COMMITTEE T1
CONTRIBUTION

Document Number T1Q1.5/91-133

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Preliminary Analysis of Subjective Test Results

****

ISSUE ADDRESSED: Video Quality

****

SOURCE: NTIA/ITS - Arthur Webster

****

DATE: September 30, 1991

****

DISTRIBUTION TO: T1Q1.5

****

KEYWORDS: Video Teleconferencing, Video Telephony, Video Quality, Subjective Quality, Objective Quality

****

DISCLAIMER:

****

1. Introduction

As part of the effort to develop an effective objective video quality assessment system (i.e. one that measures video quality with test equipment)(see Ref. [1]), the Institute for Telecommunication Sciences (ITS) conducted a subjective viewing test utilizing a panel of human viewers to rate video quality. The results of this subjective test will be used to determine a set of objective measures which correlate well with video quality as perceived by humans (see Ref. [2]). We already have a large set of proposed objective measures and this subjective test will help us to select an optimal subset of measures and to disregard the rest. This contribution presents an overview and an analysis of the subjective test and its results.

2. Overview

The subjective performance assessment of a video system depends to a large degree on the input video scene. This is particularly true for digital systems which use data compression. Thus, many scenes are required to completely characterize the performance of the particular video system. ITS conducted this subjective test utilizing a large data set consisting of 36 test scenes and 29 different video systems. The test scenes span a wide range of user applications including still scenes, limited motion graphics, and full motion entertainment video. A large portion of the test scenes are typical VTC/VT scenes (see Ref. [3]).

The 29 video systems include the 'null' system (component video) and analog video systems such as S-VHS and VHS recording and playback, RF transmission with noise, attenuated chroma, and digital video systems including 12 video codecs from 8 manufacturers operating at bit rates from 56 Kbps to 45 Mbps and bit error rates from zero to 10^{-5}. All video systems except the 'null' system include NTSC encoding and decoding.

Each test session consisted of 38 or 40 30-second test clips. Each clip consisted of nine seconds of the unimpaired video scene, three seconds of grey, nine seconds of the impaired version of the video scene, and finally a nine second period in which to mark the response form. The viewers were instructed to decide on and mark the level of impairment in the second scene, using the first scene as a reference. The five possible responses offered were: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying. This scale covers a wide range of impairment levels in a nonlinear fashion and is specified as one of the standard scales in CCIR Recommendation 500-3.

The selection of the 158 clips used in the test (out of the 1044 clips available) was made both deterministically and randomly. Some deterministic selections were made so that the results can be compared with other subjective tests that have used some of the same test scenes (see Ref. [4]). Random selections were made from a distribution table that pairs more VTC/VT systems with typical VTC/VT scenes and more entertainment systems with entertainment video scenes. The impaired video scenes span a wide range

of quality levels. They can therefore be useful as a basis for validating technology independent (i.e. independent of coding algorithm and transport architecture) objective measures of video quality.

. This subjective test was conducted in accordance with CCIR Recommendation 500-3 (see References [2] and [5]).

The test was conducted during the summer of 1991. Forty-eight viewers were selected from the U.S. Department of Commerce Laboratories phone book here in Boulder, Colorado. Each viewer completed four viewing sessions during a single week -- one session per day. Each session lasted approximately twenty-five minutes. Viewers rated a total of 132 unique clips of video out of 158 clips actually viewed. Twelve clips were used for training and fourteen clips were used as consistency checks. The consistency checks were of two types: inter-session checks (two clips were shown in each of the four sessions) and intra-session checks (two clips per session were repeated in that session).

3. Analysis of Test

The five responses were mapped to the numbers 1 through 5 as follows: 5 = imperceptible, 4 = perceptible but not annoying, 3 = slightly annoying, 2 = annoying, and 1 = very annoying. Note that linearly spaced numbers have been assigned to a nonlinear scale. Thus, the averaged quality scores shown in the following figures represent one method of producing composite subjective quality scores from viewer histograms.

Figure 1 shows the distribution of average scores (solid line) and their standard deviations (dashed line) for all the clips rated. This plot shows that the selection of test clips did, in fact, cover the full range of quality which this test sought to measure.

Figure 2 shows the histogram of the set of all scores. This figure indicates that the test contained somewhat more clips in the higher quality range. The grand mean of the test is 3.21. If the histogram was exactly flat, the grand mean would be 3.0.

Figure 3 plots the average quality score for twelve different clips as the number of viewers increases from 1 to 48. This shows that the mean fluctuates very little after about 35 or 40 viewers and therefore the number of viewers we used (48) was adequate to insure a valid average score.

4. Analysis of Data

This subjective test was designed to produce data that would be most useful for designing a technology independent objective quality assessment system. This test was not designed to produce results that would allow one to directly compare the quality of different video systems (since only 3 or 4 different video systems out of the 29 were used for each of the 36 scenes). However, some broad statements can be made about the

trends we see in the test data.

Figure 4 shows the spread between the maximum and the minimum viewer-averaged scores for seven particular video systems (and the 'null' system -- i.e. the source video compared with itself). It is clear that, especially for low bit rate codecs, the perceived quality of the system under test is a function of the input signal. Figures 5 through 12 plot histograms of the viewer scores for the video systems shown in Figure 4. Note that the histograms for a fixed video system are functions of the input video test scene. This dependence of performance on user-application is more apparent at low bit rates than high bit rates (e.g. compare Figure 6 with Figure 10). When subjectively testing higher quality video systems, such as NTSC and 45 Mbps contribution quality codecs, one may need to design test scenes which specifically attempt to point out a particular (possibly known) weakness in that system.

5. Conclusion

The first full-scale subjective viewing test performed at ITS is complete. The test results indicate that enough viewers were used to obtain stable average scores and that the test evenly covered a wide range of quality levels. The data supports the idea that the subjective quality of a video system is strongly dependent on the test scene. The set of test clips we used for this test spans a wide range of quality levels for the video systems under test. Therefore, this data set will be appropriate to use in the design of a technology independent objective quality assessment system.

ITS objective measures of video quality will now be extracted from the clips used in the subjective test. Results on the correlation between the subjective and objective measures will be presented to T1Q1.5 at the next meeting.

6. References

[1] Institute for Telecommunication Sciences, "Objective Quality Assessment of Digitally Transmitted Video", Committee T1 contribution T1Q1.5/91-118.

[2] Institute for Telecommunication Sciences, "The Development and Correlation of Objective and Subjective Video Quality Measures", Committee T1 contribution T1Q1.5/91-119.

[3] Institute for Telecommunication Sciences, "Progress Report on Subjective and Objective Quality Assessment of VTC/VT Systems", Committee T1 contribution T1Q1.5/91-123.

[4] Institute for Telecommunication Sciences, "Correlation Between ITS Objective Measures and Subjective Video Quality: Preliminary Results on a Set of 15 Scenes", Committee T1 contribution T1Q1.5/91-124.

[5] Voran, Stephen, "The Development of Objective Video Quality Measures That Emulate Human Perception", Globecom 91 Proceedings.

Figure 1. Subjective Viewing Test: NTIA1 - Distribution of Scores

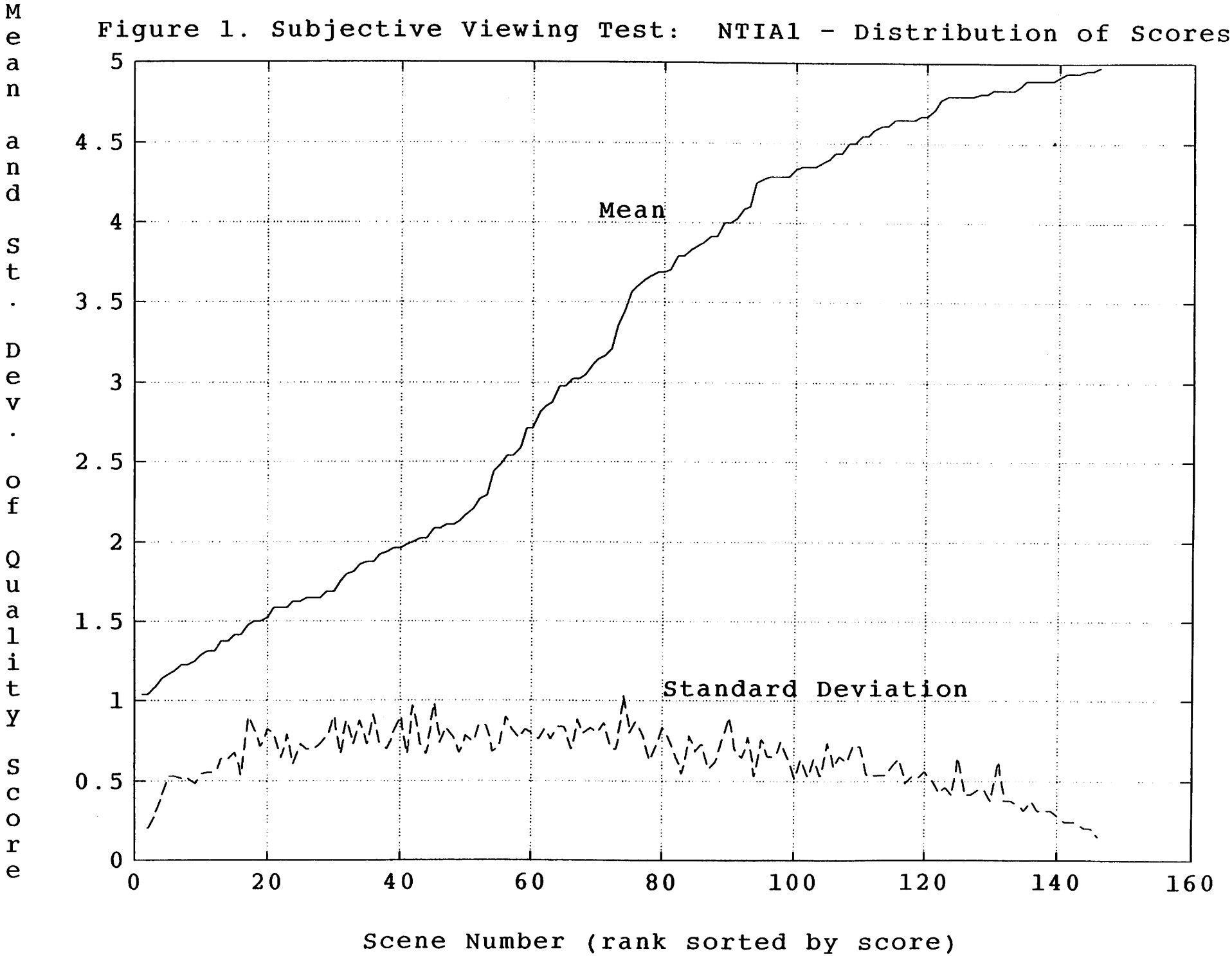

Figure 2. Histogram of 48 Viewers - All 146 Clips

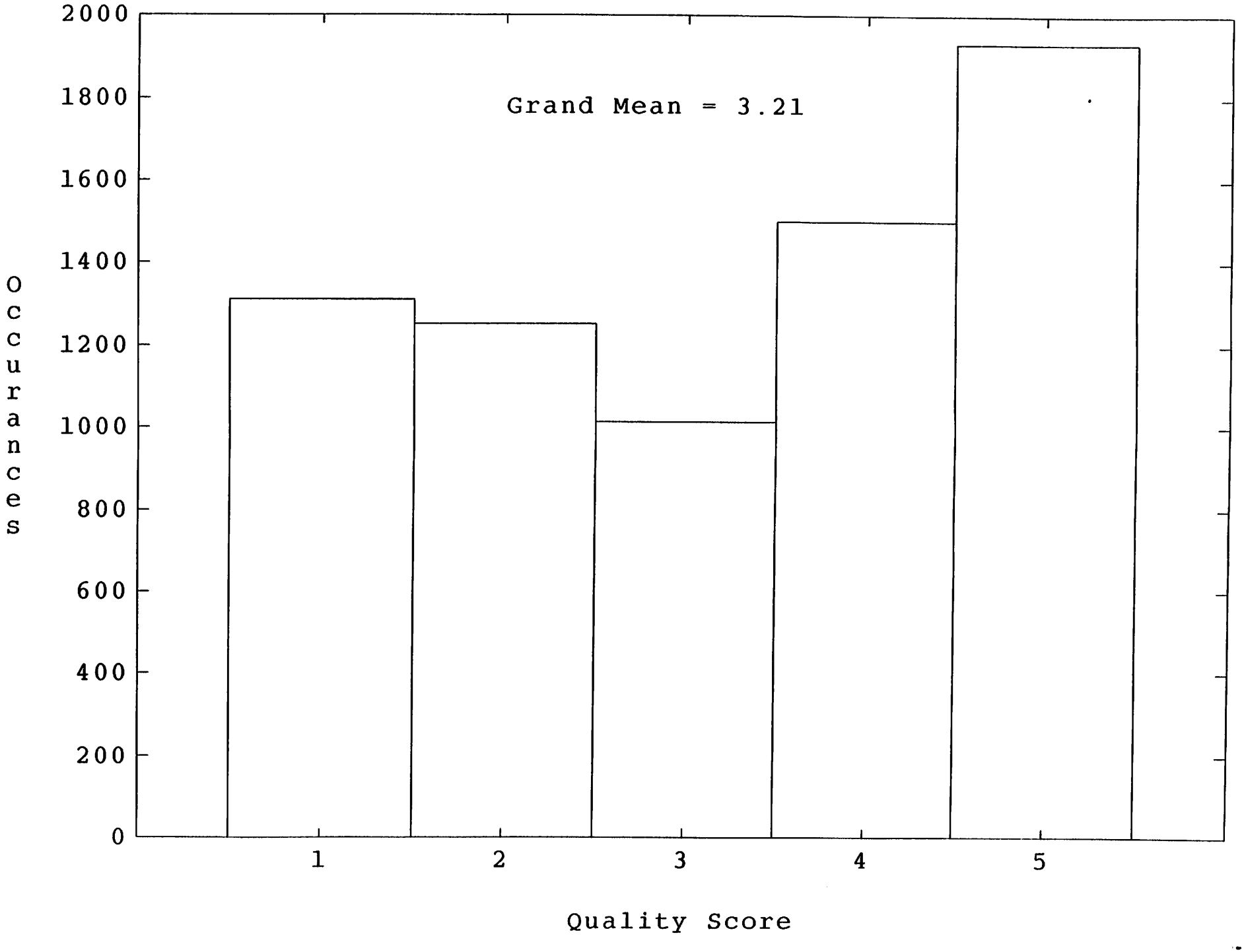

Figure 3. Average Quality Scores as Number of Viewers Increases

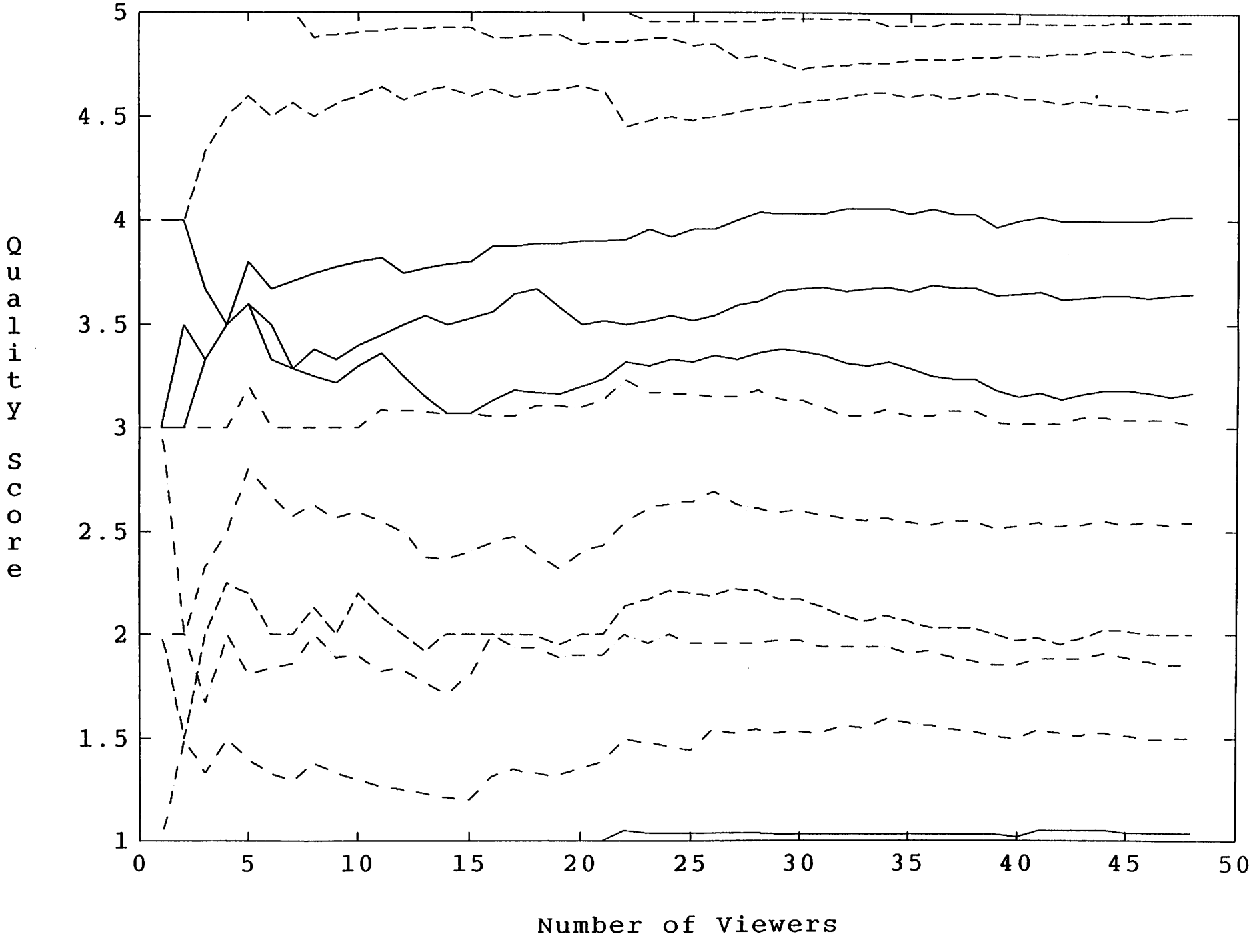

Figure 4. Max and Min of Quality Scores for 8 Video Systems

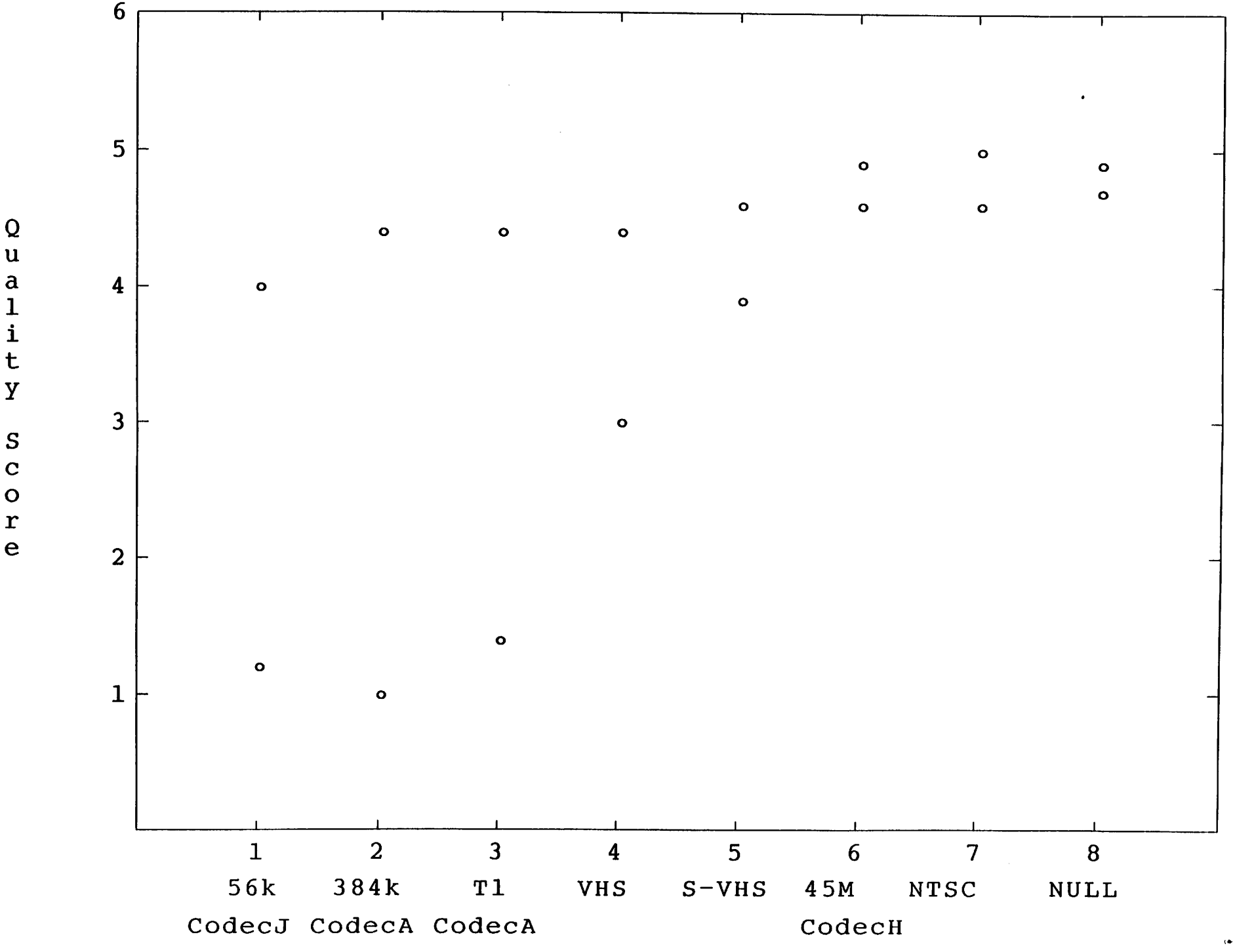

Figure 5. Histograms -- 4 Scenes Codec J 56 Kbs

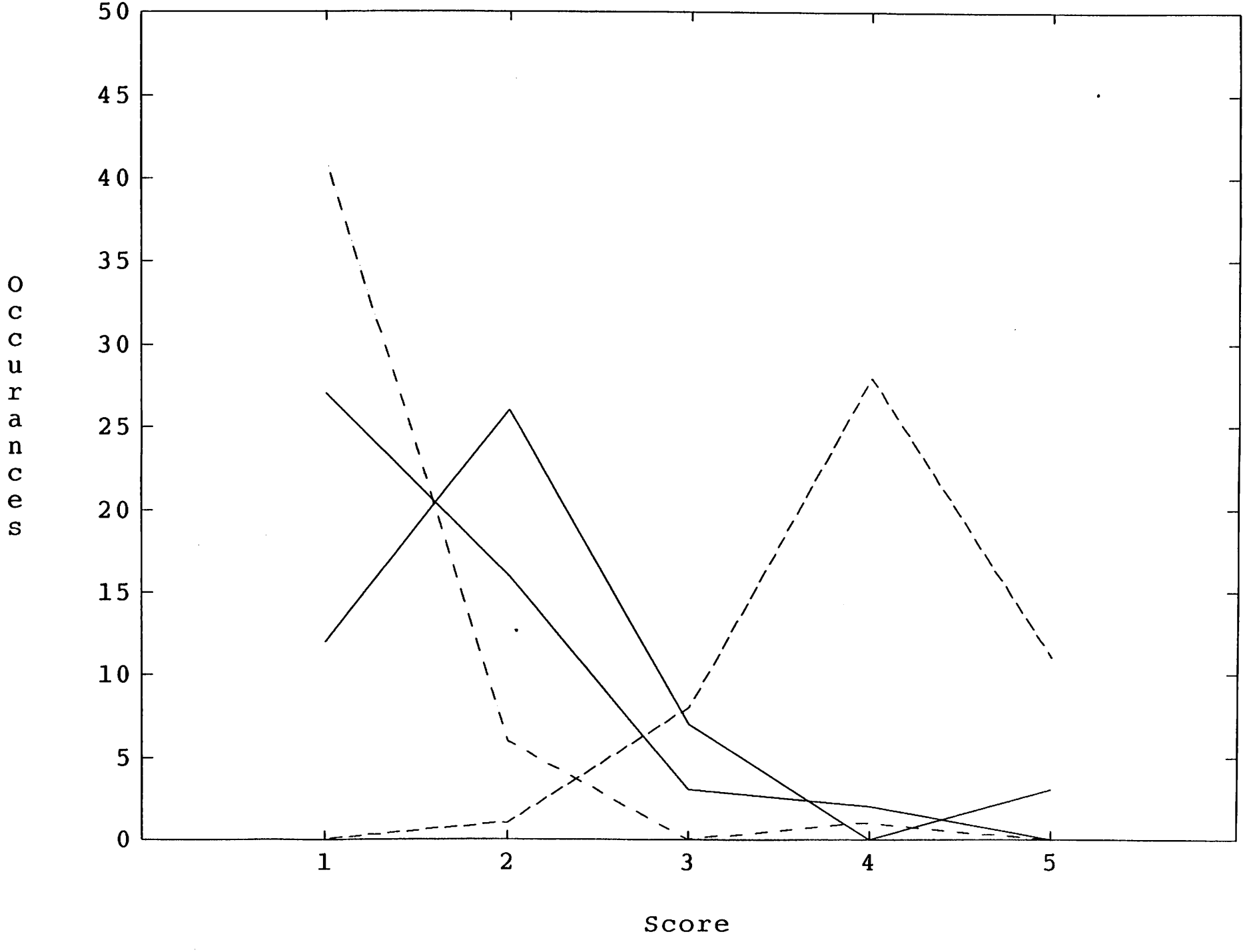

Figure 6. Histograms -- 5 Scenes Codec A 384 Kbs

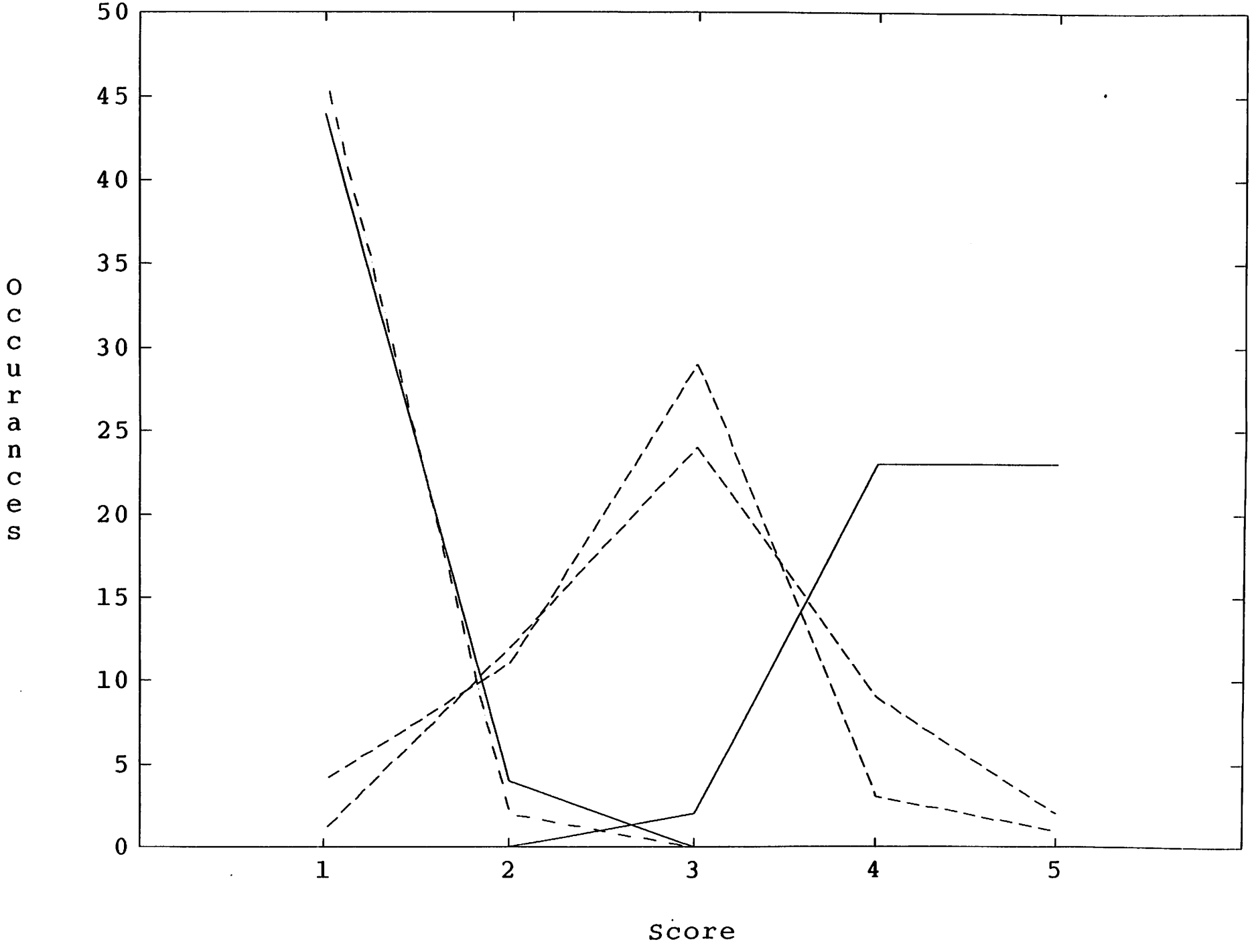

Figure 7. Histograms -- 4 Scenes Codec A 1536 Kbs

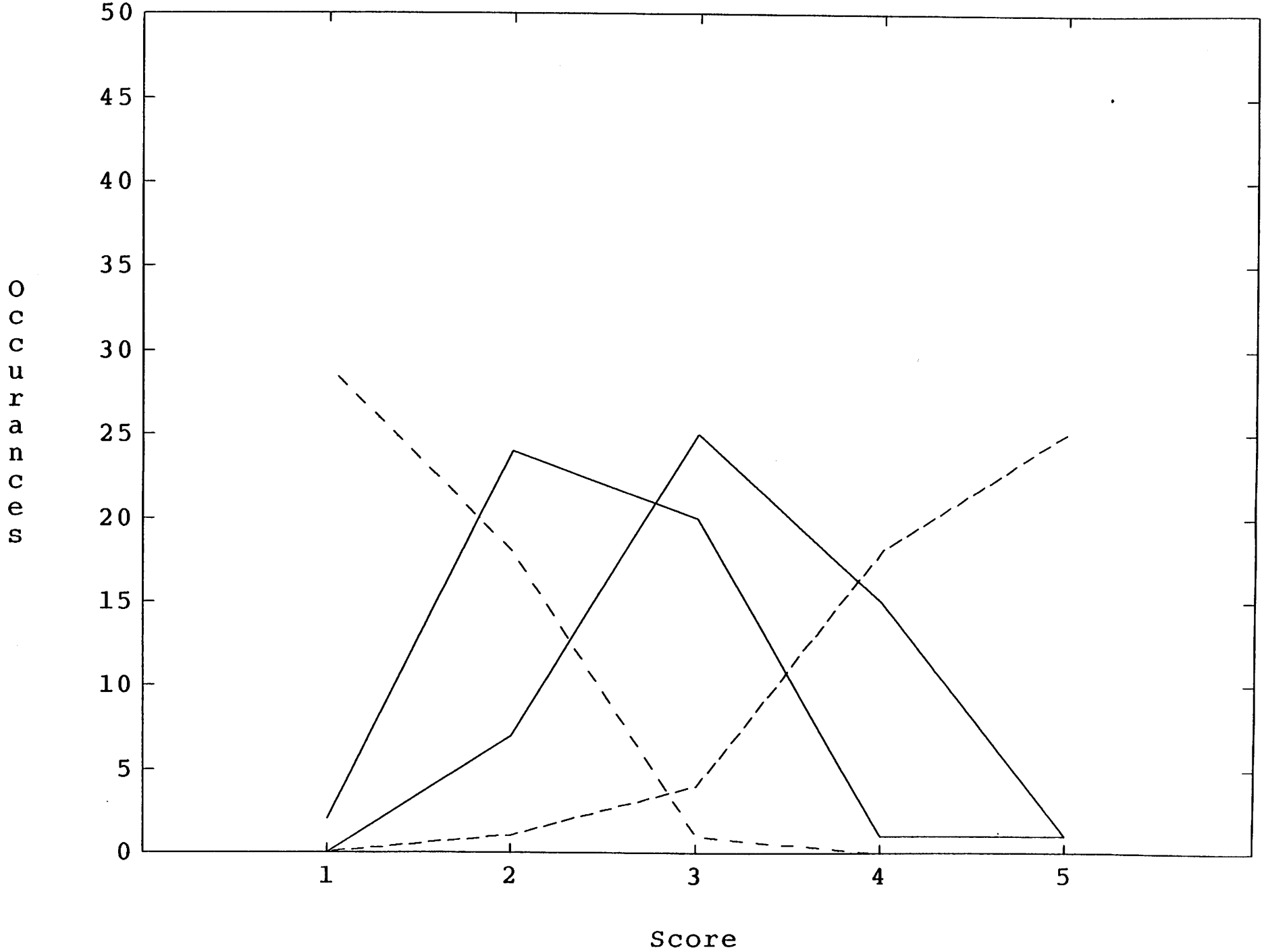

Figure 8. Histograms -- 8 Scenes VHS

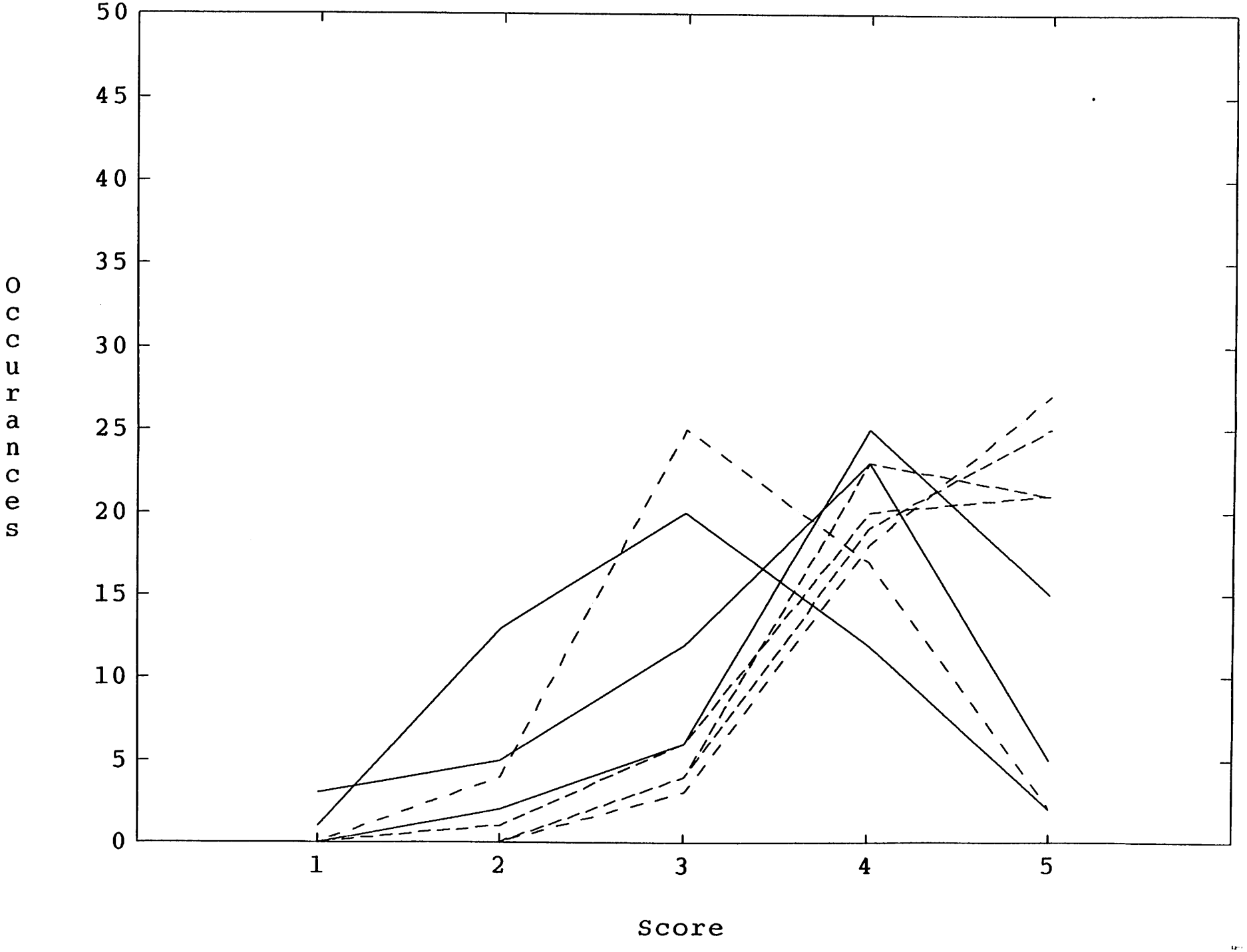

Figure 9. Histograms -- 7 Scenes S-VHS

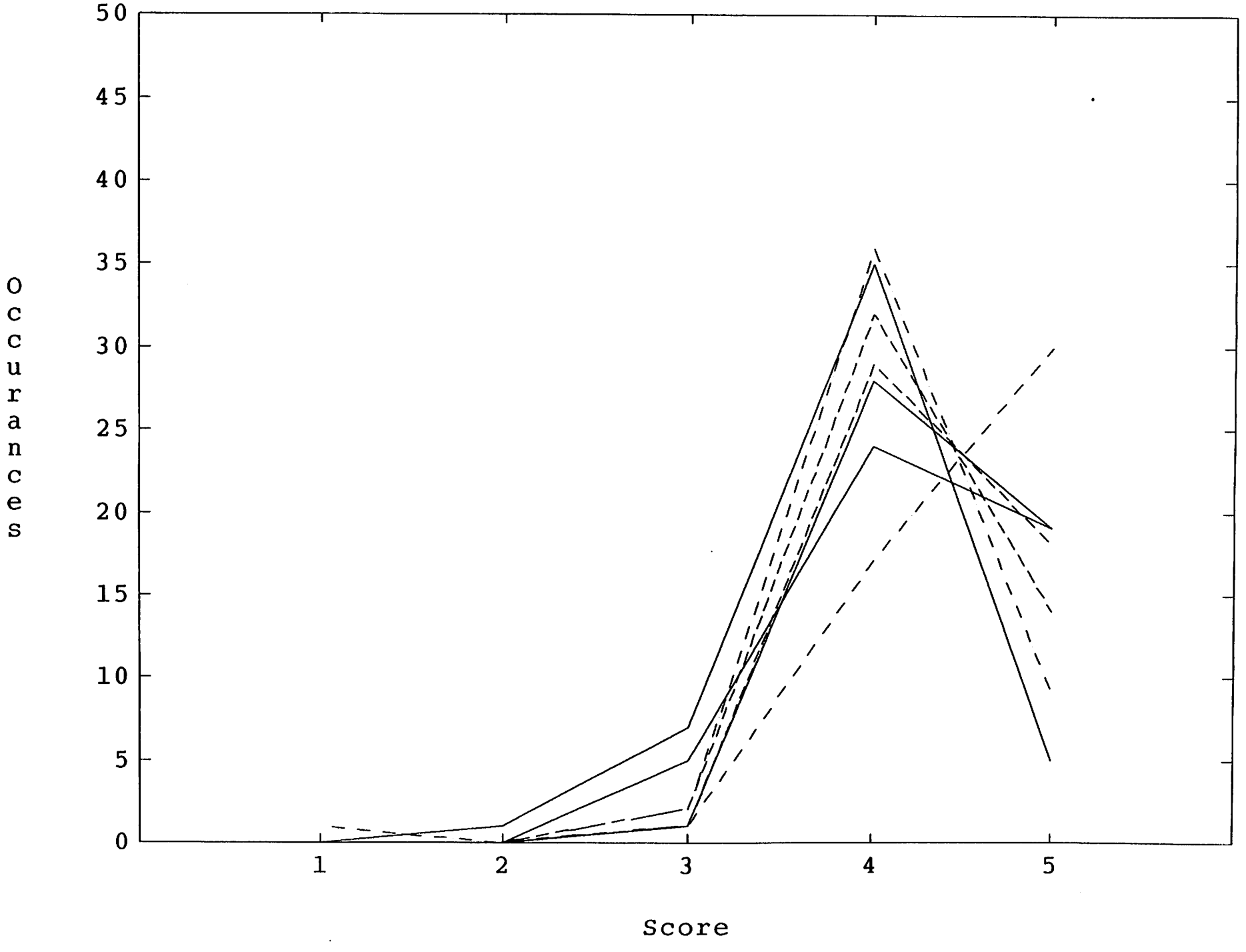

Figure 10. Histograms -- 6 Scenes Codec H 45 Mbs

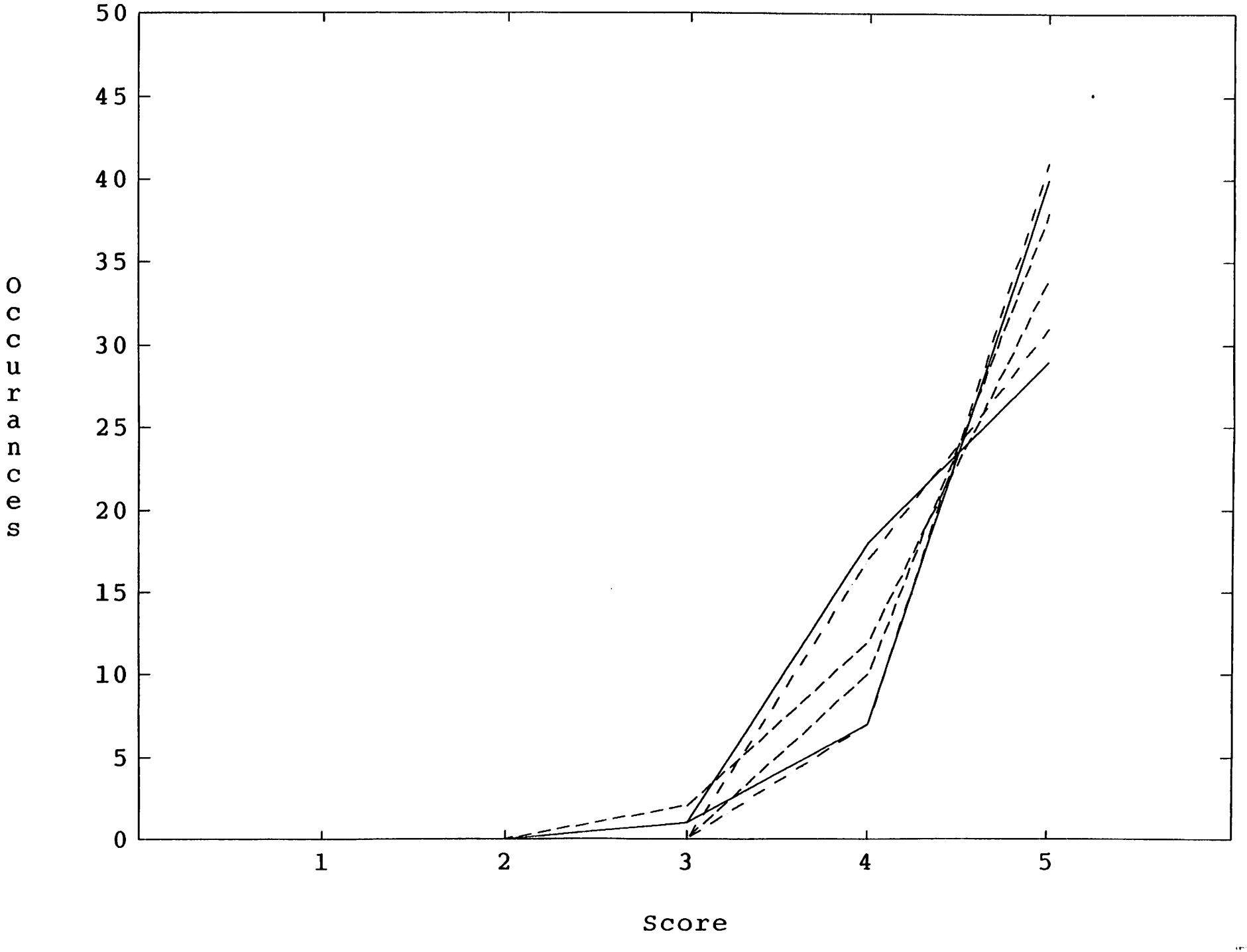

Figure 11. Histograms -- 6 Scenes NTSC

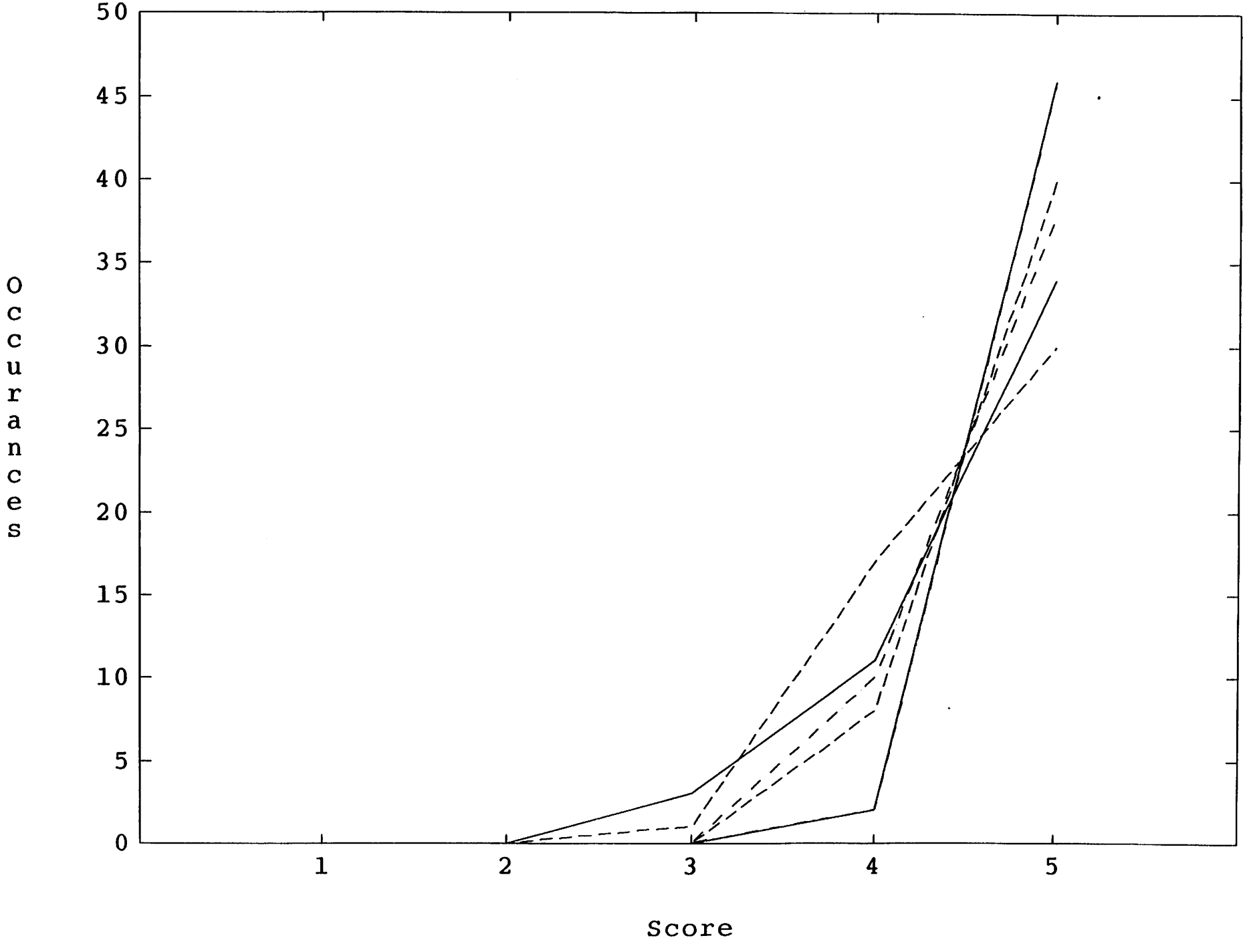

Figure 12. Histograms -- 6 Scenes NULL System