Document No: T1Y1.1/90-539 Rev. 1
July 16, 1990
DRAFT

COMMITTEE T1Y1.1
CONTRIBUTION

****

STANDARDS PROJECT: Digital Encoding of System M•NTSC Television Signals for Broadcast Quality Transmission at the DS3 Rate

****

TITLE: Statistical Analysis of Results of Test Program for Selecting Codec Standard for Broadcast Quality NTSC Television at DS3

****

SOURCE: Edwin L. Crow, Mathematical Statistician, ITS.N3
Institute for Telecommunication Sciences
National Telecommunications & Information Administration
Boulder, CO 80303

****

DATE: July 16, 1990

****

This document describes the statistical analysis for testing the significance of the differences among subjective scores of four EPAs (codecs) recorded by expert viewers in the test program developed and carried out by the T1Y1.1 Experts Group for broadcast quality video transmission at the DS3 rate. The results of the analysis are summarized. The basic computations were carried out by Michael Wagner. The results were analyzed in consultation with him, Ross J. Owens, and Richard Quinn.

Any suggestions or comments on this document can be addressed to:

Ralph C. Brainard
AT&T Bell Labs
Room 4C534
Holmdel, NJ 07733
201-949-4147
FAX 201-949-3697

and

Edwin L. Crow
ITS.N3
U.S. Dept. of Commerce
325 Broadway
Boulder, CO 80303
303-497-3452
FAX 303-497-5993

STATISTICAL ANALYSIS OF RESULTS OF TEST PROGRAM FOR SELECTING CODEC STANDARD FOR BROADCAST QUALITY NTSC TELEVISION AT DS3

1. INTRODUCTION

The test program for selecting a codec (EPA, short for "Embodiment of Proposed Algorithm") standard was described in detail in Document T1Y1.1/90-502, dated January 19, 1990 (and perhaps later revisions). It has been carried out with 26 to 28 expert viewers at two laboratories, and the data have been processed by Michael Wagner at the University of Maryland BSOS Computer Laboratory. The statistical analysis to determine the significance of the differences among EPAs, especially of the apparently best EPA from the others, was presented in Appendix 7 of the above document. The present document summarizes in Section 2 the results of the analysis that was made and describes in Section 3 that analysis and why it differs to some extent from that in Appendix 7. Section 4 gives conclusions and a recommendation.

2. SUMMARY OF STATISTICAL RESULTS

The test program was carried out as planned except that for economic reasons the same viewers were not used entirely for all of the four different quality sessions. The number of viewers varied only from 26 to 28 in the four sessions, 11 or 12 at a distance near the screen, 15 or 16 at a greater distance. There were 49 individual viewer scores for each EPA (2 more, for error bursts, were also recorded but discarded by the Committee before the present analysis for reasons unknown to this writer) and of somewhat fewer corresponding reference scenes. These scores are on a scale of 0 to 50, and the analysis was concerned only with the 49 differences between each EPA score and its paired reference score. As a result there were 4x49x27 = 5292 difference scores in all, enough to provide quite a precise mean score for each EPA. Since there is negligible difference between the scores at the two viewing distances, those scores are pooled for the summary in Table 1. The first column of standardized scores for each quality category is linearly related to the means of the viewer scores but the scale is changed to make a perfect picture 100 (0 on the original scale) and a completely bad picture 0 (50 on the original scale); the direction is reversed, so that the best EPA has the highest score, not the lowest. The order within each quality category is from best to worst. The slightly curved lines bracket the EPAs that are not statistically significantly different from one another Thus the only EPA significantly different from the other three in any category is C, in Basic Quality and Multipass. In error susceptibility, the pair A and C were significantly better than B and D.

TABLE 1
STANDARDIZED MEAN VIEWER SCORES OF FOUR EPAS

Basic Quality			Post Processing			Error Susceptibility			Multipass			Weighted Combination		
D	99.0	99.2	D	98.8	99.1	A	98.8	98.2	B	91.3	87.4	A	94.2	92.2
B	98.8	98.2	C	98.7	98.8	C	97.6	96.4	D	86.5	80.8	B	93.9	91.7
A	98.1	97.5	B	97.8	97.5	B	69.6	60.9	A	85.4	79.4	D	92.4	90.2
C	91.9	89.6	A	97.7	98.1	D	65.8	58.8	C	71.4	64.1	C	87.8	84.6

The significance was actually tested by transforming all the original difference scores, say x, to log(x+28), the term 28 being added to make all of the individual transformed scores positive. The second column of standardized scores for each quality category is a linear function of the mean log (x+28), with the scale reversed and changed to make a perfect picture score 100 and a completely bad picture score 0 (as in the first column). It is seen that these standardized log scores are in the same order as the original scores except for a negligible interchange of B and A in Post Processing.

The first column under Weighted Combination is obtained by applying the weights specified by the Committee (18/49 for BQ, 12/49 for PP, 4/49 for ES, and 15/49 for Multipass) to the standardized original mean scores in the previous columns. The second column under Weighted Combination is obtained by applying the same weights to the standardized mean log transformed scores in the previous columns.

The mean viewer scores for each sequence are plotted in Figures 1-4. It is seen that EPA C is distinctly worse than the others in three sequences for Basic Quality and in two sequences for Multipass. Otherwise the behavior is very consistent, but the consistency includes poorer performance by B and D in Error Susceptibility. The expected steady degradation with further passes through the EPAs is confirmed.

3. DISCUSSION OF THE ANALYSIS

The analysis in Appendix 7 of Document T1Y1.1/90-502 was based on the assumption of independent normal distributions of the viewer sequence difference scores with the same (true) standard deviations. Inspection of the mean scores and their standard errors (standard deviations of individual viewer scores divided by the square root of the number of viewers, 26, 27, or 28) showed a large variation in both means and standard deviations. Wagner, Quinn, and Owens questioned whether an unweighted mean of sequence means would provide a good estimate of quality, and this is confirmed by statistical theory. The assumption of the same (true) standard deviations appeared untenable, and, although immediate information on the shape of the distribution of individual scores was not available, it seemed likely that the assumption of normality would be violated also.

To investigate the type of transformation of sequence scores that could result in normal homoscedastic (equal standard deviations) distributions, the viewer standard deviation for each sequence and EPA was plotted in Figure 5 against the corresponding mean. The standard deviation increases almost linearly with mean over most of the range aside from random variation that is consistent with the 95% confidence limits for a single standard deviation for a sample of 27 from a normal distribution (drawn at mean 7.5 on Figure 5). This suggests that a logarithmic transformation of the data might render both means and standard deviations more nearly equal, but still preserve the order of the means. In fact, differences of means are ultimately of interest, and differences of large means would not, after transformation, overwhelm differences of small means.

However, some of the means are slightly negative, and many individual difference scores could be considerably negative, so a constant had to be added before taking logarithms; adding 27 would make all scores positive, but for convenience 28 was added to make all the logarithms positive also.

The relative sizes of front row and back row means and standard deviations (of the original difference scores) were analyzed to see if there were systematic differences. The results are shown in Table 2. The only one of the eight relative frequencies differing significantly from 1/2 at the 2-sided 5% significance level is that for Post Processing means. The mean of the 24 back row means in this case is 0.35 while that of the 24 front row means is 1.56. The respective root-mean-square standard deviations are 3.24 and 4.82. While the mean difference in means, 1.21, is statistically significant with the large number of data, 24x28, it is believed justifiable to pool the front and back row data in this as well as the three other quality categories because the mean difference is small relative to the random variation. All results in this document have the data from both viewing distances pooled.

TABLE 2
RELATIVE FREQUENCIES WITH WHICH BACK ROW MEANS AND STANDARD DEVIATIONS EXCEED THOSE IN FRONT ROW

Quality	Means	Standard Deviations
Basic Quality	23/36	23/36
Post Processing	4/24	7/24
Error Susceptibility	10/16	10/16
Multipass	36/60	34/60
All	73/136	73/136

The confidence intervals presented in Document T1Y1.1/90-502 for testing the significance of differences in means,

$$\bar{x}_{1..} - \bar{x}_{2..} \pm qsn^{-1/2},$$

are a generalization of Student t confidence intervals for the differences of two means. They have to be generalized further here because some of the standard deviations of EPA scores being compared differ from each other, even after the log transformation. Thus the above formula becomes

$$\bar{x}_{1..} - \bar{x}_{2..} \pm q(s_1^2 + s_2^2)^{1/2} (2n)^{-1/2},$$

where s_1 and s_2 are the standard deviations of the n (26, 27, or 28) viewer scores composing $\bar{x}_{1..}$ and $\bar{x}_{2..}$ respectively (or $s_1 n^{-1/2}$ and $s_2 n^{-1/2}$ are the standard errors of $\bar{x}_{1..}$ and $\bar{x}_{2..}$). (The two means compared could have different n's in general, but they are the same here.) Furthermore the q in the latter formula is slightly larger than in the former, being based on 26 degrees of freedom (d.f.) rather than 104 (C.W. Dunnett, "Pairwise multiple comparisons in the unequal variance case," Journal of the American Statistical Association, 75 (Dec. 1980), 796-800).

The significance testing is shown here for the Basic Quality mean log scores (adjusted by dividing by the mean of the four means):

EPA	A	B	C	D	
Mean	0.996	0.994	1.020	0.991	
Std. Devn.	0.013	0.010	0.024	0.017	
Std. Error	0.0025	0.0019	0.0046	0.0033	(n=27)

Testing the difference between the apparently best EPA, D, and the next best, B, gives the ratio

$$\frac{0.994 - 0.991}{(0.0033^2 + 0.0019^2)^{1/2}} = 0.79$$

If this ratio is less than the appropriate $2^{-1/2}q$ (or, equivalently, the confidence interval is longer than the difference in means, 0.003 here), then the means do not differ significantly. The q depends on the confidence level specified (specified by the Committee as 0.95, equivalent to a two-sided significance level of 0.05), the number of means, 4, and the effective d.f., 26 in this case. Here $2^{-1/2}q$ is 2.743 (H.L. Harter, "Tables of range and studentized range", Annals of Mathematical Statistics, 31 (Dec. 1960), 1122-1147), so the two means do not differ significantly. The only EPA differing significantly in the mean from the others is C, which differs significantly from all the others.

Document T1Y1.1/90/90 541 ("Report on the Video Transmission Quality Evaluation" by Ross J. Evans and Richard Quinn, July 9, 1990) tabulates the mean raw scores and their standard errors. Significance tests can also be made with these based on large-sample normal theory, but they would not be as reliable as the above tests on the log transform. Still it is pleasing that they give exactly the same results, shown by the curved lines in Table 1.

Just as the standard error of each of the EPA mean quality category scores was found by calculating the corresponding score for each viewer, the standard error for each Weighted Combination was found by calculating the corresponding Weighted Combination for each "viewer." Twelve viewers did participate in all quality categories. Fourteen others in each quality category were made to correspond at the same respective viewing distances but otherwise at random from the 14 to 16 remaining viewers to calculate a standard deviation among 26 "viewers" and thence a standard error of each mean by dividing by the square root of 26. These standard errors are somewhat but not effectively larger than those calculated from the formula for the variance of a linear combination using the quality category standard errors.

4. CONCLUSIONS AND RECOMMENDATION

The testing was carried out successfully, very close to the test plan. The statistical analysis to test significance was modified slightly by transforming the test scores to logarithms to bring the data more closely to the statistical assumptions necessary for testing whether the nominally best EPA is significantly better than others. However, formal analysis of the original scores gave exactly the same significance results. The nominally best EPA (in the Weighted Combination of all four quality categories) is the same, EPA A, whether original or transformed scores are used, but is not significantly better than B or D. EPA C is significantly worse than the other three.

It is therefore recommended that on the basis of the specified subjective testing EPA A be named the nominally best EPA but that it be noted that it is not statistically significantly better than B or D in the Weighted Combination.. That is, there is no clear winner in the Weighted Combination. However, B and D are significantly poorer than A and C on Error Susceptibility, so that A is the only EPA not differing significantly from the best on any of the four categories. Hence A has a clear edge on all the others if one looks at all four individual quality categories.

6/20/90

Figure 1. Basic Quality

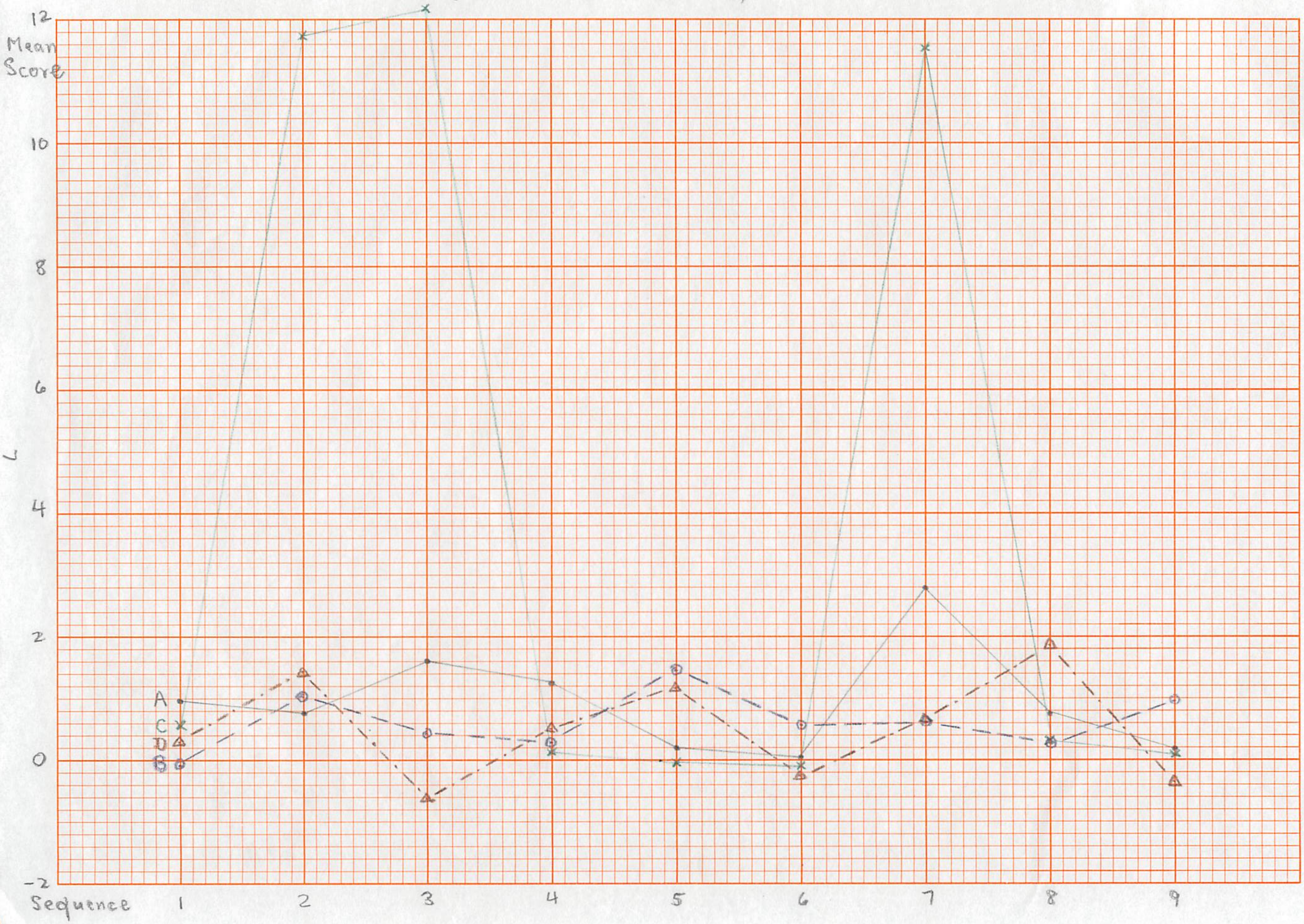
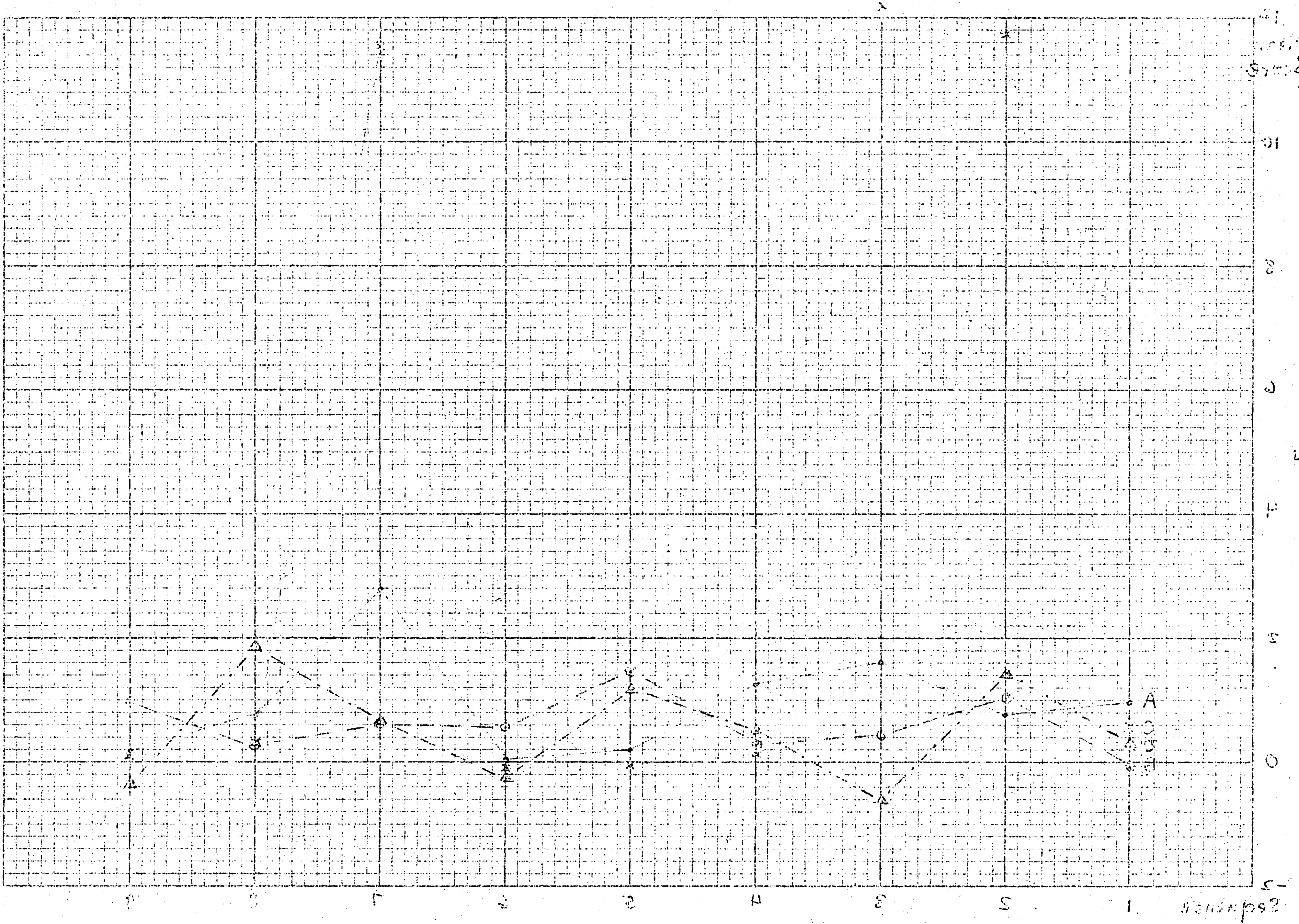
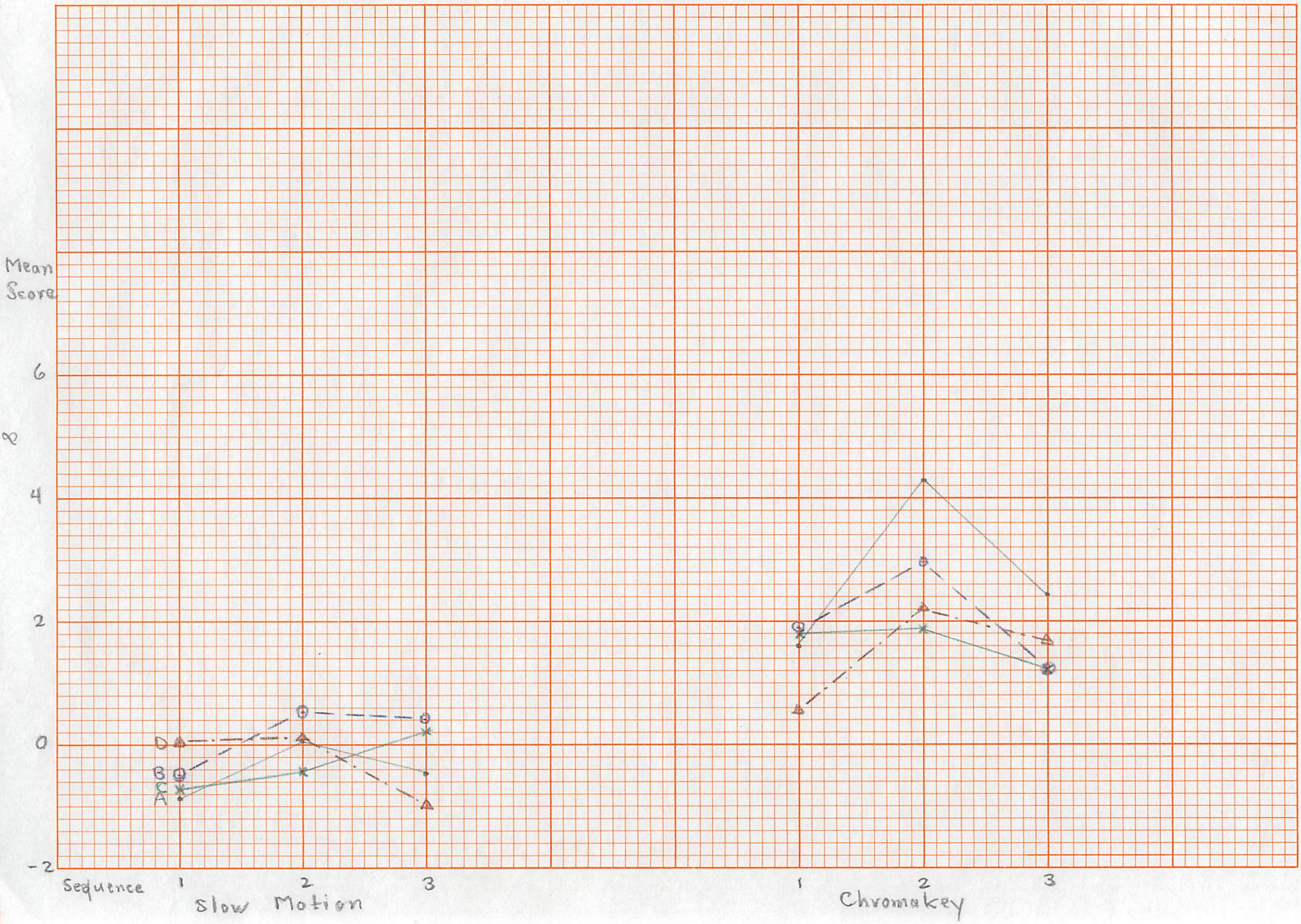


Figure 1. Basic Ability



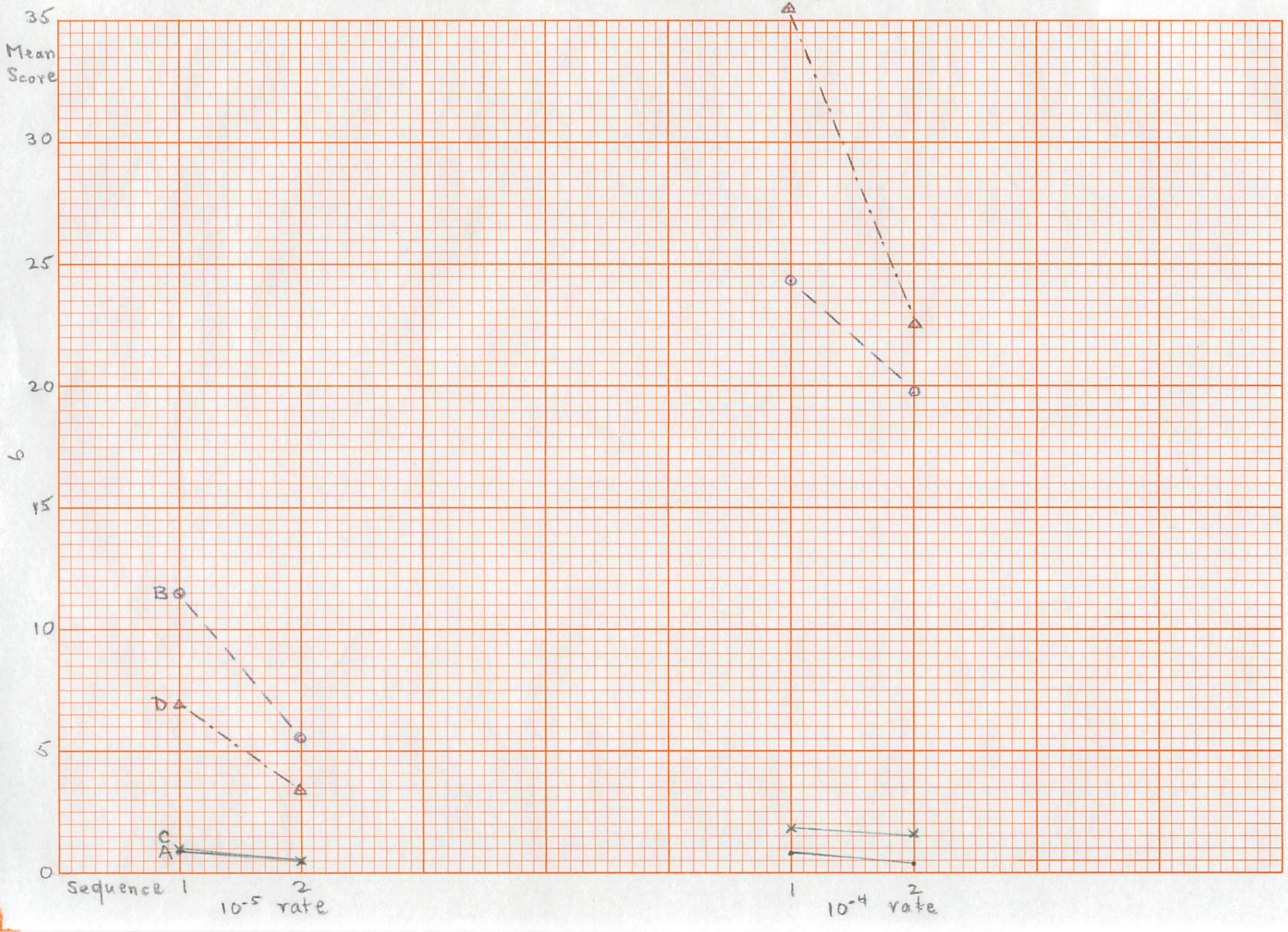
6/20/90

Figure 2. Post Processing



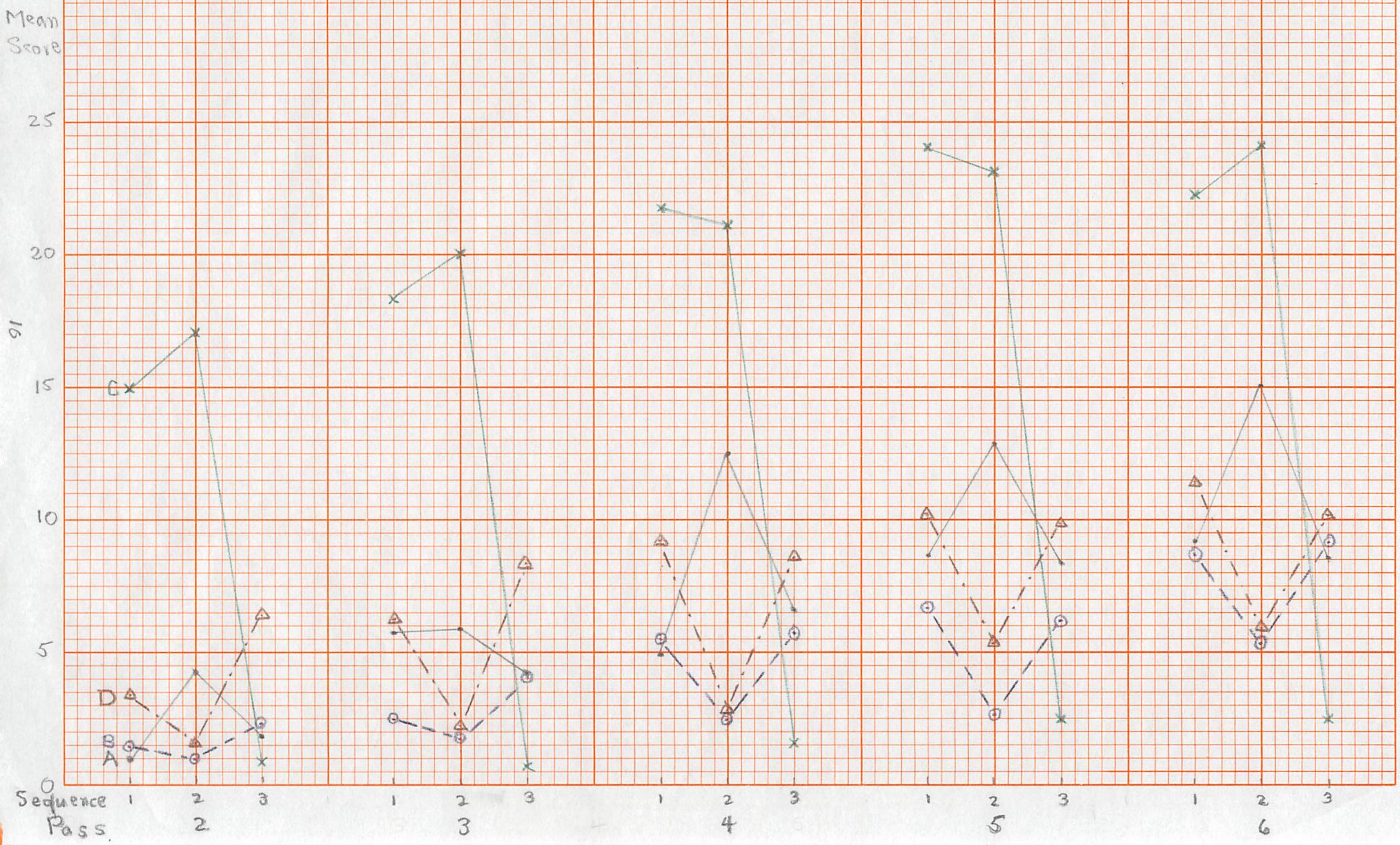
6/20/90

Figure 3. Error Susceptibility



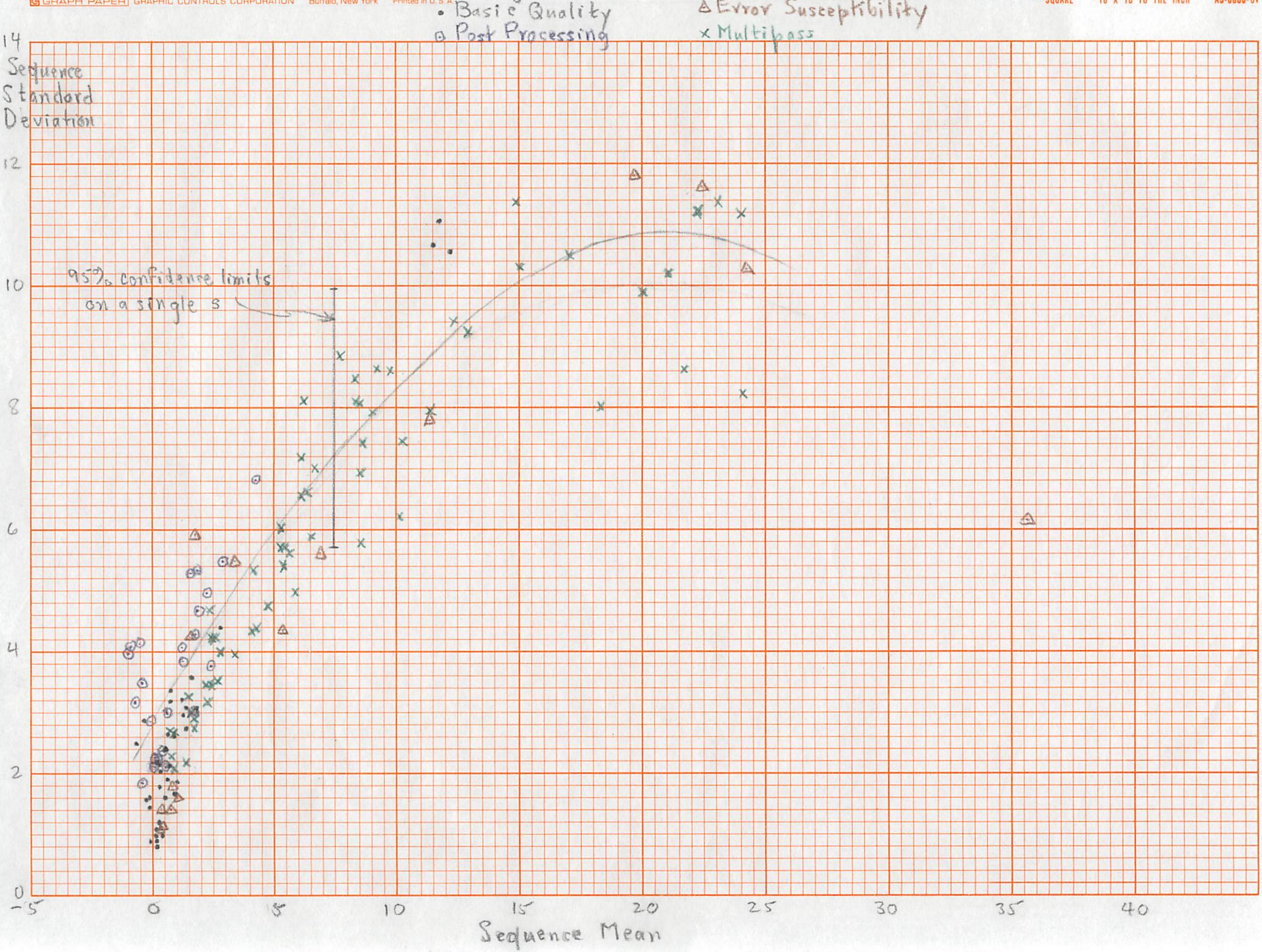
6/21/90

Figure 4. Multipass



6/25/90

Figure 5


• Basic Quality
⊙ Post Processing
△ Error Susceptibility
× Multipass
Sequence Standard Deviation
95% confidence limits on a single s
Sequence Mean