RESEARCH DEPARTMENT

ANALYSIS OF DATA FROM SUBJECTIVE TESTS EMPLOYING GRADED SCALES OF OBSERVER PREFERENCE

Report No. A-037

(1955/23)

THE BRITISH BROADCASTING CORPORATION
ENGINEERING DIVISION

RESEARCH DEPARTMENT

ANALYSIS OF DATA FROM SUBJECTIVE TESTS EMPLOYING GRADED SCALES OF OBSERVER PREFERENCE

Report No. A-037

(1955/23)

W Proctor Wilson

E.R. Wigan, B.Sc. (Eng.), A.M.I.E.E.

(W. Proctor Wilson)

Report No. A-037

ANALYSIS OF DATA FROM SUBJECTIVE TESTS EMPLOYING GRADED SCALES OF OBSERVER PREFERENCE

Summary Title Page

SUMMARY 1

1 INTRODUCTION 1

2 METHOD OF ANALYSING CLASSIFIED DATA 2

2.1. The Median Grading 2

2.2. Reliability of the Median Grade 3

3 SUMMARY OF RE-ANALYSIS OF THE DATA OF REPORT A-032/2 5

3.1. Preference for F.M. Versus A.M. 5

3.2. Interpretation of the Reliability Limits Set to X_m 5

3.3. Comments on the Reliability Limits Shown in Fig. 3 6

4 CONCLUSION 8

APPENDICES 10

PRIVATE AND CONFIDENTIAL

September 1955

Report No. A-037

(1955/23)

ANALYSIS OF DATA FROM SUBJECTIVE TESTS EMPLOYING GRADED SCALES OF OBSERVER PREFERENCE

SUMMARY

This report deals with the statistical analysis of questionnaires in which observers enter their opinions under a series of graded classifications, such as "Bad", "Indifferent", "Good".

It is shown that the significance of the opinions expressed can be put into simple and realistic terms. The technique is demonstrated by applying it to the questionnaires detailed in an earlier report on the preference of listeners for F.M. versus A.M. transmission.

1. INTRODUCTION.

It is common practice to assess the merit of a transmission system by asking a number of observers, viewers or listeners, to "grade" the quality of the reproduction in terms of a prearranged scale of classification; for instance, visual or audible distortion might be classified as "Not Perceptible", "Just Perceptible", "Perceptible", "Poor", "Gross". A rough estimate of the quality of reproduction can be made from the totals scored under each of these headings by a large number of observers.

This type of test would be much more valuable if deductions from the data could be made with greater precision, and more particularly if the degree of precision could be estimated by a statistical approach of some kind. A method of doing this is the subject of this report.

In order to test the method, data has been abstracted from Research Report No. A-032/2, Final Report on "Wrotham A.M./F.M. Listening Tests". Graded scales of preference were included in the questionnaires on which the report is based. Broad conclusions were drawn from the tabulated data presented. It will be seen from what follows that by recasting this data some fairly precise conclusions can be reached, and that the reliability of these conclusions can be tested statistically.

Although the object of the re-analysis was primarily to test the technique, rather than to review the conclusions reached in the report, a brief re-assessment of the A.M./F.M. trials is included here since it was found that the re-analysis brought out very clearly several important points.

The success of this re-analysis justifies the choice of graded scales of preference in an investigation of audible distortion now in hand.

2. METHOD OF ANALYSING CLASSIFIED DATA.

2.1. The Median Grading.

To illustrate the method, consider Section 2.6 of Report No. A-032/2, which gives the totals of the answers to questions on fading:

TABLE 1

Range from Wrotham	40 to 70 miles
Total number of observers	23
Total reports of fading	16
(Whence we deduce "Nil" reports as	7)
Fading reported "Slight" (SL)	6
Fading reported "Marked" (M)	6
Fading reported "Severe" (SEV)	4

This data is not an entirely realistic statement of the observers' opinions; it is obvious for instance that classification "M" will contain all observations of fading which exceeded "Slight" but fell short of "Severe". The six observations under class "M" could therefore be justly divided into two equal parts, half representing those falling between "SL" and "M" and half those between "M" and "SEV". Proceeding in this way we get Table 2:

TABLE 2

(Fading at 40 to 70 miles)

Class Division	NIL	SL	M	SEV
Answers to Questionnaire	7	6	6	4

Redistributed scores falling into class intervals:	3·5	6·5	6	5	2
Total scores above each class mark:	23	19·5	13	7	2
% based on total of 23:	100	84·5	56·4	30·4	8·7
Deviate* (Table of Erf.):	+ ∞	+ 1·02	+ 0·16	− 0·5	− 1·36

This method of tabulation is designed to show how the observers' scores are distributed; to deduce the "median grading", i.e. that grading which divides the total of scores into equal parts; to estimate the spread of the scores about the median; and to bring to light any abnormalities which may exist in the distribution.

*Erf. is an abbreviation of Error Function, and Erf. $(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}\,dx$.

The "deviate" is derived, as described in Appendix A, from Erf. (percentage/100) but is not equal to it.

Fig. 1 shows that by choosing equal intervals of the x-axis to represent the various grades, and plotting the deviates of Table 1 on the y-axis, a straight line can be drawn through the points. For reasons discussed in Appendix B, the x-axis is scaled from 0 to 1·0. The "median grading" (X_m) is 0·43 where the deviate is zero, corresponding to a 50% score, which means that as many observers would judge the grading to be higher than 0·43 as would judge it to be lower.

Since the data as plotted in Fig. 1 falls on a straight line, it can be assumed that the distribution of the scores (when redistributed into class intervals) is at least approximately Gaussian, and normal. This conclusion is supported by similar plots of other data taken from the report.

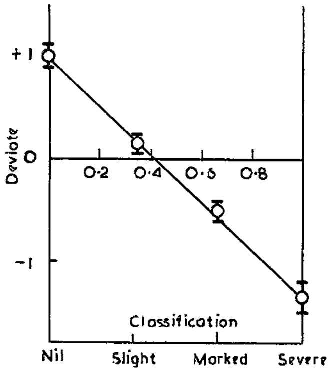


Fig. 1 - Analysis of fading reported at 40/70 miles

$N = 23$

$X_m = 0{\cdot}425$

$S = 0{\cdot}43$

2.2. Reliability of the Median Grade.

It will be obvious that the "average grade" could have been calculated from the first line of Table 2 giving "weights" to each grade in accordance with the 0 to 1·0 scale of Fig. 1;

$$\text{i.e. "average grade"} = \frac{7(0) + 6(0{\cdot}33) + 6(0{\cdot}66) + 4(1)}{23} = 0{\cdot}435.$$

However, the figure shows that we are dealing with a distribution of data which is approximately Gaussian and symmetrical. It is therefore to be expected that the median (X_m = 0·43 from the figure) and the average just calculated should coincide.

Using well-known statistical techniques the "reliability" of the median* can therefore be computed, giving an estimate of where the median would fall if the test were repeated by another group of twenty-three observers. Such an estimate must necessarily be somewhat speculative since the behaviour of the second group of observers and their receivers cannot be predicted. Lacking any other data it has to be assumed that these observers will be subject to the same kind of variability as was the original group. The estimate becomes more doubtful as the number of observers in a group gets less (for example the report includes one group of only eight observers).

The reliability is estimated by computing the range over which the median might fluctuate were such group tests to be repeated many times. The inverse slope of the line in Fig. 1 (0·43) is proportional to the standard deviation (S) of the Gaussian distribution of the grading assigned by the twenty-three observers. It can be shown that there is a 95% chance that the median grading of a repeated test would be within the range

$$X_m \pm \frac{2S}{\sqrt{N-1}}$$

*In what follows, formulae strictly applicable only to the *average* have been used, and from these the reliability of the *median* has been calculated. The data is too meagre to establish the identity of median and average with any precision, but identity is presumed.

if N, the number of observers, is large (greater than fifty). If N is as small as eight, the range is increased by roughly 10%. Similarly there is a 68% chance that the median will lie between

$$X_m \pm \frac{S}{\sqrt{N-1}} ,$$

again increased 10% when N = 8. Here the increased range will be used throughout so that estimates of reliability will be, in most cases, pessimistic.

The estimates of reliability cannot be very precise in any case since they are derived from the slope of a line which passes among a series of points on a graph. However, the position of these points is subject to uncertainty for a reason peculiar to data obtained by the "grading" of observations, when the number of observations is small. This arises as follows:

Take as an example the table on page 9 of the report which gives the grading of the hiss heard on F.M. receivers at a range of 70 to 90 miles.

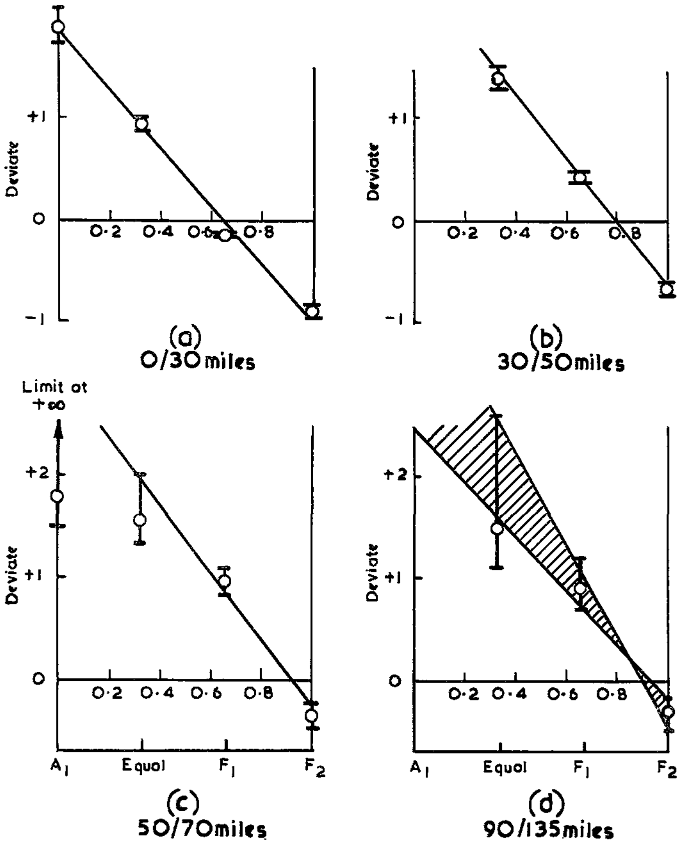
TABLE 3

Classification	IP	JP	P	SD	D
Observed Gradings	6	2	1	3	0

Adjusted to class interval:	3	4	1·5	2	1·5	0
Totals above intervals:	12	9·0	5·0	3·5	1·5	0
Totals as %:	100	75	41·5	29	12·5	0
Deviates:	+ ∞	+ 0·67	− 0·22	− 0·55	− 1·16	− ∞

It can be seen that if one observer had made an error and placed his estimate of hiss as "Perceptible" (P) instead of "Slightly Disturbing" (SD), the entry in the top line of the table would read 6, 2, 2, 2, 0. This has a large effect upon the entries under deviates; in particular the −1·16 entry becomes −1·39. Any attempt to fit a straight line to the plot of deviates must take account of the possibility of such errors, and this is done by assuming that the figures in line 3 of Table 3 are subject to uncertainties of ± 0·5. When the resulting percentages are converted finally to deviates in the last line of the table, corresponding upper and lower limits can be set to the quantity to be plotted. Throughout this report the horizontal bars indicate these limits. Notice that occasionally one limit may be at infinity as in Fig. 2(c).

It will be seen that the provision of these limits makes the choice of the line of best fit very much easier; in fact, in this figure it is scarcely possible to make a better fit. In other cases (e.g. in Fig. 2(d) where N = 8) several lines are possible, but it is found that in general the value of the median grade is scarcely altered. The uncertainty introduced by the limits is therefore not so much in the median grade as in the standard deviation associated with the median. Where this occurs the line has been chosen which gives the most pessimistic estimate of the reliability of the median, and its slope is used to compute the limits set out in the next section and in Fig. 3.



(a) 0/30 miles

(b) 30/50 miles

(c) 50/70 miles

(d) 90/135 miles

Fig. 2 - Preference FM/AM

3. SUMMARY OF RE-ANALYSIS OF THE DATA OF REPORT A-032/2.

3.1. Preference for F.M. Versus A.M.

As an experiment the data given in the report on the preference expressed for F.M. versus A.M. reception was treated in exactly the same way as the tabulated data on hiss, interference and fading. To carry out the analysis it was necessary to invent classifications which were: "A.M. Slightly Preferred" (A1); "No Preference" (EQUAL); "F.M. Slightly Preferred" (F1); "F.M. Markedly Preferred" (F2). The few preferences expressed for A.M. did not distinguish between A.M. and A.M.L. so no special classification was needed for these cases.

It will be seen from Figs. 2(a) to 2(d) that the plot of the deviates fit a straight line remarkably well, due allowance being made for the ± 0·5 limits indicated by the horizontal bars.

3.2. Interpretation of the Reliability Limits Set to X_m.

To illustrate the use of the reliability limits take the case of Fig. 2(d):

Range	90 to 135 miles
Number of observers	N = 8
X_m	0·93*
S	0·37*

$$95\% \text{ limits on } X_m \text{ are } 0{\cdot}93 \pm (1{\cdot}1)\frac{(2S)}{\sqrt{7}} = 0{\cdot}93 \pm 0{\cdot}308$$

$$68\% \text{ limits are half as wide } = 0{\cdot}93 \pm 0{\cdot}154$$

*Taking the lower line which gives a pessimistic estimate of reliability.

These limits mean that there is a 2·5% chance that X_m will lie above (0·93 + 0·308) = 1·238, or lie below (0·93 − 0·308) = 0·622; and a 16% chance that X_m will be above (0·93 + 0·154) = 1·084, or be below (0·93 − 0·154) = 0·776.

The classification divisions lie at:

A1	0·0
EQUAL	0·33
F1	0·66
F2	1·00

So we conclude, taking the lower 95% limit (ignoring the upper limit as meaningless in this context), that the odds are about 40/1 against any eight listeners at 90 to 135 miles range giving it as their overall opinion (median grading) that their preference for F.M. is only "Slight", (X_m = 0·66) and astronomical odds against their expressing no preference at all (X_m = 0·33).

3.3. Comments on the Reliability Limits Shown in Fig. 3.

In Fig. 3 the result of re-analysing the data in the report (see Table 4) is set out graphically. The shaded areas represent the 68% reliability limits and the dotted lines, outside, the 95% limits. There is a 2·5% chance that the median grade might lie above the upper dotted line and an equal chance that it might lie below the lower dotted line. The median grading (X_m) assigned in the tests reported lies in the middle of the shaded area.

The opening out of the limits at large ranges is not due, as might be supposed, to the greater variability of siting of the receivers and incidence of interference; this is proved by examination of the relevant values of S (Table 4) which show no marked or systematic increase with range. The spreading of the limits is in fact almost entirely due to the reduction in the number of observers (N) at large ranges, for the spread is proportional to $1/\sqrt{N-1}$. The fact is that at these ranges the data is too meagre to yield any highly precise conclusions, hence the wider limits. However at ranges of 70 miles or less N is never less than 15 and the 95% limits on X_m are not too widely spread. The exception to this is in the data on impulsive interference.

Dealing with impulses first it is at once obvious from Fig. 3(d) that no sharp distinction whatever can be drawn between A.M. and A.M.L.; although at short ranges the median grading is lower with the limiter than without it, the diversity of opinion, as shown by the values of S, is so great that the differences between the medians is quite insignificant in the sense that there is a high probability that a repetition of the tests would reverse the position. At 60 miles range there is a 16% chance of the interference with either system being graded as more than "Perceptible" by a group of fifteen observers and there is about 10% chance of it being graded as "Slightly Disturbing".

With an F.M. system, on the other hand, roughly the same odds apply to a range of 85 to 90 miles, whereas at 60 miles the chance of the interference being graded as above the "Just Perceptible" margin is less than 16%.

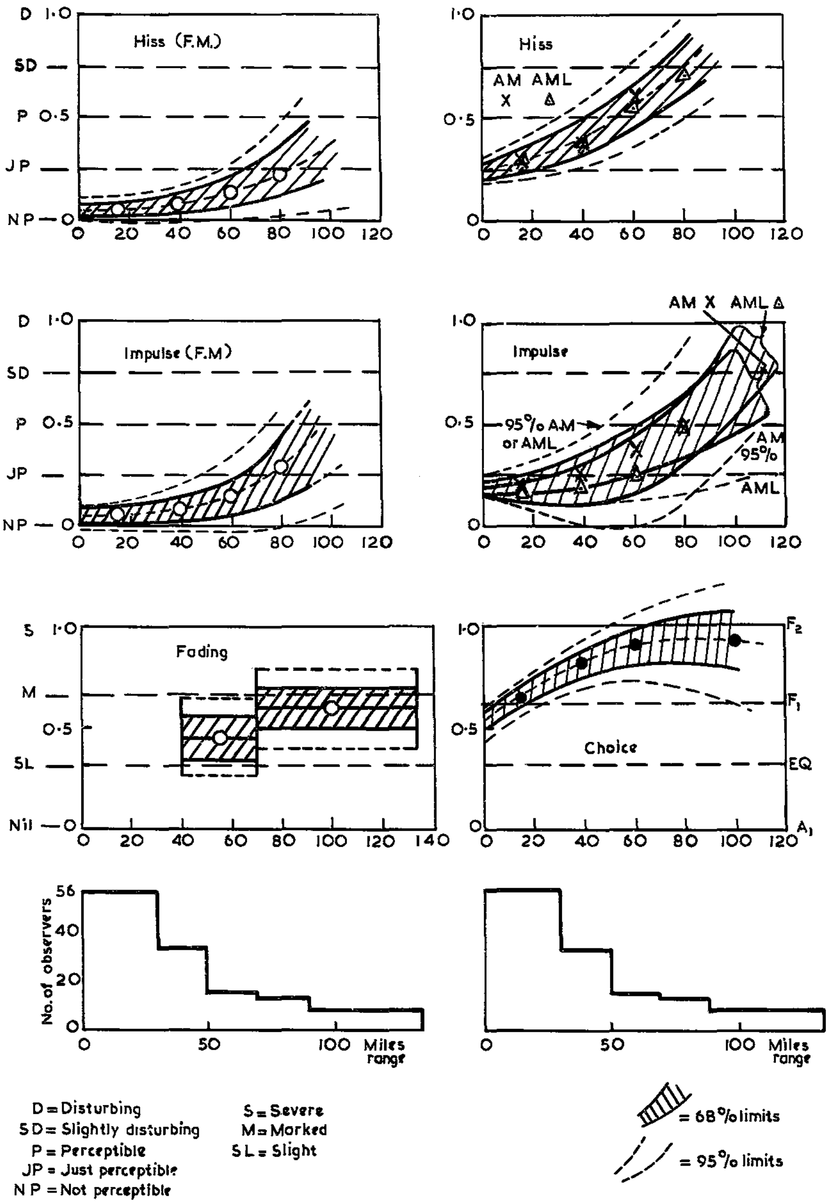


Fig. 3 - Summary of data

The graph depicting the grading of general preference for F.M. versus A.M. or A.M.L. is interesting in suggesting a significant preference for F.M. even at minimum range. This however is very probably an artifact produced by the grouping of observers in the first 30 miles of range. The median grading assigned by this group was 0·65, and this is plotted on the graph at 15 miles, the middle of the range involved. It is more than likely that the preference for F.M. was expressed only by those in the outer fringe of·the 30 miles.

There is however, at 60 miles, a perfectly unambiguous preference for F.M. which is so marked that the odds are 100 to 1 against the group preference of fifteen observers falling as low as the F1 margin, "F.M. Slightly Preferred".

4. CONCLUSION.

It is not the purpose of these notes to re-assess in detail the F.M./A.M. trials, but rather to illustrate by the statements of the previous section the order of precision with which conclusions can be drawn from subjective tests using a graded scale of preference. The incomplete analysis made here should be sufficient to demonstrate the point.

TABLE 4

Summary of Data

Symbols:

N = Number of observers.
X_m = Median grading assigned by N observers.
S = Standard deviation of the N individual gradings.
S_m = Standard error of X_m increased by 10% (to allow for the worst case in which N = 8) = $(1{\cdot}1)S/\sqrt{N-1}$.

Note 1: All quantities are in terms of a scale of grading from 0 to 1·0.
Note 2: The 95% reliability limits lie at $X_m \pm 2\ S_m$, and the 68% reliability limits lie at $X_m \pm S_m$.

PREFERENCE FOR F.M. VERSUS A.M.:

RANGE Miles	N	X_m	S	S_m
0/30	56	0·65	0·344	0·055
30/50	32	0·81	0·321	0·064
50/70	15	0·90	0·295	0·087
90/135	8	0·93	0·370	0·153

Class Boundaries:

	A1	EQUAL	F1	F2
at	0	0·33	0·66	1·00

FADING Miles	N	X_m	S	S_m
40/70	23	0·43	0·43	0·101
> 70	20	0·60	0·426	0·108

Class Boundaries:

	NIL	SLIGHT	MARKED	SEVERE
at	0	0·33	0·66	1·00

HISS Miles	N	X_m			S			S_m		
		F.M.	A.M.	A.M.L.	F.M.	A.M.	A.M.L.	F.M.	A.M.	A.M.L.
0/30	56	0·04	0·30	0·30	0·234	0·333	0·333	0·035	0·049	0·045
30/50	32	0·06	0·38	0·36	0·313	0·392	0·500	0·062	0·077	0·099
50/70	15	0·14	0·61	0·54	0·227	0·526	0·455	0·061	0·151	0·134
70/90	12	0·23	0·72	0·70	0·416	0·327	0·313	0·138	0·108	0·104

IMPULSE Miles	N	X_m			S			S_m		
		F.M.	A.M.	A.M.L.	F.M.	A.M.	A.M.L.	F.M.	A.M.	A.M.L.
0/30	56	0·05	0·21	0·17	0·250	0·384	0·267	0·037	0·057	0·040
30/50	32	0·08	0·28	0·18	0·345	0·435	0·468	0·068	0·086	0·092
50/70	15	0·12	0·38	0·26	0·289	0·606	0·556	0·085	0·178	0·263
90/135	8	0·29	0·52	0·50	0·465	0·370	0·460	0·154	0·123	0·152

Class Boundaries for HISS and IMPULSE:	IP	JP	P	SD	D
at	0	0·25	0·5	0·75	1·0

APPENDIX A

Conversion of Observers' Scoreś to Erf. Deviates

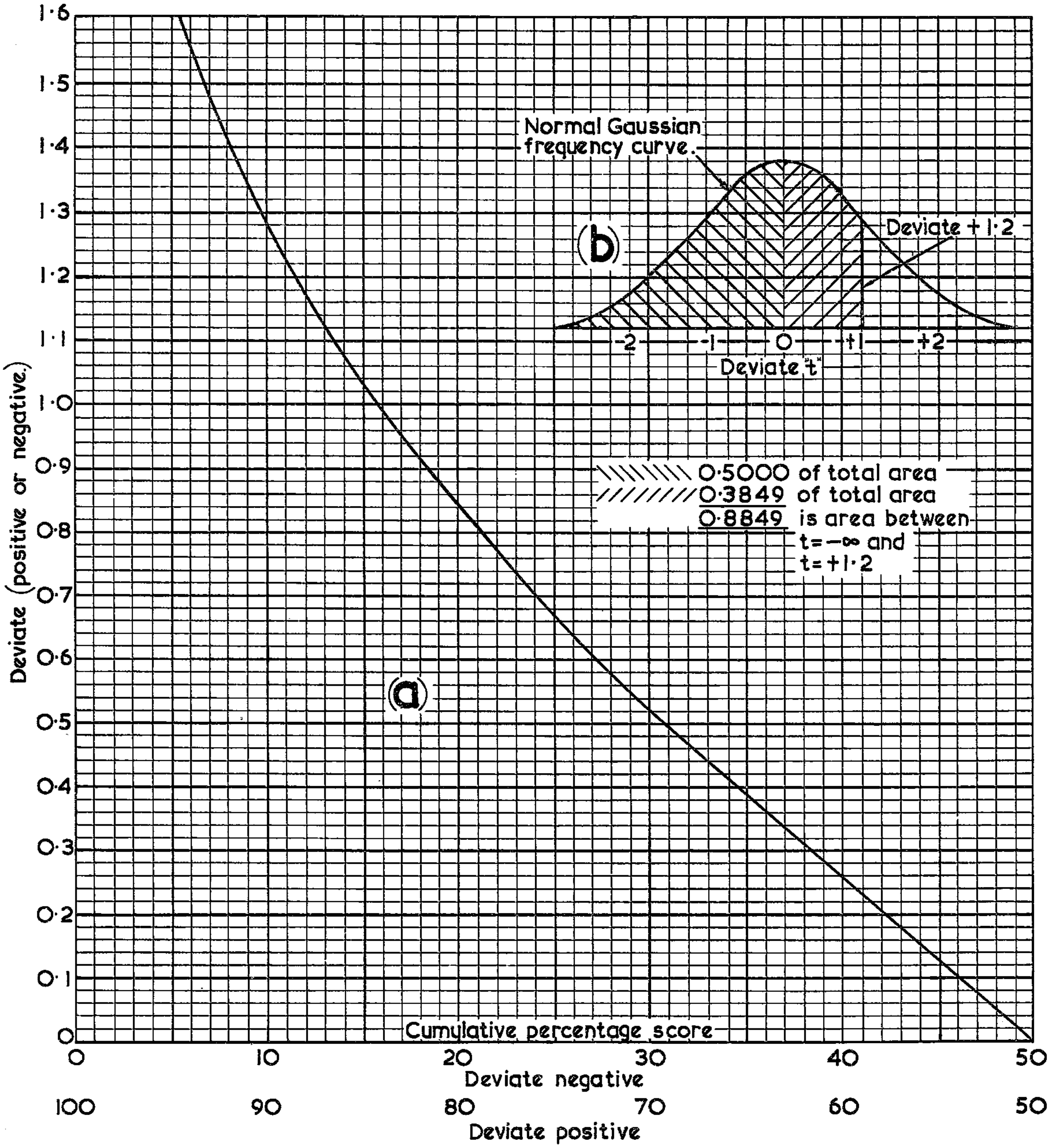


Fig. 4 - Total shaded area represents the cumulative % score

Fig. 4(a) shows the relationship between the deviate and the cumulative percentage derived from the "adjusted" data of the observers' report. Tables of the "Cumulative Normal Frequency Distribution" from which Fig. 4(a) is constructed can be found in any reference book on statistics; but they must be used with care. For example, Table 8.6 on Page 180 of "Statistical Methods" (Snedecor) gives the fraction of the area under the Normal Frequency Curve of a Gaussian distribution expressed as parts in 10 000, the area being that lying between the median and the ordinate at a distance t along the x-axis of the Frequency Curve. The fractional area we are concerned with (corresponding to the percentage in the penultimate line of Table 2 for instance) is the area lying between the t ordinate and $t = -\infty$.

To construct Fig. 4(a) proceed as follows: To plot the point corresponding to the deviate + 1·2 for instance, look up the cumulative normal frequency as 0·3849 and add 0·5, yielding 88·49% (see Fig. 4(b)). By symmetry, the one curve of Fig. 4(a) serves for both positive and negative values of t, corresponding to percentages respectively higher and lower than 50%.

APPENDIX B

Choice of X-Axis Scale

There is no *a priori* justification for the spacing of the class-marks at equal intervals on the x-axis of the deviate plot. It is clear from the figures, however, that such an arrangement leads to a very simple presentation of the data and that the probability function represented by the straight line drawn among the plotted points is at least approximately Gaussian. The "fit" is in some cases remarkable. Only five of the thirty deviate plots are shown here. Some of the thirty failed, as in Fig. 2(a), to yield a line passing through the ±0·5 limits but very few failed if the limits were made ±1·0.

It will be noticed that an arbitrary scale of 0 to 1·0 is used in all figures in spite of the differences in the number of class intervals. It may well be that the class interval should be used as a unit. However if this is done the reader cannot appreciate at a glance, as he can with the 0 to 1·0 scale, where the median grade lies relative to the lowest or highest class assigned. On the other hand with the 0 to 1·0 scale, the reader can appreciate that a grading of 0·5 represents the mid-grade of all those available.

The class interval is, however, an extremely artificial unit. It is most desirable that data obtained from a questionnaire based on, say, four classes should be directly comparable with one based on seven classes. It seems likely that the use of a 0 to 1·0 scale fulfils this purpose provided certain conditions are met. A controlled experiment would be necessary to prove it, but in general terms the argument is as follows:

Although the data obtained from questionnaires are derived from the observers' purely subjective assessments, the factor which influences the data is purely objective. For instance the quantities X_m and S in Fig. 1 are a measure of the mean value and the variability of the fading at the twenty-three receiving stations reported on. Here we are assuming that each observer can be relied upon to make a true assessment of the grade of fading he experienced. If the judgement of all observers is poor their errors will increase S, but are unlikely to affect X_m greatly. The class descriptions include all possible, i.e. they extend from Nil to Severe, and are numbered from 0 to 1·0. If we set up a new test in which the extreme classes were the same, the number of intervening classes could have no influence upon the value of the mean and the variability, provided they were expressed in the scale of 0 to 1·0.

It is clearly of prime importance that the extreme classes bear the same, or effectively the same, title. The observer will draw no distinction between Nil and Not Perceptible, or between Severe and Gross for example. Provided the classifications used in both questionnaires are complete the conditions are met. The classification of Fig. 2 fails on this test; it could reasonably be contended that "A.M. Markedly Preferred" should have been included to the extreme left. This failure does not in any way invalidate the conclusions of Section 3; it has bearing only if a new assessment were carried out in which the questionnaire was couched in different terms.