

# **An Evaluation of Quality Assurance Tests for JANET Videoconferencing Services and Recommendations for Improvements to the Testing Programme**

## **Management Summary**

This memorandum reports the findings of a retrospective evaluation of Quality Assurance tests for the JANET Videoconferencing Switching Service over a period of approximately twelve months. A total of 147 test results were analysed for completeness of data, general validity of the test procedure and potential confounds arising due to differences between individual raters.

During the survey period, a large number of cases recorded incomplete data in respect of independent or dependent variables, or both. During the course of the study period, the test procedure was modified to include items of independent variables that had not originally been recorded, and this accounts for the majority of missing data values. In other cases, dependent variables (measured or rated items) have been omitted for a number of reasons, most notably including the premature abandonment of tests in which a FAIL outcome resulted.

Despite the large number of cases in which data were incomplete, no major threat to the validity of the testing procedure has been revealed. A chi-squared test for rater bias in the test outcome did not produce a statistically significant result. The results of the subjective rating tests for audio and video quality were also analysed for a rater effect using a one-way ANOVA model. The result of this test on the audio rating was not statistically significant, however a significant ( $p < 0.01$ ) result was found in the case of the video rating.

A number of recommendations are made for improvements to the test procedure. These recommendations are intended to: -

improve the completeness and accuracy of test results (Recommendations 1 and 2);

ensure that test conditions more closely reflect those used in actual conference operation (Recommendation 3);

improve the rating scales used for subjective tests (Recommendations 4 and 5);

promote greater consistency between raters (Recommendations 6 and 7);

ensure that test results are recorded in a consistent and meaningful form (Recommendations 8 to 10).

## **Recommendations**

### ***Recommendation 1***

*Test sessions should continue until it is impossible to proceed further.*

### ***Recommendation 2***

*The audio encoding used during the test session shall be recorded as one of the following: G.711, G.728, G.722 (for H.320 systems).*

### ***Recommendation 3***

*System tests should be conducted at the maximum bit rate supported both by the measurement site and the system under test.*

#### **Recommendation 4**

The binary items for **echo** and **doubletalk** should be changed to use the following five point Likert scales: -

##### **Echo**

- 5 Imperceptible
- 4 Perceptible, but not annoying
- 3 Slightly annoying
- 2 Annoying
- 1 Very annoying

##### **Doubletalk**

- 5 No impairment
- 4 Perceptible impairment, but not annoying
- 3 Slightly annoying
- 2 Annoying
- 1 Very annoying

#### **Recommendation 5**

Likert rating scales for subjective quality of audio and video should be changed to use the five point scale specified in ITU-R BT.500, viz: -

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

#### **Recommendation 6**

Measures should be taken to ensure a high degree of rater consistency in the use of subjective rating scales. Instructions to raters should be revised accordingly. In rating, no allowance should be made for system configuration or bit-rate when rating scales are completed.

#### **Recommendation 7**

PASS/FAIL criteria should be explicitly and deterministically defined, and these criteria should be made available to site managers.

#### **Recommendation 8**

The reason(s) for failure should be tabulated.

#### **Recommendation 9**

It is recommended that a forms-based interface be used to ensure a standard format in the recording of test results.

#### **Recommendation 10**

It is recommended that the reason for test be recorded: i.e. whether initial acceptance, routine re-test or re-test after reported problems.

# An Evaluation of Quality Assurance Tests for JANET Videoconferencing Services and Recommendations for Improvements to the Testing Programme

## Background

Quality assurance testing for sites participating in JANET Videoconferencing Services was first introduced during operation of the pilot Super JANET Video Network. When the present Janet Videoconference Switching Service (JVCSS) replaced the pilot service in 1997, the continued operation of a QA testing programme was one of the Mandatory Requirements for the procurement of a Management Centre to operate the service. With the subsequent introduction of the broadband Scottish MANs Videconferencing Network, the testing programme was extended to include the sites connected to this new service.

A successful QA test must be completed before a site is registered to use the JANET Videoconferencing Services, and thereafter at regular intervals. Additionally, further QA tests are conducted where a complaint has been received in regard to a site's performance in a scheduled conference. Sites may also request additional QA tests, for example when changes are made to a site's equipment or conference room environment.

The present study was initiated to evaluate the QA testing programme and to make recommendations for the improvement of the QA test procedure.

## Survey sample

The survey included all QA tests for the JVCSS performed by the Management Centre, over the period of March 1998 to March 1999. A total of 147 tests were surveyed, of which 113 (90.48%) resulted in a PASS outcome, and 14 (9.52%) resulted in a FAIL outcome.

## Completeness of data

Complete data were recorded in only 31 cases. There were a number of reasons for missing data, most notably the format of the test form was amended during the evaluation period: the original form did not record the audio coding or connection bit rate and did not require the rater's identity to be recorded. However, in 41 cases one or more of the measured or rated values were missing. The incidence of missing data for independent and dependent variables is summarised in the following contingency table: -

Rater, encoding and bit rate present.	One or more measured or rated values missing		Row total
	YES	NO	
NO	21	75	96
YES	20	31	51
Column total	41	106	147

Test outcome	Measured or rated values missing (N=41)
PASS	34 (30%)
FAIL	7 (50%)

There is evidence, both implicit and from explicit comments recorded on certain test forms, that a number of FAIL tests were abandoned prematurely. Whilst there may be cases where it is not possible to complete a test, premature abandonment of a test may result in further problems remaining undetected until a retest is performed.

## **Recommendation 1**

*Test sessions should continue until it is impossible to proceed further.*

## **Reasons for failure**

Of the 14 tests resulting in a FAIL outcome, in 12 cases mic level was too low, in one case failure was attributed to intermittent distracting noise (from nearby lift plant) and the remaining case suffered from gross impairments of video quality.

## **Effect of rater on test outcome**

The tests were further categorised by rater; the number of pass and fail outcomes by rater are given in the following table. In 41 cases, the identity of the rater was not recorded.

<i>Rater</i>	<i>N</i>	<i>Pass</i>	<i>Fail</i>
RP	56	49	7
PL	18	17	1
WP	32	30	2
Anon	41	37	4

For the non-anonymous cases, the chi-squared statistic was used to test for a rater effect on outcome, but this was not statistically significant. (Chi-squared = 2.572; N.S. for DF = 2,1).

## **Audio coding**

The audio codings used in tests were recorded as follows:

G.711	3
G.722	49
Unknown	95

It is noted that in the present sample only.

o7(e pr)mat(lt is)-have be3( th)-1.ne prnclt iso7(e pr)d1.7(o1.)-1.2o3

## Mute level

The mean value of mute level was -48.39 dBm (N=113). In 47 cases (41.6%) a mute level less than -48 dBm was recorded and in 65 cases (57.5%), the mute level was equal to, or greater than, -48 dBm. The modal value was -48 dBm (37 cases, 32.7%). The distribution is given in the following table:

dBm	< -53	-53 -51	-50 -48	-47 -45	-44 -42	-41 -39
	2	27	56	20	7	1

The high incidence of mute level being recorded as -48dBm is attributable to the SNR of an 8 bit encoding system. However, it is indicated that this is not consistent with the number of test cases specifically recorded as G.711 operation, and there is evidence to suggest that the audio format has not been correctly recorded in a number of cases.

## Mic level

Since some installations comprise multiple microphones, the analysis was restricted to the level of the first microphone. Mic level is the most common reason for failures. There is a single instance of inconsistency where a level of -4dBm was recorded in a test resulting in a PASS outcome. In other cases, sites have been failed at -4dBm or -2dBm. The distribution is given in the following table:

dBm	-10	-8	-6	-4	-2	0	1	2	3	4
PASS				1		37	9	44	3	39
FAIL	1	1	4	4	1			1		
All cases	1	1	4	5	1	37	9	45	3	39

## Ambient noise level

The mean value of ambient noise level was -44.63 dBm (N=144). The distribution is summarised below.

dBm	-56	-53	-50	-47	-44	-41	-38	-35	-32
	-53	-51	-48	-45	-42	-39	-36	-33	-30
	1	10	38	40	31	9	6	4	5

It appears that no site has been failed the test on the sole basis of this measurement. Levels as high as -30dBm have been recorded without the test being failed. Whether a limit on ambient noise should be prescribed is a matter for further study. It should be noted that this static measure is sensitive only to continuous noise. In the only case failed for background noise, the measured level was a respectable -46dBm and the noise in question was due to the operation of adjacent lift plant, and consequently of an intermittent nature.

## Echo

Echo is a common problem in teleconferencing audio, and a common reason for re-test requests. However, only 8 tests resulted in a positive result being recorded for echo. The use of the present binary variable is therefore of no use in determining the test outcome. A more suitable approach would involve the use of a 5 point disruption rating scale for echo and would permit extremely poor performance to be identified.

## Double talk

In this case, no instance of doubletalk being unsatisfactory has been recorded and the measure in its present form is of limited value. This item could usefully be amended to use a Likert type scale.

### **Recommendation 4**

*The binary items for echo and doubletalk should be changed to use the following five point Likert scales: -*

#### Echo

- 5 Imperceptible
- 4 Perceptible, but not annoying
- 3 Slightly annoying
- 2 Annoying
- 1 Very annoying

#### Doubletalk

- 5 No impairment
- 4 Perceptible impairment, but not annoying
- 3 Slightly annoying
- 2 Annoying
- 1 Very annoying

## Audio Rating

This item is seen to suffer from a number of problems: some test results contain notes suggesting that raters make allowances for system configuration (e.g. desktop) when determining ratings; in other cases, raters have changed the semantic labels (e.g. changing good to fair) before completing a rating; furthermore it is likely that many 4 or 5 ratings are largely based on objective measures already performed. The data suggest the possibility of a rater effect, however this was not statistically significant at the 5% level.

### **Analysis of Variance: Audio Rating by Rater**

<i>Source of variation</i>	<i>SS</i>	<i>DF</i>	<i>MS</i>	<i>F</i>	
Between raters	4.02	2	2.01	2.61	(N.S.)
Within raters	76.73	100	0.77		
Total	80.75	102			

One problem lies in the semantic labels used, in particular that the central anchor (good) is not neutral. A suitable five point rating scale is specified in ITU-R Recommendation BT.500: this is reproduced below.

## Video rating

The observations made for audio rating apply equally to video rating. In addition, a test for rater effect produced a significant result ( $p < 0.01$ ).

### Analysis of Variance: Video Rating by Rater

Source of variation	SS	DF	MS	F
Between raters	3.87	2	1.93	6.43 (p<0.01)
Within raters	29.66	100	0.30	
Total	33.53	102		

The video rating also includes an incidence of impairment scale. The data for this scale have not been analysed.

### **Recommendation 5**

*Likert rating scales for subjective quality of audio and video should be changed to use the five point scale specified in ITU-R BT.500, viz: -*

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

### **Recommendation 6**

*Measures should be taken to ensure a high degree of rater consistency in the use of subjective rating scales. Instructions to raters should be revised accordingly. In rating, no allowance should be made for system configuration or bit-rate when rating scales are completed.*

## **Other recommendations**

The current test procedure suffers from the lack of explicit criteria for the PASS and FAIL outcomes. It is believed that there would be advantages in the criteria being published and made known to sites.

### **Recommendation 7**

*PASS/FAIL criteria should be explicitly and deterministically defined, and these criteria should be made available to site managers.*

It also recommended that where a site fails a test, the reason(s) for failure should be explicitly recorded. At present the reason for failure may only be determined implicitly from the test record.

### **Recommendation 8**

*The reason(s) for failure should be tabulated.*

A major problem encountered in this study has been the inconsistency of test records as a result of the open format of the items on the test form. This problem may be avoided by use of a forms-based interface for the collection of data.

### **Recommendation 9**

*It is recommended that a forms-based interface be used to ensure a standard format in the recording of test results.*

The present test records do not distinguish between the various reasons for a test being conducted. It is felt that this information is of statistical value when reviewing the work performed by the Management Centre.

### **Recommendation 10**

*It is recommended that the reason for test be recorded: i.e. whether initial acceptance, routine re-test or re-test after reported problems.*

## **QA tests for the Scottish MANs Videoconferencing Network**

A modified form of the QA test procedure is employed on the Scottish MANs Videoconferencing Network (SMVCN). SMVCN tests were not included in the present study. However, most of the recommendations made in the present study apply also to the test procedure used for the SMVCN.

## **Extension of QA testing to cover operation of H.323 systems**

The present test may be applied to H.323 operation with few changes. Modifications to the items for audio coding and connection bit rate will be necessary and there will be a requirement to include additional measures to determine network QoS performance.

## **Conclusions**

Whilst a number of recommendations have been made for improvements to the testing programme, in general no major problems, i.e. those which constitute a threat to the validity of the programme, with the current test procedure have been revealed. The recommendations made are intended to: improve the completeness and accuracy of test results (Recommendations 1 and 2), ensure that test conditions more closely reflect those used in actual conference operation (Recommendation 3), improve the rating scales used for subjective tests (Recommendations 4 and 5), promote greater consistency between raters (Recommendations 6 and 7) and ensure that test results are recorded in a consistent and meaningful form (Recommendations 8 to 10).

It should be noted that the validity of the testing programme is dependent also upon the reference systems and instrumentation used at the management centre and the consistency of environmental conditions in effect during testing sessions. In particular, the type and picture line-up of video monitors, viewing distance and ambient lighting; the calibration of instrumentation used for *measurements*; and the level settings of microphones and sound monitoring, room acoustics and prevailing ambient noise level in effect during subjective assessment of audio performance. However, consideration of these factors was not included in the terms of reference for the present study and these issues have not been investigated.