Agreement Among International Laboratory's Subjective Video Quality Assessments Using Several Rating Methods

Source: A.C.Morton, AT&T +1 732 949 2499 mailto:acmorton@att.com

1. Introduction

In 1997, four laboratories studied the effects that different sets of experimental stimuli have on subjective video quality assessments as measured by three different rating procedures[Context]. Two sets of test conditions contained either predominantly weak impairment levels, or predominantly strong impairment levels. Both sets included four common test conditions for analysis, to determine if the context in which the common stimuli occur would influence the subjective ratings. Viewers rated one set of test conditions; either the set with strong impairments and four common conditions, or the set with weak impairments and common conditions, using one of three rating scales [BT.500]:

1. DSCQS - Double Stimulus Continuous Quality Scale
2. DSIS - Double Stimulus Impairment Scale
3. Comparison - A bipolar continuous comparison scale constructed for this experiment

The study concluded that ratings of the common test conditions using the DSCQS method were the least susceptible to the context of the stimuli present. Ratings with other methods changed to some degree depending on the strong or weak context. Context independence is an important criterion for the VQEG (Video Quality Experts Group) experiments, since it would complicate the objective methods to take the experimental context into account.

Because this study included results from four laboratories in different countries and utilized different rating methods, it provides an opportunity to examine some additional questions currently facing the members of VQEG.

- What is the best subjective method, when the results of multi-national laboratories will be combined?
- What method permits the greatest discrimination between test conditions with similar impairment levels?
- What is the best method, when the results will be used to evaluate objective estimates of the video impairments?

A method that produces highly consistent ratings, among both viewers and laboratories, while minimizing contextual effects, is the answer to VQEG's needs.

2. Analysis and Discussion

We first examine the consistency (or agreement) among the ratings collected at different laboratories. As in the original analysis [Context] we use the four common test conditions in the strong and weak sets as a basis for comparisons, permitting a further examination of the effects of experimental context. However, the VQEG analysis will use the ratings of individual stimuli (scene-condition combinations), rather than aggregating all scenes over each test condition, so we use the scene-condition ratings here.

VQEG analysis plans include several different statistical metrics to test the consistency of objective and subjective measurements [Obj_Plan]. We will use two metrics in this analysis, the Pearson correlation coefficient and (unweighted) RMS error.

Table 1 summarizes the findings for Lab-to-Lab comparisons with each rating method and experiment (strong or weak impairment sets). We give the range for each metric over the six comparisons between the four labs. To aid comparison, we report RMS Errors as a percentage of the full rating scale.

Table 1 Summary of Agreement among Labs

	Strong Impairments		Weak Impairments	
Method	Correlation	RMS Error, %	Correlation	RMS Error, %
DSCQS	0.975 - 0.937	7.7 - 12.0	0.993 - 0.960	4.3 - 18.1
DSIS	0.995 - 0.985	5.0 - 10.2	0.993 - 0.969	4.8 - 10.2
Comparison	0.989 - 0.961	6.5 - 16.4	0.990 - 0.978	6.9 - 10.8

Although DSIS results correlate best among methods with the strong impairment set, the RMS error ranges have considerable overlap. Furthermore, DSIS has no advantage with the weak impairment set. It appears that all of these methods can produce results with reasonably good agreement across labs. Tables giving each lab-to-lab comparison and figures illustrating two independent pairs are found in the annex.

Note that analysis of only the common test conditions may be optimistic from the perspective of agreement, since two of the four conditions are near the end of range where subjective ratings tend to have smaller variance (than for mid-range conditions).

In the author's opinion, Lab-to-Lab Agreements represent an important benchmark in the evaluation of objective measurement methods. It would be possible to make this benchmark a more formal part of VQEG's evaluation process, by adding a few details in the Subjective and Objective Test Plans. It is proposed that VQEG discuss and adopt this or a similar benchmark.

3. Conclusions

This contribution examines additional criteria for VQEG's selection of a subjective test method, beyond the contexual effects studied in [Context]. We find that all methods foster reasonable agreement among multi-national laboratories, when the criteria for comparisons are some of VQEG's metrics for evaluating objective measures.

This contribution proposes that VQEG adopt lab-to-lab comparisons as a benchmark in the evaluation of objective measurement methods. Further work would be necessary to examine the relative discrimination power of these rating methods.

References

[BT.500] ITU-R BT.500-7, "Methodology for the Subjective Assessment of the Quality of Television Pictures," 1997.

[Context] "Investigation of Contextual Effects," ITU-R WP 11E Contribution, Canada, France, Germany, Switzerland, April 4, 1997.

[Obj_Plan] "Evaluation of New Methods for Objective Testing of Video Quality: Objective Test Plan," M. Ravel, Editor, February 1998.

Annex - Details of the Comparisons

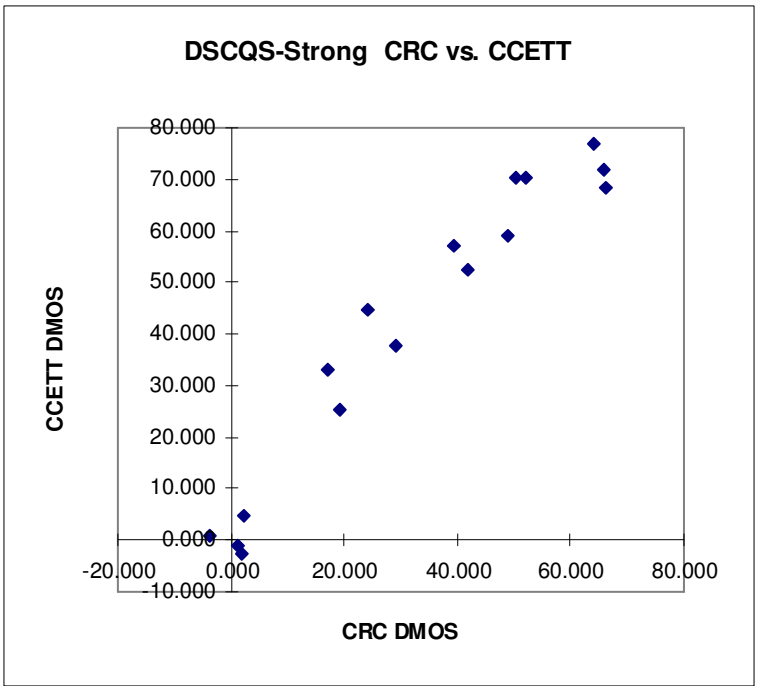
DSCQS Method, Strong Impairment Set

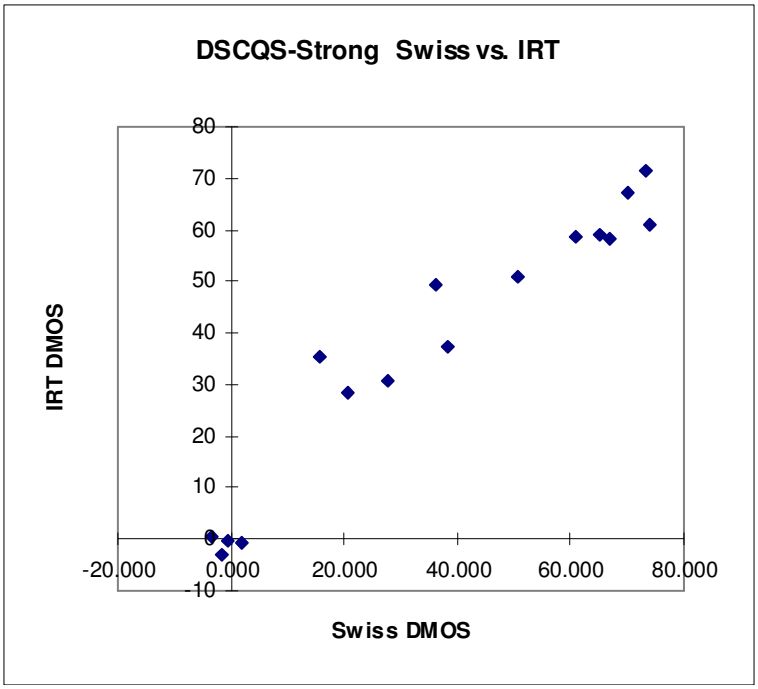
Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.9692	1		
Swiss	0.974588	0.97335	1	
IRT	0.93734	0.972932	0.964362	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	11.95867	0		
Swiss	8.81012	7.90927	0	
IRT	10.28854	7.742678	7.751723	0





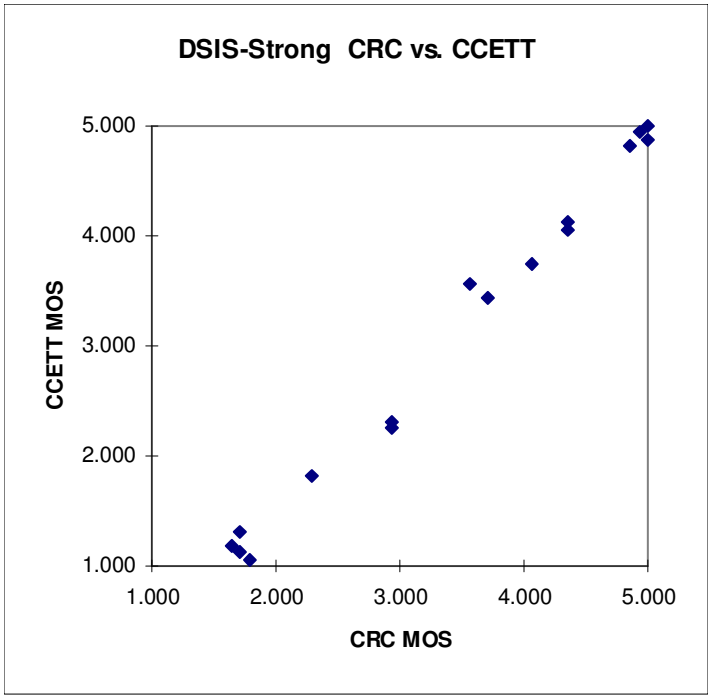
DSIS Method, Strong Impairment Set

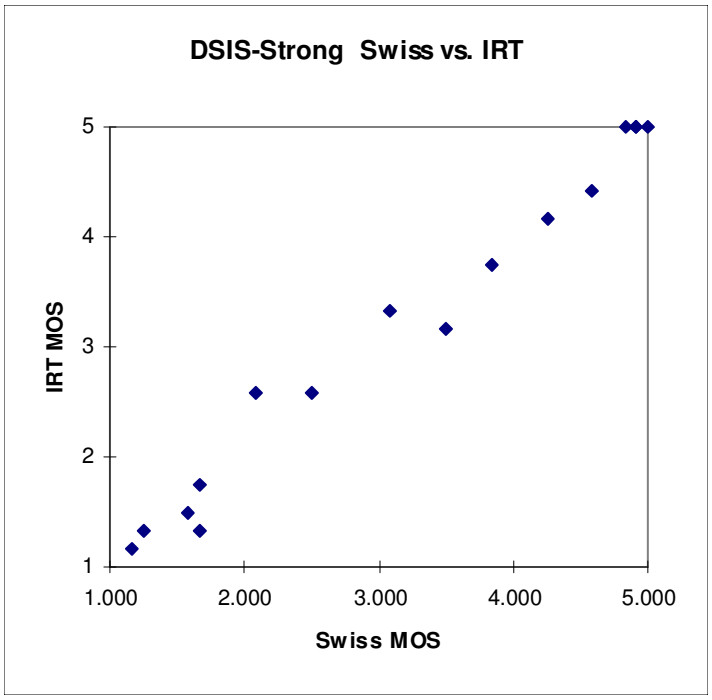
Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.995036	1		
Swiss	0.986991	0.984941	1	
IRT	0.995323	0.991114	0.990122	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	0.406311	0		
Swiss	0.371965	0.262181	0	
IRT	0.321092	0.21381	0.200789	0





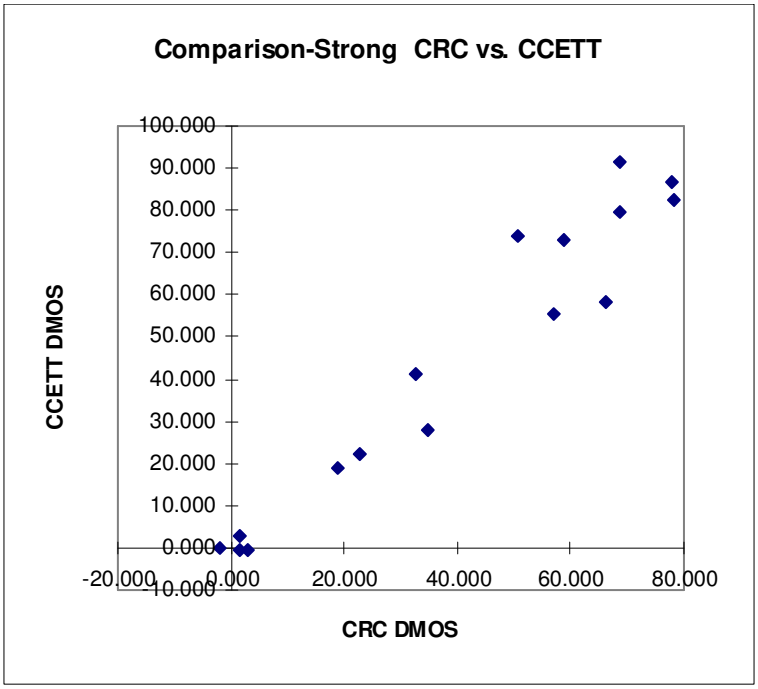
Comparison Method, Strong Impairment Set

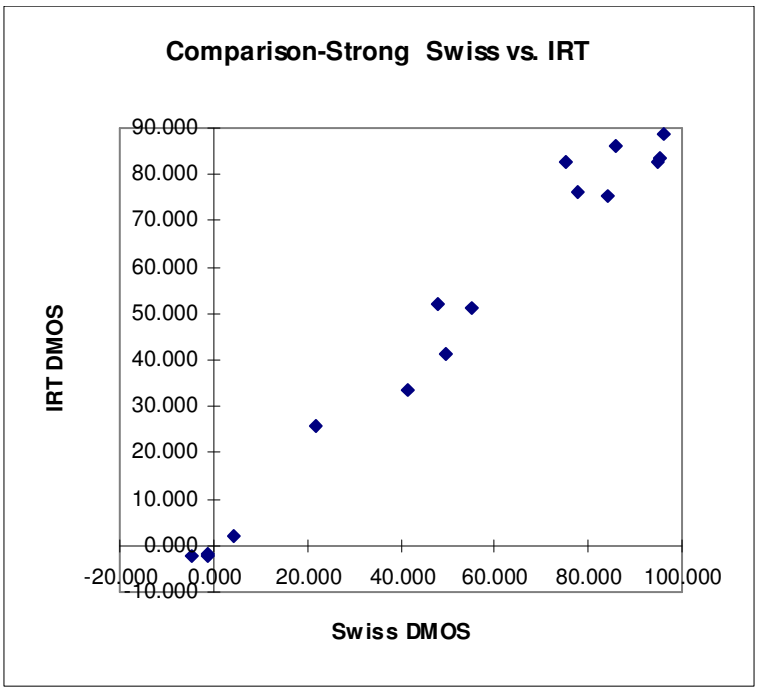
Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.969142	1		
Swiss	0.964912	0.961316	1	
IRT	0.970782	0.971334	0.988586	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	10.25422	0		
Swiss	16.38789	12.2705	0	
IRT	12.78747	9.042465	6.474986	0





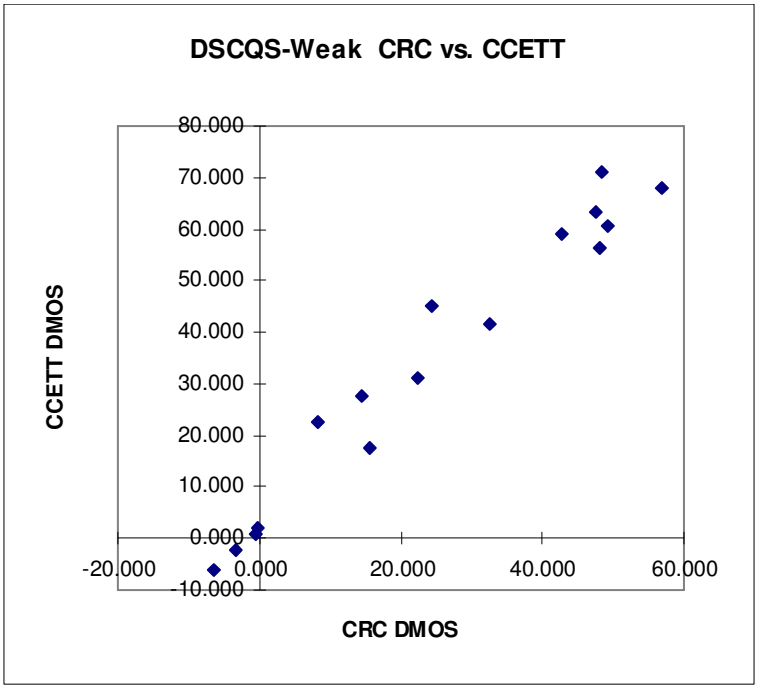
DSCQS Method, Weak Impairment Set

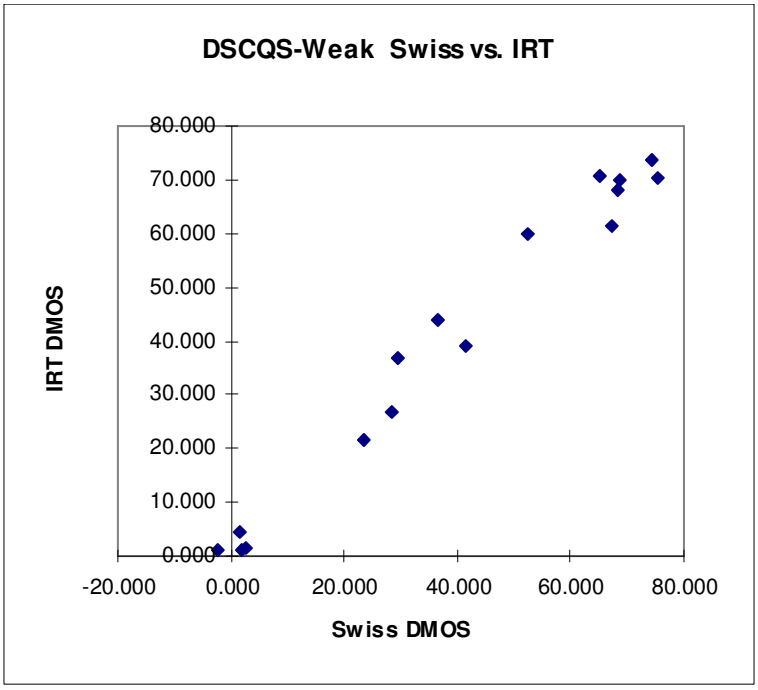
Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.979301	1		
Swiss	0.974295	0.992951	1	
IRT	0.959921	0.983576	0.988548	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	12.05743	0		
Swiss	16.94892	5.939421	0	
IRT	18.0975	7.597364	4.297394	0





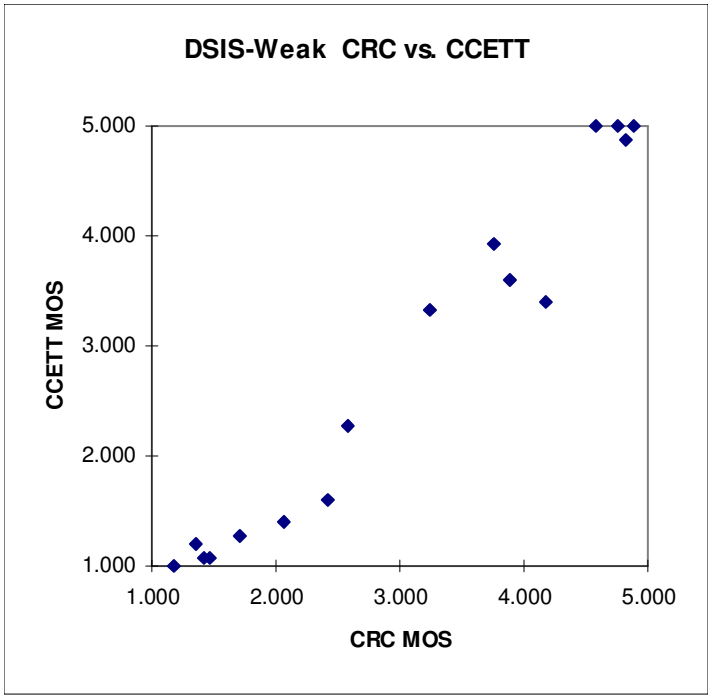
DSIS Method, Weak Impairment Set

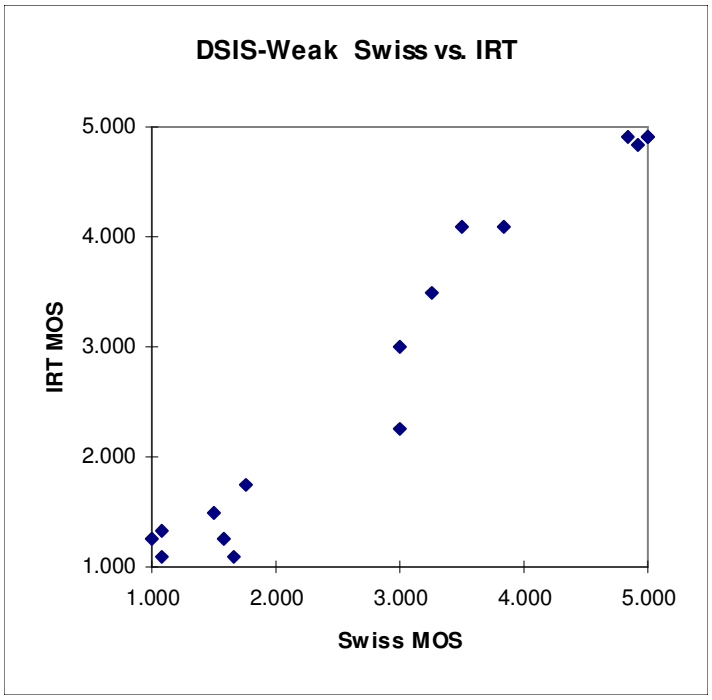
Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.981015	1		
Swiss	0.969228	0.987839	1	
IRT	0.982201	0.992901	0.97801	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	0.40983	0		
Swiss	0.397235	0.261604	0	
IRT	0.359197	0.193961	0.319369	0





Comparison Method, Weak Impairment Set

Correlation Coefficient, r

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	1			
CCETT	0.98529	1		
Swiss	0.977987	0.987061	1	
IRT	0.978413	0.981124	0.989761	1

RMS Error

	CRC	*CCETT*	*Swiss*	*IRT*
CRC	0			
CCETT	7.676937	0		
Swiss	10.77693	6.859154	0	
IRT	7.784559	8.985078	10.68943	0