp@ml	Transmission errors
9	21.18	3	mp@ml	
1	23.23	n/a	n/a	Multi-generation Betacam
11	30.30	3	mp@ml	Transmission errors
14	33.35	2	mp@ml	Horizontal resolution reduction
16	34.12	1.5	H.263	CIF, Full Screen
13	37.74	2	sp@ml	
15	45.75	0.768	H.263	CIF, Full Screen

Table 1. Mean DMOS score for each HRC from the VQEG experiment.

5.2.2 SRCs

We propose to only examine 60 Hz conditions (SRC 13-22 as shown in Figure 2). Further, we propose to exclude SRC 20, as it was a still image. This yields a total of 9 SRCs.

5.3 Viewing Conditions

Viewing conditions will follow ITU Rec. 500 recommendations. We will use two viewing distances of 3H and 5H. Two viewing distances will be used because it is known that viewing distance has a large effect on perceived quality, and we would like to test the models' ability to predict this. The 5H distance is chosen to match that used in the VQEG experiments.

5.4 Total Observing Time

A summary of total observing time is given in Table 2. For the conditions selected, we anticipate approximately 75 hours of observing time. It should be noted that this time will be distributed over a number of observers. Each observer will be present for about 1.5 hours, and will complete only the thresholds corresponding to about 2 SRC-HRC conditions.

HRC	5
SRC	9
distances	2
thresholds	5
minutes/threshold	10
replications	1
total minutes	4500
total hours	75

Table 2. Total Observing Time

5.5 Observers

Observers will be non-experts with normal or corrected-to-normal acuity and normal color vision. Each observer will complete two SRC-HRC conditions at a single viewing distance. The pair of conditions will be chosen so that the anticipated number of thresholds to be measured adds up to about 10. Two observers, serving as research associates on this project, will complete all 75 hours of observation. Their results will provide a useful comparison to the data from the other observers.

5.6 Experimental Design

Since each observer will be available for only about 1.5 hours, each will complete only a small fraction of the complete experiment. As noted above, 1.5 hours will suffice to collect about 7 thresholds. A general strategy will be to select two different SRCs, and two different HRCs, whose expected number of thresholds adds up to 7. These expectations will be based on the MDOS scores form the VQEG experiment, as described above.

A consequence of this design is that no condition will be repeated on a single observer (except for the Research Associates data) and there will be no way of estimating effects due to individual differences, suc as acuity or sensitivity. This could be remedied by increasing the number of iterations to 2. To compensate, the number of SRCs might be reduced from 9 to 5.

5.7 Psychophysical Procedures

QUEST will be used to measure thresholds, and 32 trials will be used for each threshold. The method of concatenated thresholds will be used. Other methods will be as described in the pilot experiments.

5.8 Comparison With DMOS Data

The previous VQEG research project(VQEG, 2000) obtained differential mean opinion scores (DMOS) for each of the 320 distinct SRC-HRC conditions. The most straightforward analysis would be to correlate the JND values for each measured HRC to the DMOS scores obtained by VQEG. This would provide a potential "calibration" of DMOS scores in terms of JND. Another interesting analysis would be to compare the relative variability of JND and DMOS scores.

6. Experiment 2

6.1 General

In Experiment 1 we made use of blends to estimate the scaling function for each HRC, and as the end point of that scale, the JND value for the full (w=1) HRC. In Experiment 2 we attempt arrive at these numbers with a method that does not require the use of blends. Here we again make use of the logic of Thurstonian scaling. We note that stimuli which do not arise from an obvious one-dimensional physical intensity scale (such as the weight in a blend) nevertheless, in the Thurstone scheme, give rise to sensory magnitudes that are ordered along the sensory scale. Thus we can use pair comparisons among a complete set of HRCs for a given SRC to derive the sensory scale.

6.2 Source materials

Here we will use the same 5 HRCs and 9 SRCs as in experiment 1. If pilot experiments suggest that more HRCs are needed, they will be drawn from the other VQEG HRCs.

6.3 Viewing Conditions

Viewing conditions will be identical to those for Experiment 1. Viewing distances of 3H and 5H will again be used.

6.4 Observers

This experiment, because it will require a large number of trials from a single observer, will be conducted first using the authors as observers. If the method proves successful, additional observers will be used.

6.5 Psychophysical Procedures

Each trial will again be a two-alternative forced-choice presentation of two videos: a.x and a.y. A block of trials will again be 32 trials. Simulations will be conducted to determine how many blocks will be required to establish the scale for each SRC. Because comparisons between stimuli far apart on the sensory scale are not informative, we will make use of a recently developed method that adaptively selects the pairs to be presented based on prior results(Silverstein & Farrell, 1998).

6.6 Data Analysis

The maximum likelihood method described above will be used to derive the scale value for each HRC from the pair comparison data. These scale values, in units of JND, will be compared to the comparable numbers derived from Experiment1, as well as to the DMOS scores from the VQEG study. Finally, they will be compared to the predictions of the DVQ model.

The data from this experiment also provide an internal test of the assumptions that underlie the Thurstone model. In essence, the additivity of JNDs can be tested. If HRCs a and b are one JND apart, and b and c are one JND apart, then a and c should be two JNDs apart. Whether this is so will be manifest in the value of the error term in the maximum likelihood fit.

7. Efficient Adaptive Estimation of Sensory Scales

In experiment 1 we propose to use the method of concatenated thresholds (MCT) to measure impairment scales. The pilot experiment has demonstrated the utility of this method. However, it has several drawbacks. One is that it typically requires that the thresholds be measured in a particular sequence, from a reference of w=0, to references of progressively greater w. This may bias the results in unknown ways, for example, if the observer gets better (or worse) as data collection progresses. A second possible drawback is that it may be efficient. If the goals is to estimate the underlying scale function, the placement of trials implied by MCT may not be optimal.

As a separate research project, we propose to develop an optimal method for scale estimation. The general idea is as follows. We have described previously how the Thurstone model predicts the probability a correct response to a particular pair of blends (Equation 1). If a particular parametric form is assumed for the scale function, a likelihood function can be constructed for the data collected, and the parameters can be optimized so as to maximize the likelihood. After some data have been collected, we can consider how presentation of any possible pair will influence the parameter estimates. This influence can be expressed as a narrowing of the posterior density for each parameter, or as a gain in information (reduction in entropy). The best pair will be that which has the highest expected information gain. Although the mathematics may be opaque, the principle is simple. A pair that are too far apart will provide little information, because the observer will always get the right answer. A pair that are too close will be uninformative, because the observer will perform at chance. Intuitively, there is an optimal separation. Likewise the location (midpoint) of the pair will have an impact on the information gained, since if all the trials were previously at one end of the scale, more information will be gained by testing the other end.

The method we propose will be based upon a mathematical analysis of the likelihood function and of information gain in the case of multiple parameters. We have made some initial progress in this problem. We are also aware of other efforts to address this problem, which we will study further (Jesteadt, 1980; Levitt, 1992; Kiessling, Schubert & Archut, 1996; Keidser, Seymour, Dillon, Grant & Byrne, 1999).

8. Subjective Laboratories

The following laboratories may have the facilities and the willingness to participate in these studies.

8.1 Sarnoff Research Center

Contact: Jeffrey Lubin.

8.2 Tektronix, Inc.

Contact: Ann-Marie Rohaly.

8.3 National Telecommunications and Information Administration (NTIA)

Contact: Arthur Webster webster@its.bldrdoc.gov
Contact: Stephen Wolf
url: http://www.its.bldrdoc.gov/n3/video/Default.htm

8.4 Communications Research Centre of Canada

Contact: Phillip Corriveau
url: http://www.crc.ca

9. References

Brunnstroem, K., Eriksson, R. & Ahumada, A. J. (2000). Spatio-temporal discrimination model predicting IR target detection Proceedings, Human Vision and Electronic Imaging IV, 3644, pp. 403-410.

ITU-R. (1998). Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures International Telecommunications Union BT.500-8.

Jesteadt, W. (1980). An adaptive procedure for subjective judgments. *Percept Psychophys 28*(1), 85-8.

Keidser, G., Seymour, J., Dillon, H., Grant, F. & Byrne, D. (1999). An efficient, adaptive method of measuring loudness growth functions. *Scand Audiol 28*(1), 3-14.

Kiessling, J., Schubert, M. & Archut, A. (1996). Adaptive fitting of hearing instruments by category loudness scaling (ScalAdapt). *Scand Audiol 25*(3), 153-60.

Levitt, H. (1992). Adaptive procedures for hearing aid prescription and other audiologic applications. *J Am Acad Audiol 3*(2), 119-31.

Nakasu, E., Aoki, K., Yajima, R., Kanatsugu, Y. & Kubota, K. (1996). A Statistical Analysis of MPEG-2 Picture Quality for Television Broadcasting. *SMPTE Technical Conference 105*(11), 702-711.

Silverstein, D. A. & Farrell, J. E. (1998). Quantifying perceptual image quality Proceedings, Image processing quality and capture, Portland Oregon, The Society for Imaging Science and Technology, pp. 242-246.

Thurstone, L. L. (1959). *The Measurement of Values*. Chicago: University of Chicago Press.

VQEG. (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment.

Watson, A. B. & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys 33*(2), 113-20.

Watson, A. B. & Solomon, J. A. (1997). Psychophysica: Mathematica notebooks for psychophysical experiments. *Spatial Vision 10*(4), 447-466.