NASA CONTRACTOR REPORT

NASA CR-2107

NASA CR-2107

THE METHODS OF PAIRED COMPARISONS AND MAGNITUDE ESTIMATION IN JUDGING THE NOISINESS OF AIRCRAFT

by Frank R. Clarke and Karl D. Kryter

Prepared by
STANFORD RESEARCH INSTITUTE
Menlo Park, Calif. 94025
for Langley Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • AUGUST 1972

1. Report No. NASA CR-2107	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle The Methods of Paired Comparisons and Magnitude Estimation in Judging the Noisiness of Aircraft		5. Report Date August 1972
		6. Performing Organization Code
7. Author(s) Frank R. Clarke and Karl D. Kryter		8. Performing Organization Report No.
		10. Work Unit No.
9. Performing Organization Name and Address Stanford Research Institute Menlo Park, CA 94025		11. Contract or Grant No. NAS1-10017
		13. Type of Report and Period Covered Contractor Report
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546		14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The point of subjective equality in regard to perceived noisiness for each of 14 pairs of aircraft noises was obtained using both the magnitude estimation technique and the method of paired comparisons. Both methods gave approximately the same estimates of the points of subjective equality for the noise pairs, and both showed similar correspondence to predictive physical measures. Nevertheless, the two methods appear to differ in some ways. Based on verbal report the method of paired comparisons appears to have greater face validity to the listeners. However, the magnitude estimation technique appears to be more efficient; for a given level of reliability it requires approximately 50% of the testing time required by the paired comparison method. The functions relating physical intensity to the estimated magnitude of subjective noisiness had slopes ranging from about .61 to .29 for the aircraft noises employed in this study, indicating a required change of about 5 to 10 dB for a doubling in subjective magnitude. Some physical units of noise measurement were found to be very predictive (standard errors of estimate as low as 1.9 dB) of the subjective judgments of noisiness.

17. Key Words (Suggested by Author(s)) Perceived noise Subjective noise Noisiness of Aircraft		18. Distribution Statement Unclassified - unlimited	
19. Security Classif. (of this report) unclassified	20. Security Classif. (of this page) unclassified	21. No. of Pages 39	22. Price* $3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

THE METHODS OF PAIRED COMPARISONS AND MAGNITUDE ESTIMATION IN JUDGING THE NOISINESS OF AIRCRAFT

By Frank R. Clarke and Karl D. Kryter

Stanford Research Institute

INTRODUCTION

The main purpose of the experiments reported below was to compare results of noise judgment tests as obtained by the magnitude estimation technique with results obtained by the method of paired comparisons. Listeners were asked to judge how "noisy, annoying, or unwanted" they would find the sounds if heard regularly as a part of their living environment. A secondary purpose of the study was to compare how well different units of so-called Perceived Noise Level (PNL) and Effective Perceived Noise Level (EPNL), which are based on physical measures of the noise, predict the subjective judgments.

From a theoretical point of view, the method of paired comparisons appears to have great face validity. The operations are straightforward: two noises are presented and the listeners are simply asked to state which of the pair is the least annoying. By varying the level of one member of the pair, a psychometric function (percent listeners preferring one member of the pair as a function of the level of the variable member) is obtained. From this function the point of subjective equality may be estimated, i.e., we may determine the relative levels of the two sounds at which they are considered equally annoying. Despite the relative simplicity of this method there are problems associated with it. There is a significant "time error" associated with the listener's judgment, namely, other things being equal, with relatively intense sounds listeners will tend to judge the most recently presented member of the pair as less acceptable than the first member of the pair. Secondly, the response of the listener is merely "Prefer A" or "Prefer B," and information regarding <u>how much</u> more or less acceptable one of the pair is relative to the other, is lost.

The magnitude estimation technique may provide both theoretical and practical advantages when compared to the paired-comparison method though it has its own unique problems of interpretation. In the magnitude estimation technique, a "standard noise" is presented, and the listeners are told to assign some particular number (usually 10) to indicate the annoyance value for that noise. They are then asked to assign numbers to subsequent test noises, such that the ratio of each number to 10 (the value assigned to the standard) indicates the ratio of the relative annoyance of the test noise to the standard noise. Thus, if the listener feels the test noise is twice as annoying as the standard, he assigns it the value of 20, if only one-third as annoying, 3.3, and so forth. There are obvious philosophical arguments for and against such a technique, and such arguments have constituted a significant proportion of the subject matter of psychophysics for the last 100 years. These arguments will not be presented here. Empirically, it has been found that results will differ from experiment to experiment, depending upon the choice of intensity of the standard noise, the number assigned to the standard noise, the relative frequency with which the standard is presented, and other variables. Despite these difficulties, the magnitude estimation technique has some appealing features: the listeners' responses inherently contain more information than the response obtained from the paired-comparison method in that they indicate a relative magnitude, not simply a greater than or less than judgment. Thus, data obtained by the magnitude estimation technique lead to a more convenient and efficient scaling of the relative annoyance of various noises.

Because the magnitude estimation procedures show the percent change in the magnitude of the psychological attribute as a function of a degree of change in the physical stimulus, the subjective value to people from a change in the level of a noise can be expressed in quantitative terms; this is not possible with paired-comparison test data. It has generally been found that a change of approximately 10 dB in the stimulus intensity is equivalent to a doubling or halving, as the case may be, of the magnitude of subjective impression of the loudness or the noisiness of the stimulus; however, this is a matter of some dispute and further magnitude estimation data on this question are needed. Indeed, the full usefulness of the results of paired-comparison tests rests to some extent upon the establishment of valid subjective scales as derived from some form of subjective magnitude judgment tests.

If the information obtained by the magnitude estimation technique is sufficiently reliable (and valid), operating economies might result through use of this method when compared to the paired-comparison method. Also it is possible that the "time error" would have less deleterious effect on data when the comparison stimuli is far removed from the standard noise than in the method of paired comparisons in which the comparison is always made to the immediately preceding stimulus. Clearly, both methods have advantages and disadvantages relative to one another, and it is not surprising that both methods are in common use.

Immediate practical interest in the comparison of the two methods arises from the fact that one method was used to evaluate the effect of nacelle treatment developed by The Boeing Company (ref. 1) while the other method was used to evaluate nacelle treatment developed by the Douglas Aircraft Company (ref. 2). It is necessary to determine if the two methods are measuring "the same thing."

PROCEDURE

Stimulus Materials. Recordings for 14 pairs of aircraft flyovers were selected. Recordings were taken from landings and takeoffs actually employed in the Boeing Moses Lake tests, (5 pairs), the Wallops Station tests, (5 pairs), and from recordings employed in the Douglas tests conducted at SRI, (4 pairs). Boeing and Douglas tapes included both treated and untreated nacelles. The Boeing recordings were provided by The Boeing Company, the Douglas recordings were provided by the Douglas Aircraft Company, while the Wallops Station recordings were provided to SRI by NASA Langley. Table I gives additional data describing the stimuli. In addition, a standard noise was employed in this experiment. The standard noise was 4 seconds in durations with fixed intensity over that period. This noise was generated by shaping the output of a white noise generator such that it had a low frequency roll off of 3 dB per octave and a high frequency roll off of 6 dB per octave. The 3 dB downpoints were at 63 Hz and 500 Hz respectively.

Originally, three paired-comparison test tapes were constructed, each divided into two parts: Boeing (B1, B2), Douglas (D1, D2), and Wallops (W1, W2). Each part of a test contained twenty noise pairs: 15 flight pairs and 5 pairs of standard noises. For each pair of aircraft flights the standard was always presented at a fixed level, while a comparison occurred at the estimated point of subjective equality, as well as a value +4 dB and - 4 dB relative to that point. For each pair of the so-called standard noises one of the pair was always at a fixed level, while the other was either at that level, ±4 dB, or ±8 dB. Each of five pairs of flights occurred three times in part one of each test, such that the comparison stimulus was represented at each of its three values, -4, 0, and +4 dB relative to the estimated point of subjective equality. Pairs of standard noises occurred five times in part one of each test, such that each of the five comparison intensity levels was represented once. The order of items within each pair was determined systematically so that there were ten items presented in the order Standard followed by Comparison and ten items presented in the order Comparison followed by Standard. Part II of each test was a "mirror image" of Part I, i.e., pair number 20 on Part I became pair number 1 on Part II, with the order Standard followed by Comparison (or vice versa) reversed. The original design also called for two magnitude estimation tests, one of these was in fact Part I of the Wallops test, but with numbering appropriate for 40 items rather than 20 pairs. The other consisted of

Table I

DESCRIPTION OF AIRCRAFT NOISES EMPLOYED IN THIS STUDY

Item	Code No.	Aircraft	Operation	Dur.	Presentation Level dBC Stand.	Presentation Level dBC Comp.
1	D3	DC-8	TO	20.0		83,87,91
2	D4	DC-8	TO	22.0	91	
3	D7	DC-8	L	13.0		76,80,84
4	D8	DC-8	L	18.5	86	
5	D15	DC-8	TO	21.0		77,81,85
6	D16	DC-8	TO	23.0	89,93	
7	D17	DC-8	TO	31.5		86,90,94
8	B1	707-320B	L	14.0		78,82,86
9	B2	707-320B	L	14.0	91	
10	B3	"	L	14.5		77,81,85
11	B4	"	L	15.0	91	
12	B5	"	TO	36.0		89,93,97
13	B6	"	TO	26.5	97	
14	B7	"	TO	20.5	96	
15	B8	"	TO	31.5		93,97,101
16	B9	"	TO	32.5		83,87,91
17	B10	"	TO	26.0	92	
18	W1	880	TO	24.0	87-92	
19	W2	720	L	14.5		83,87,91
20	W3	C141	TO	17.5		87,91,95
21	W5	Jetstar	TO	13.5		85,89,93
22	W6	990	L	10.0		77,81,85
23	W7	CH47	FB	13.0		87,91,95

TO = Takeoff, sometimes simulated
L = Landing, sometimes simulated
FB = Flyby

five Boeing standards, three Douglas standards, and their associated comparison stimuli each at the three intensity levels employed in the paired-comparison tests. In addition to these 32 aircraft flight items, there were eight presentations of the standard noise, four at the level of the original standard, the remaining four being at ±4 dB and ±8 dB. The order of these items was determined randomly. It was to prove that three pairs of Boeing flights, one pair of Douglas flights, and two pairs of Wallops flights resulted in psychometric functions not crossing the 50% point. Two new paired-comparison test tapes of 20 pairs each and two new magnitude estimation tapes of 40 items each were prepared for subsequent testing to remedy this fault and to provide some additional data.

On all test tapes items were separated by 5 seconds, during which subjects recorded their responses and item numbers (or, when appropriate, pair numbers) were announced. Duration of tests varied from approximately 13.5 minutes to 15.5 minutes. Magnitude estimation tests were preceded by two presentations of the standard noise with an assigned noisiness score of 10.

Experimental Subjects. Twenty-two paid volunteers drawn from the community at large served as subjects in this experiment. Ages ranged from 19 to 67 with a median age of 40. Sixteen of the subjects were female, six male. Subjects were run in groups of six to eight with some exceptions due to illness and the need for make-up sessions. None of the subjects had previous experience with paired-comparison tests or magnitude estimation tests.

Instructions to the Subjects. The subjects were told of the general nature of the tests (to judge the relative acceptability of different noises), and were read the specific instructions regarding each of the experimental methods prior to the start of the experiment. The instructions were also printed on their answer sheets (see Appendix A).

Acoustic Environment. All tests were conducted in an anechoic chamber which had 21-inch long fiberglass wedges on all six surfaces. Measured from the tips of the wedges the internal dimensions of the anechoic chamber were 8.5 by 17.75 by 8 feet. The noises to be judged were presented via two Altec-Lansing A7-500 speaker systems each driven by an 80-watt McIntosh power amplifier. Conventional playback circuitry was employed with the exception of artificial quieting of the system noise between stimulus presentations and the use of an equalization network designed to provide as flat as possible frequency response at the listener positions within the room. Each speaker system was directed at four subjects seated in an arc of radius of 8-1/2 feet. The chord of each arc was approximately 5 feet. The sound pressure level of octave bands of noise with center frequencies ranging from 63 to 8000 cycles varied by less than ±2-1/2 dB at any listener position. A low-pass filter with 3 dB downpoint at 8000 Hz was used to minimize tape hiss. As all signals were recorded on the test tapes at a constant value of dBD_2, signal intensity was varied for the test items by means of an attenuator in the circuit.

Originally, the experimental design was well counterbalanced both with respect to test tapes and test method. The need for a make-up session to obtain data for those cases where the point of subjective equality was poorly estimated resulted in less than optimal counterbalancing for some purposes but appropriate counterbalancing was maintained for the comparison of the two methods.

Noise Analysis. Physical measures of noises were computed from one-third octave band sound pressure levels sampled and averaged over 1/2-second time intervals. A General Radio Type 1921 Real-Time Analyzer was used to produce, each 1/2 second, sound pressure level measurements in 24 one-third octave bands covering the frequency range 50 to 10,000 Hz. These data were recorded and processed in digital form. The end results of the analysis include the time-histories of sound pressure levels in each of the 24 bands and the so-called maximum and effective values of various weighted measures dBA, dBC, dBD_2, PNdB, PNdBM, and PNdB and PNdBM corrected for tonal content by two procedures and designated by the subscripts t1 and t2. These units and related frequency weightings and calculation procedures are given in refs. 3 and 4.

DATA ANALYSIS AND RESULTS

Paired Comparisons. For each pair of flights six paired-comparison judgments were made by each listener: standard vs. comparison at each of three intensity levels of comparison (spaced 4 dB apart) in both orders (standard followed by comparison, and vice versa). Psychometric functions showing percentage of listeners who preferred the standard as a function of the level of the comparison stimulus were plotted separately for the case in which the standard preceded the comparison and for the reverse order. These curves appeared to be relatively well fit by normal ogives with common slopes for the two presentation orders but differing slopes for the various pairs of flights. Thus with the restriction that the curves of the two presentation orders be fitted with a common slope, normal ogives were fitted to the data using the least square error criterion. The point on the fitted curve at which fifty percent of the listeners preferred the standard was taken as the point of subjective equality and was obtained for each presentation order separately. The difference in the point of subjective equality for the two stimulus orders was taken as the estimate of the "time error" and these data are reported in Table II. The average time error over the 14 pairs of stimuli was 3.6 dB.

Data for the two orders of presentation were then combined by averaging normal deviate scores whenever possible and by averaging the proportions in the six instances where one of the proportions was either 1.0 or 0. Normal ogives were fitted to the resultant functions according to a least square error criterion and are plotted with data points indicated by solid dots in Figs. 1 through 14. Points of subjective equality were calculated based on the fitted functions, and these values as well as the slope of the functions are also reported in Table II.

Magnitude Estimation Treated as Paired-Comparison Data. Magnitude estimation data may be reduced to paired-comparison data by the simple assumption that if a listener assigns a higher annoyance rating to one stimulus than to another, he has, in effect, judged that stimulus to be the more annoying. Of course for the resultant data to be valid and reliable, additional assumptions would be required, but our data will provide an empirical check upon the relative validity and reliability of this procedure and the assumptions need not be detailed at this point. Differing amounts of magnitude estimation data were obtained for the varying pairs of aircraft flights, ranging from a) one judgment per subject for the standard and one judgment per subject for each of the three levels of the variable comparison stimulus for a total of four judgments per listener per noise pair to b) a repetition of the paired-comparison test pairs but requiring magnitude estimation on the part of the subjects giving a total of 12 judgments per stimulus pair per subject.

There are three aircraft pairs for which 12 magnitude estimation judgments were obtained using the paired-comparison tapes. These data, after the magnitude estimation ratings were translated into greater than or less than judgments, were analyzed in exactly the same manner as were the paired-comparison data described above. The "time error" for these pairs were reported in Table III. Whereas the data are too sparse for statistical significance, indications are that the time error in the magnitude estimation technique is considerably less than that observed for the paired-comparison method. For the three flight pairs for which comparable data exist the average time error in paired comparisons was 4.9 dB as compared to 0.7 dB for magnitude estimation. The fitted psychometric functions and data points (open circles) are shown in figures 1, 2, and 3.

Three pairs of aircraft flights were presented in the paired-comparison format, with approximately one-half of the listeners hearing

Table II

SUMMARY STATISTICS FOR PAIRED-COMPARISONS DATA

Pair Number	Relative Intensity of Comparison Stimulus dB	Flight Pair	Aver. % Prefer Standard	Time Error	Slope of Psychometric Function	Relative P.S.E.
1	+4	D16,D17	(77)	8.5	.15	-1.61
	0		67			
	-4		32			
2	+4	W1,W6	(95)	6.4	.18	-5.13
	0		79			
	-4		60			
3	+4	W1,W7	70	-0.2	.18	1.11
	0		41			
	-4		18			
4	+4	W1,W2	(80)	1.7	.24	-0.03
	0		62			
	-4		13			
5	+4	W1,W3	87	3.0	.25	-0.58
	0		56			
	-4		20			
6	+4	W1,W5	(88)	-0.6	.40	1.39
	0		20			
	-4		(2)			
7	+4	B6,B5	70	4.0	.23	2.25
	0		21			
	-4		10			
8	+4	B7,B8	83	2.0	.31	0.99
	0		35			
	-4		6			
9	+4	B10,B9	57	2.1	.16	3.08
	0		29			
	-4		13			
10	+4	B2,B1	74	3.9	.16	0.38
	0		44			
	-4		25			
11	+4	B4,B3	75	4.6	.21	0.92
	0		41			
	-4		15			
12	+4	D4,D3	(98)	3.6	.29	-3.01
	0		83			
	-4		38			
13	+4	D8,D7	71	6.0	.21	1.10
	0		46			
	-4		12			
14	+4	D16,D15	86	5.1	.09	-7.02
	0		67			
	-4		65			

each pair in one order (e.g., standard followed by comparison) while the remaining listeners heard the pair in the reverse order. Thus, for each pair the standard was heard three times and each of the levels of the comparison noise was heard once. The resultant three points on the psychometric functions were fit by a normal ogive using a least square error criterion, and the data (open circles) and the fitted curves are shown in figures 4, 5, and 6.

The above data were obtained using the same tapes as used for the paired-comparison tests (of course the numbering system was changed from pair numbers to individual item numbers). The magnitude estimates which were converted into paired-comparison data were based on contiguous items only, i.e., the data for the standard member of the pair were not averaged over the multiple presentations of the standard.

For the remaining eight pairs of aircraft noises data were obtained in the minimal configuration, i.e., a single magnitude estimation per listener per standard and a single magnitude estimation per listener for each of the three levels of the comparison stimuli. Order of presentation was determined randomly and it was rare for a particular standard to be contiguous with one of its comparison stimuli. In addition, a retest for five of these pairs was obtained one week after the original data were collected. For the three pairs for which there was no retest, the magnitude estimation data were converted to paired-comparison data simply by comparing for each listener the magnitude he assigned to the single standard to the magnitude assigned to each of the three levels of the comparison, and rating the comparison as more or less annoying than the standard. The resulting psychometric functions were plotted, fit by a least square error criterion and these data (open circles) are shown in figures 7, 8, and 9. Associated slopes and points of subjective equality are tabulated in Table III. For the five pairs for which a retest was available, the data were analyzed in two ways. Comparisons for each test separately were made as above, the percentage values for the two tests were averaged, and the resultant psychometric function was fitted with a normal ogive by a least square error criterion. These data are not plotted in this report but percentage values, slopes, and points of subjective equality are reported parenthetically in Table III. Data were also analyzed in a manner to reduce the relative contribution of intra-listener variability; namely, test and retest data were combined by first taking the geometric mean of the two responses to each stimulus by each listener and then converting the magnitude estimates to paired-comparisons as described above. These data are plotted as open circles in figures 10 through 14 and associated values are given in Table III.

Magnitude estimation data interpreted in the conventional manner

Table III

SUMMARY STATISTICS FOR MAGNITUDE ESTIMATION DATA

TREATED AS PAIRED COMPARISONS

Pair Numbers	Relative Intensity of Comparison Stimulus dB	Flight Pair	n*	Aver. % Prefer Standard	Time Error	Slope of Psychometric Function	Relative P.S.E.
1	+4	D16,D17	12	70	1.8	.22	1.56
	0			36			
	-4			12			
2	+4	W1,W6	12	94	1.8	.23	-2.74
	0			70			
	-4			41			
3	+4	W1,W7	12	64	-1.7	.20	2.33
	0			31			
	-4			10			
4	+4	W1,W2	6	96	na	.34	-1.29
	0			73			
	-4			16			
5	+4	W1,W3	6	93	na	.27	-1.75
	0			73			
	-4			25			
6	+4	W1,W5	6	93	na	.40	0.36
	0			41			
	-4			4			
7	+4	B6,B5	4	80	na	.24	0.06
	0			59			
	-4			14			
8	+4	B7,B8	4	73	na	.22	1.42
	0			34			
	-4			18			
9	+4	B10,B9	4	50	na	.19	3.63
	0			30			
	-4			7			
10	+4	B2,B1	(4) 8	(74) 80	na	(.25) .35	(1.80) 1.82
	0			(26) 23			
	-4			(9) 2			
11	+4	B4,B3	(4) 8	(84) 89	na	(.28) .42	(0.84) 1.13
	0			(33) 32			
	-4			(10) 0			
12	+4	D4,D3	(4) 8	(94)100	na	(.24) .21	(-2.19)-2.60
	0			(66) 70			
	-4			(35) 39			
13	+4	D8,D7	(4) 8	(80) 82	na	(.19) .25	(0.04) 0.72
	0			(43) 36			
	-4			(25) 14			
14	+4	D16,D15	(4) 8	(85) 82	na	(.22) .22	(-0.53)-0.04
	0			(49) 48			
	-4			(25) 20			

*n is the number of magnitude estimates per listener upon which the data for a given pair are based.

were treated as follows: (a) in those cases in which there was but a single response per listener to a given stimulus the geometric mean over the twenty-two listeners was computed, and (b) in those cases in which there were multiple presentations of a given stimulus the geometric mean was taken over all responses by all listeners to result in a single magnitude estimate. The resultant functions were closely approximated by the power law and were so fit by the least square error criterion without *a priori* restrictions on the slope of the function. Resultant functions are shown in figures 15 through 28 with solid dots showing the data for the "comparison stimuli" and with the magnitude estimate for the standard plotted as an open circle on the fitted power function. The abscissa value at that point is taken as the estimate of the point of subjective equality and is reported separately in Table IV. The exponents of the power-law function are reported in Table V. It may be noted that these values average 0.42 which is somewhat higher than typically reported.

DISCUSSION

There is no straightforward way to evaluate the validity of the two methods for determining what the relative annoyance of various aircraft noises would be if heard in real life rather than in the laboratory. Formally, the validity of a test is its correlation with some external criterion. A good external criterion for annoyance is lacking though for practical purposes, the degree of annoyance of various environmental noises is thought to be reflected by the attitudes of citizens regarding these noises in real life. Even this is not always a good criterion as many different factors may affect attitudes towards given noises. More important from a practical point of view is the determination under the real-life conditions, of the physical character of the noises as present at the ears of the citizens from whom attitude information has been obtained. Needless to say, in our limited experiments we made no attempt to evaluate the formal validity of either method.

Yet something can be said about the relative validities of the two methods. Obviously, if both methods gave exactly the same results we could conclude that they were equally valid even though we would not know explicitly what the degree of validity might be for a specific purpose. Both methods have reasonable "face validity," i.e., the defining operations appear reasonable and when a significant proportion of people state that noise A is more annoying than noise B, we have little reason to doubt that this is the case. Or when the average magnitude rating assigned noise A is greater than that assigned noise B there is little

Table IV

POINTS OF SUBJECTIVE EQUALITY

(dB, ARBITRARY ZERO DIFFERENT FOR DIFFERENT PAIRS)

FOR MAGNITUDE ESTIMATION DATA ANALYZED IN TWO WAYS

AND FOR PAIRED-COMPARISON DATA

Pair Number	Flight Pair	MET Power Func.	MET as PC	P.C.
1	D16,D17	1.5	1.6	-1.6
2	W1,W6	-2.8	-2.7	-5.1
3	W1,W7	2.3	2.3	1.1
4	W1,W2	-1.3	-1.3	0.0
5	W1,W3	-1.2	-1.8	-0.6
6	W1,W5	0.6	0.4	1.4
7	B6,B5	0.0	0.1	2.2
8	B7,B8	2.0	1.4	1.0
9	B10,B9	4.0	3.6	3.1
10	B2,B1	2.0	1.8	0.4
11	B4,B3	1.4	1.1	0.9
12	D4,D3	-1.8	-2.6	-3.0
13	D8,D7	1.0	0.7	1.1
14	D16,D15	0.1	0.0	-7.0

reason to think that A is not more annoying than B. Despite the fact that both methods have reasonable face validity, there are no logical a priori reasons to conclude that both are measuring "the same thing." Hence, the empirical data reported herein.

First, we note that the two different methods of analyzing the magnitude estimation data given very similar results (see Table IV). In none of the 14 comparisons has the absolute value of the difference arising from the two methods for obtaining the point of subjective equality exceeded 0.8 dB and 11 of the 14 results differ by less than 0.5 dB. We shall use the magnitude estimation data as converted to paired-comparison data to compare to the paired-comparison data per se simply because both types of data can then be plotted on a common ordinate system. We see from figures 1 through 14 and more directly from Tables II, III, and IV, that the two methods give quite similar results. Points of subjective equality are within 2 dB of one another in ten out of 14 instances, and within 2.4 dB in 12 of the 14 instances. Two results differ by a fairly large margin: 3.2 dB and 7.0 dB respectively. There is no obvious explanation for the difference of 3.2 dB. Examination of figure 1 shows that both psychometric functions are "reasonably well-behaved" and the observed difference may well be real. In the second case the observed result is probably inaccurate for unknown reasons. As can be seen in figure 14 when the comparison stimulus was raised in intensity from -4 dB to 0 dB (re arbitrary level) the percent preferring the standard stimulus only increased from 65 percent to 67 percent. This is atypical of results obtained with all other pairs in this experiment as well as our experience in previous experiments.

Table VI shows r.m.s. error obtained when each of 18 physical measures of the noise stimuli are used in an attempt to predict the points of subjective equality. Root-mean-square errors for predicting the results of one psychophysical measure from the results of a different psychophysical technique are also shown. These results will be discussed more extensively later in this section. However, while we are discussing the relative validity of the two methods two points will be made. First, there is smaller error in predicting the results of paired-comparison data from the results obtained with the magnitude estimation technique or vice versa than there is in predicting either measure from most of the physical measures employed in this study. Second, when all 14 pairs are considered, typically the r.m.s. error in predicting psychophysical results from physical measures is smaller with the magnitude estimation technique than for the paired-comparisons method. If we omit the one pair thought to give atypical results in the paired-comparison method and base our computation on the remaining 13 pairs, the difference in predictability of measures resulting from the two methods is dramatically reduced.

Table V

VALUE OF EXPONENT IN POWER LAW FUNCTION $\Psi = k\phi^n$

AS OBTAINED BY LEAST SQUARE ERROR FIT TO MAGNITUDE ESTIMATION DATA

Aircraft	Exponent
D3	.43
D7	.48
D15	.43
D17	.33
B1	.44
B3	.49
B5	.44
B8	.33
B9	.40
W2	.44
W3	.44
W5	.61
W6	.36
W7	.29

Table VI

ROOT-MEAN-SQUARE ERROR IN PREDICTING PSE OBTAINED BY VARIOUS PROCEDURES FROM ONE ANOTHER AND FROM VARIOUS PHYSICAL MEASURES

	MET	MET as PC	PC (n=14)	PC (n=13)
MET	----	0.36	2.40	1.52
MET as PC	0.36	----	2.36	1.49
PC(14)	2.40	2.36	----	----
PC(13)	1.52	1.49	----	----
Max dBA	3.27	3.44	4.91	4.14
Max dBC	5.65	5.77	7.23	6.09
Max dBD_2	1.95	2.07	3.10	2.63
Max PNdB	2.00	2.16	3.39	2.79
Max $PNdB_{t1}$	2.98	2.94	2.59	2.55
Max $PNdB_{t2}$	2.15	2.19	2.38	2.15
Max PNdBM	1.93	2.07	3.15	2.65
Max $PNdBM_{t1}$	3.33	3.29	2.70	2.74
Max $PNdBM_{t2}$	2.13	2.16	2.27	2.01
Eff dBA	4.61	4.73	6.08	5.33
EFF dBC	6.32	6.43	7.79	6.89
Eff dBD_2	2.85	2.97	4.32	3.61
Eff PNdB	2.94	3.02	4.43	3.66
Eff $PNdB_{t1}$	1.89	1.94	2.95	2.46
Eff $PNdB_{t2}$	2.59	2.68	3.85	3.25
Eff PNdBM	2.68	2.77	4.12	3.41
Eff $PNdBM_{t1}$	2.02	2.04	2.72	2.37
Eff $PNdBM_{t2}$	2.41	2.49	3.54	3.01

Considering the data discussed so far there would appear to be no essential difference between psychophysical results obtained with the magnitude estimation technique and those obtained with the paired-comparison method. Although little can be said about the absolute validity of either method for predicting any specific external criterion variable we see no reason to conclude that the relative validities of the two methods differ significantly. There is some indication that results are slightly more lawful with the magnitude estimation technique than with the paired-comparison method but this is probably related to the reliability of the two methods rather than the validity per se. This matter will be discussed in the following paragraph.

The experiments described herein were not designed to exhaustively compare the relative reliabilities of the two techniques. Such a procedure would require extensive testing and retesting. However, based on rather simple assumptions we believe our data do provide at least preliminary information regarding the relative reliability of the two methods.

Consider a datum point in a paired-comparison experiment. Examination of the data would show at least two major sources of variability: 1) intra-listener variability reflected by inconsistencies in an individual's judgments from one repetition of a pair of noises to another, and, 2) inter-listener variability which reflect individual differences so that for a large number of trials involving the same stimuli we might find that a given individual preferred A to B with probability .8 while another individual preferred A to B with a probability of .6. In our comparison of the two psychophysical methods inter-listener variability was nonexistant through the employment of a common 22 subjects for judgments in both the paired-comparison method and the magnitude estimation technique. (Of course, changes in performance due to temporal factors--practice, fatigue, etc.--may introduce variability despite counterbalancing.) Intra-listener variability remains and could conceivably be different for the two psychophysical methods employed. If there were no intra-listener variability and a particular pair of stimuli were judged such that A was preferred to B 70 percent of the time, then successive repetitions of the same experiment would show that 70 percent of the listeners always preferred A to B while 30 percent of the people always preferred B to A. Intra-listener variability may be regarded as an additional random-error process and introduction of intra-listener variability in this situation would tend to depress the obtained value of .7 to some lower value with the chance value of .5 in the limit where intra-listener variability completely dominated the inter-listener variability. The limiting value would be .5 regardless of the original error-free value. Thus in comparing the two psychophysical techniques the relative slopes of the psychometric functions obtained would reflect the relative contribution of intra-listener variability. Further, as results for individual listeners are averaged over a larger and larger number of trials, the random-error process giving rise to intra-listener variability would tend to cancel out and we would expect an increase in the slope of the psychometric function as data points are based upon larger and larger numbers of judgments. Examination of figures 1 through 14 and the associated values of the slopes given in Tables II and III show that over the 14 pairs of stimuli employed in this experiment the magnitude estimation technique resulted in a steeper slope in 11 of the 14 cases while the paired-comparison method had a steeper slope in two of the cases with one pair showing the same slope for both methods. By the elementary Sign Test such a result is statistically significant approximately at the one percent level of confidence. This result is particularly compelling in that for most comparisons the number of stimulus presentations required by the magnitude estimation technique as used in these experiments was less than the number of stimulus presentations required in the paired-comparison method. For those five pairs of flights where magnitude estimation data is available both for four stimulus presentations per psychometric function and for eight stimulus presentations per psychometric function the slope of the psychometric function increased on the average though this increase is probably not statistically significant. Looking at the data over the various pairs of flights there is little or no tendency for the slope of the psychometric function to increase as the function of the number of judgments upon which the psycometric function is based. This result would be counter-intuitive if we were to consider that all flight pairs had the same slope for the psychometric function. It may be that the slope of the function is, in fact, influenced by the spectrum of the noise or by the duration of the noise. Or, even if all functions have the same slope under ideal conditions, it may be that the magnitude estimate assigned a given noise is influenced by context, e.g., by the quality or magnitude of the preceding stimulus. The experiments described herein were not designed to answer these questions; however, we can get some indication of whether the observed variations in slope result from chance fluctuations or are systematic. If the observed variations in slopes of the psychometric function are due to chance factors alone then we should expect the correlation between the slopes obtained in the magnitude estimation technique and in the paired-comparison method to be zero. If due to non-chance factors, we should then expect the correlation to be positive. The Spearman rank order correlation coefficient for the slopes in the two cases is 0.37. If we eliminate the dubious datum in the paired-comparison case referred to above, i.e., pair number 14, the correlation is 0.38. While in neither case is this correlation significant it leads us to entertain the hypothesis that slopes associated with various pairs may indeed differ in a systematic manner.

COMPARISON OF RESULTS WITH THOSE OF OTHER STUDIES

MET. In its first formulations, it was assumed that the growth in the magnitude of the subjective attribute of noisiness was the same as a function of intensity of a sound, as that usually reported for loudness --namely a 10-dB increase in intensity would result in a doubling of loudness. This is equivalent to an exponent of 0.3 in the power law function.

A number of magnitude estimation studies have been conducted to determine the scale of perceived noisiness. While perhaps the scale of perceived noisiness is such that on the average about a 10-dB change of intensity is sufficient to achieve a doubling of the perceived noisiness of the sound, there are many unexplained deviations from this value. As can be inferred from Table V, in this study we found for individual stimuli a range of from 5 to 10 dB in the increase in intensity level required for a doubling of perceived noisiness with the average stimulus requiring 7 dB. In a comparison study (ref. 5) employing five aircraft noises, we found a range of from 8 to 12 dB with a mean of 10 dB. The Boeing study (ref. 1) reported an average of 13 dB, as did Broadbent and Robinson (ref. 6). Both used aircraft noises and the magnitude estimation technique. Ollerhead (ref. 7) using a method of adjustment and random pink noise found results in the range of 7 to 10 dB. With bands of noise and a method of adjustment, Parnell, et al (ref. 8) found results ranging from 8.5 dB to 14.3 dB. The same investigators, with the magnitude estimation technique, obtained results ranging from 14 to 27 dB. However, as Parnell, et al note, this scale value is not critical to the rank ordering of the relative perceived noisiness of sounds on the basis of a unit of PNL; on the other hand, this scale value is of obvious importance in the interpretation of how much subjective value is to be ascribed to a given change in the intensity of a noise. Clearly, the above results, and many more that could be cited, are not all estimates of the same thing; they are not comparable and cannot simply be averaged to obtain a best estimate of the "true" scale value. The differences in estimates are sometimes dramatic and must be explained on a rational basis before we can conclude that we understand the manner in which perceived noisiness grows as a function of noise intensity.

P.C. Results. The majority of the aircraft noises used in the various experiments of the present study were recorded in the field at the same time subjects located in the field were making subjective judgments of the noisiness of the flyover sounds. One set of the noises was recorded but not judged in the field; however, these noises were later presented via loudspeakers to listeners in an anechoic chamber for purposes of judging their noisiness.

The results of these various tests provide, then, a means of evaluating the degree of consistency to be found between judgments made under live and more realistic field conditions and those made in the laboratory from recordings of the same noises.

It might be noted here that although field tests are generally difficult to arrange and subject to a number of uncontrollable factors in the environment, recordings of aircraft noises that are suitable for use in the laboratory must be made with great care. Indeed, it was not possible to make as many direct comparative tests of noises judged in the field as we wished because most of the recordings made during the field tests were not adequate in all respects. However, as shown in Table VII, rank ordering of PNL and EPNL predicted perceived noisiness is very similar for both the field and the laboratory studies.

In addition to a comparison of field and laboratory tests, these data provide a fairly broad basis for evaluating the accuracy and consistency of the relation between noisiness as judged and as predicted by various physical units. This evaluation is also to be found in Table VII where in addition to the data directly related to the present study we have included data for aircraft noise as obtained in the laboratory by Pearsons and Bennett (ref. 1). Included in Table VII are those physical units which have been most commonly proposed as means of predicting with some accuracy the perceived noisiness of aircraft sounds. There are available from the literature, judgment data from a number of studies of the perceived noisiness of aircraft sounds for which, unfortunately, the physical data regarding the sounds judged are not available in terms of all the units reported in Table VII (in particular for EPNdBM or $EdBD_2$) and for this reason, are not included.

Overall, these data indicate that definitely EdBA and possibly EPNdB are inferior in predicting the judged noisiness of these aircraft noises; the difference in prediction accuracy is rather small and inconsistent among the other units, although $EPNdBM_{t1}$ and $EPNdB_{t1}$ do stand out as the most accurate. It can also be observed that there do not appear to be any large differences between the data from the field vs. those from the laboratory, or paired-comparison vs. magnitude estimation test data.

Perhaps the techniques for treating the physical data that are the most controversial are those of (1) pure-tone corrections, and (2) the integration of successive .5-sec PNLs to achieve the effective or EPNL value. The present tests were not designed to evaluate those specific procedures nor were some of the other studies reported in Table VII. For example, whether or not EPNL does better than Max PNL usually depends

Table VII

RESULTS OF SOME JUDGMENT TESTS OF AIRCRAFT NOISE

CONDUCTED IN THE FIELD AND UNDER VARIOUS LISTENING CONDITIONS IN THE LABORATORY
Data are the RMS errors between judged equal noisiness and noisiness as predicted by various physical measures. P.C. = Paired Comparison, M.E.T. = Magnitude Estimation Test

			Present Study				
	Field(ref. 3) (N = 18)	P.C. Lab (**) Anechoic (N = 12)	P.C.(ref. 9) Lab Anechoic (N = 20)	P.C. Lab Anechoic (N = 13)	M.E.T. Lab Anechoic (N = 14)	Average	Rank order of average
$EPNdBM_{t1}$	3.2	2.9	2.4	2.4	2.0	2.6	1.5
$EPNdBM_{t2}$	3.2	2.8	3.0	3.0	2.5	2.9	3.0
EPNdBM	3.1	3.7	2.5	3.4	2.8	3.1	4.5
$EPNdB_{t1}$	3.8	2.7	2.2	2.5	1.9	2.6	1.5
$EPNdB_{t2}$*	4.0	2.8	3.3	3.2	2.7	3.2	6.0
EPNdB	3.9	3.8	3.5	3.7	3.0	3.6	7.0
$EdBD_2$	3.0	3.4	2.4	3.6	3.0	3.1	4.5
EdBA	4.6	5.4	4.3	5.3	4.7	4.9	8.0

* Unit prescribed for FAA noise certification (ref. 4)

** Previously unreported data from this laboratory upon whether the noises being judged differ significantly with respect to duration; if they do not then it can be expected that Max PNL will perform as well as EPNL, which is indeed the case in this study (see Table VI).

CONCLUSIONS

1. Both the method of paired comparisons [as used to collect data reported by the McDonnell Douglas Corporation (ref. 2)] and the magnitude estimation technique [as used to collect data reported by The Boeing Company (ref. 1)] provide estimates of the point of subjective equality which differ little from one another. On this basis, at least, there is no indication that one method provides more valid results than the other.

2. Both methods are capable of providing highly reliable data but the magnitude estimation technique appears to be more efficient in this respect. Based on the measures employed in this study if the two methods are to provide equally reliable data, the method of paired comparisons requires approximately twice as much testing time as does the magnitude estimation technique. This result applies in the case where matched pairs of flights are to be evaluated. In the case where N stimuli are to be scaled on a common scale, the relative efficiency of the magnitude estimation technique should be greater.

3. In the magnitude estimation tests of the present study, the accuracy with which the most accurate physical units of PNL and EPNL predicted the subjective results was slightly lower (r.m.s. error of about 2.0 dB) than found in present and previous paired-comparison tests (2.2 to 3.8 dB) with the same or similar noises.

4. Although no formal evaluation was attempted, spontaneous remarks by many listeners indicated that they felt little confidence in their magnitude estimates (despite the fact that the data shows they made such judgments reliably) and found the paired-comparison task more natural. This would suggest that special care be taken with instructions to elicit the full cooperation of the subjects when the magnitude estimation technique is employed.

5. The scales of perceived noisiness derived from the magnitude estimation tests in this study and others cited herein indicate that the scale of 10 dB per-doubling-of-noisiness presently used in the calculation of PNL is not well documented. Results have varied over

a wide range. However, more data on this question is in order before a change in present PNL calculation procedures is contemplated.

REFERENCES

1. The Boeing Company, "Study and Development of Turbofan Nacelle Modifications to Minimize Fan-Compressor Noise Radiation," Vol. VII, Subjective Evaluation Tests. NASA CR-1717, 1971, prepared by The Boeing Company under Contract No. NAS1-7129 for NASA Langley Research Center.

2. L. E. Langdon, R. F. Gabriel, and A. H. Marsh, "Investigation of DC-8 Nacelle Modifications to Reduce Fan-Compressor Noise in Airport Communities," Part VI - Psychoacoustic Evaluation. NASA CR-1710, 1970, prepared by the McDonnell Douglas Corporation under Contract No. NAS1-7130 for NASA Langley Research Center.

3. K. D. Kryter, "Possible Modifications to the Calculation of Perceived Noisiness." NASA CR-1636, 1970, prepared under Contract No. NAS1-6885 by Stanford Research Institute for NASA, Washington, D.C.

4. K. D. Kryter, "The Effects of Noise on Man." Academic Press, New York, New York (1970).

5. F. R. Clarke, and K. D. Kryter, "Perceived Nosiness Under Anechoic, Semi-Reverberant and Earphone Listening Conditions." NASA CR-2108, 1972, prepared under Contract No. NAS1-10017 by Stanford Research Institute for NASA Langley Research Center, Hampton, Virginia.

6. D. E. Broadbent, and D. W. Robinson, "Subjective Measurements of the Relative Annoyance of Simulated Sonic Bangs and Aircraft Noise." J. Sound Vib., Vol. 1, No. 2, pp. 162-174 (1964).

7. J. B. Ollerhead, "The Noisiness of Diffuse Sound Fields at High Intensities." Final Report FAA-NO-70-3, 1970, prepared under Contract No. FA68WA-1971 by Wyle Laboratories for the Federal Aviation Administration, Washington, D.C.

8. J. E. Parnell, D. C. Nagel, and H. J. Parry, "Growth of Noisiness for Tones and Bands of Noise at Different Frequencies." Technical Report FAA-DS-67-21, 1967, prepared under Contract No. FA65WA-1180 by Bolt Beranek and Newman Inc. for the Federal Aviation Administration, Washington, D.C.

9. K. S. Pearsons, and R. L. Bennett, "The Effects of Temporal and Spectral Combinations on the Judged Noisiness of Aircraft Sounds." Final Report FAA-NO-69-3, 1969, prepared under Contract No. FA68WA-1978 by Bolt Beranek and Newman Inc. for the Federal Aviation Administration, Washington, D.C.

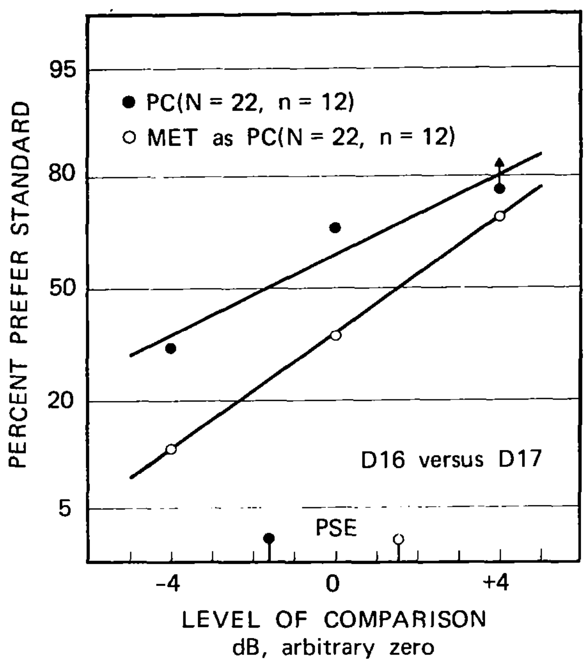

FIGURE 1*

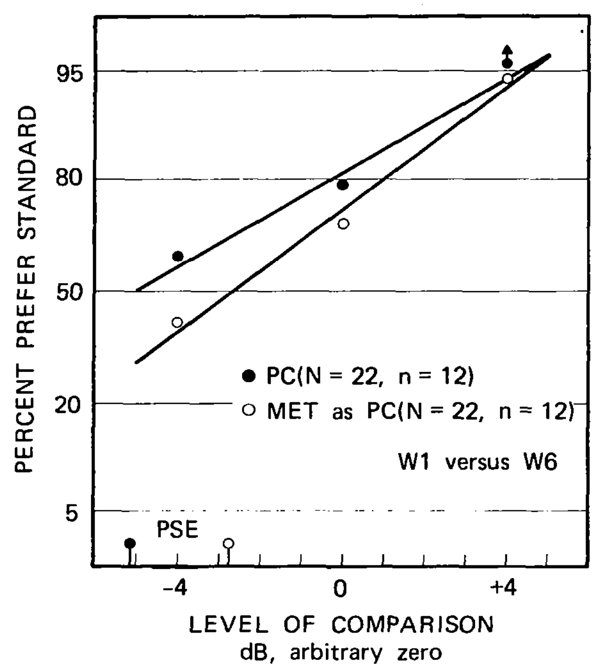

FIGURE 2*

*Psychometric functions showing the percent of listeners preferring the standard stimulus as a function of the intensity level of the comparison stimulus for each of the 14 pairs of aircraft noises. Solid points show data obtained by the method of paired-comparisons. Open circles show data obtained by the magnitude estimation technique after conversion to paired-comparison format as described in the test. The number of listeners is indicated by N, while the number of individual flights judged for each psychometric function is indicated by n. Points of subjective equality are shown on the abscissa. Arrows on points indicate that the point was obtained by averaging proportions rather than z-scores. Were z-scores to have been averaged for these points the average would have been either + or − infinity as indicated by the direction of the arrow.

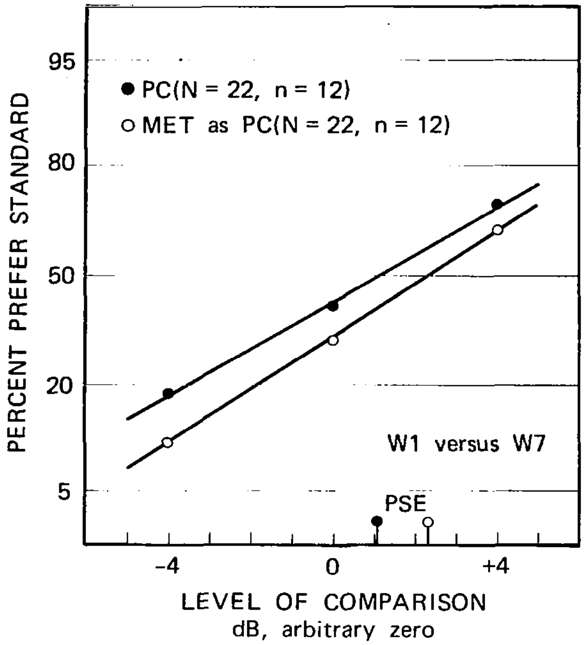

FIGURE 3*

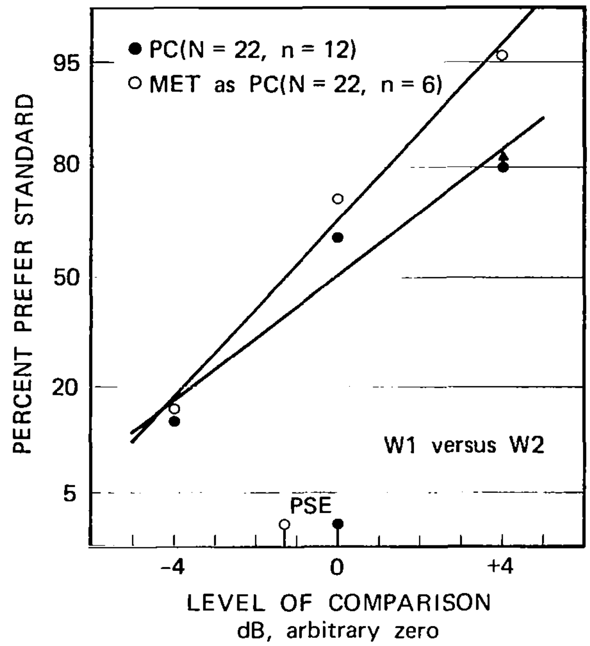

FIGURE 4*

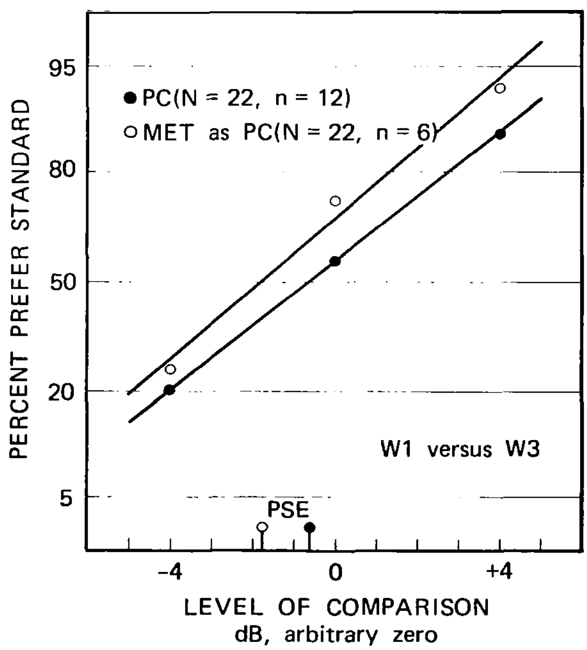

FIGURE 5*

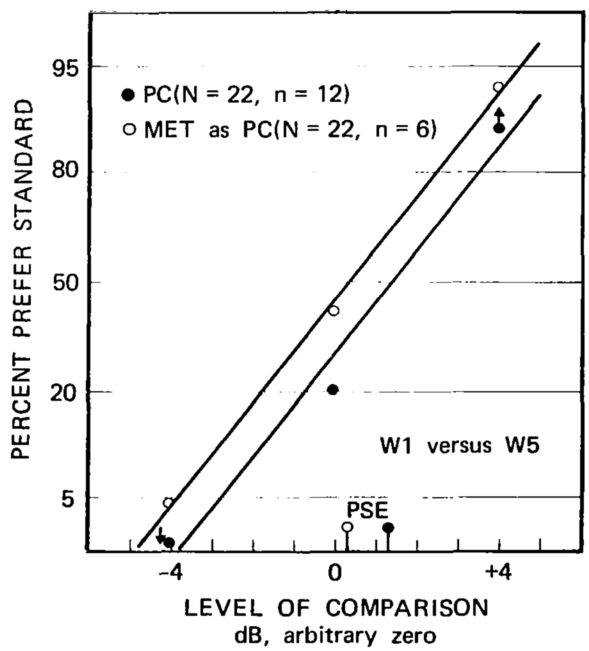

FIGURE 6*

*See page 24

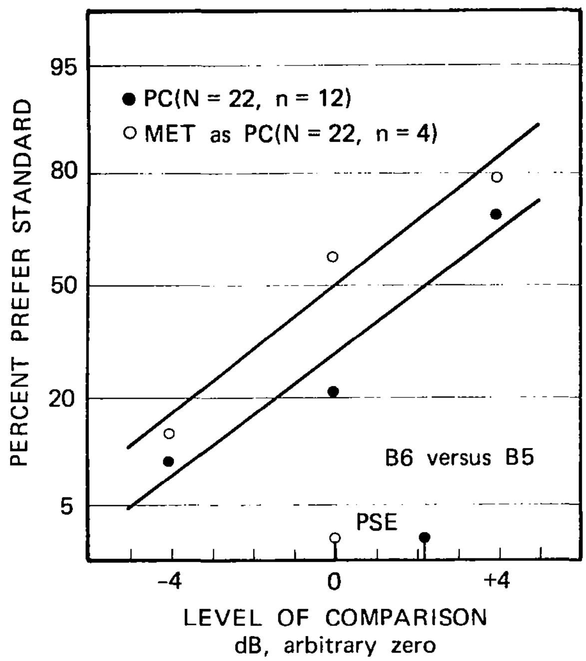

FIGURE 7*

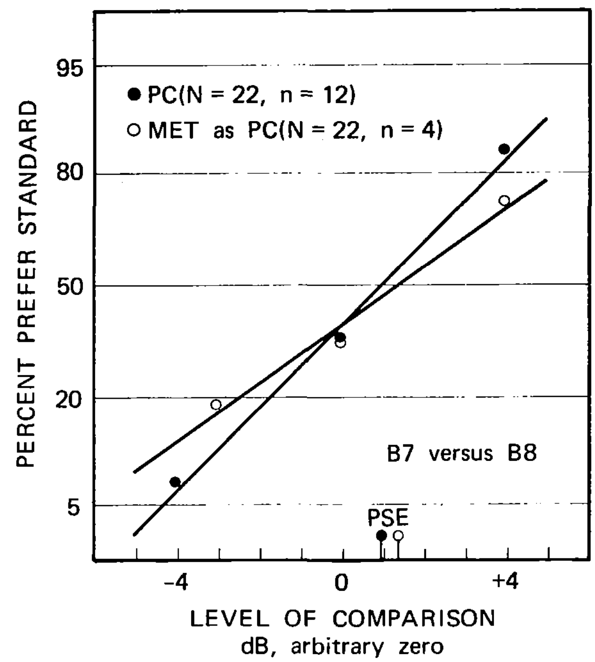

FIGURE 8*

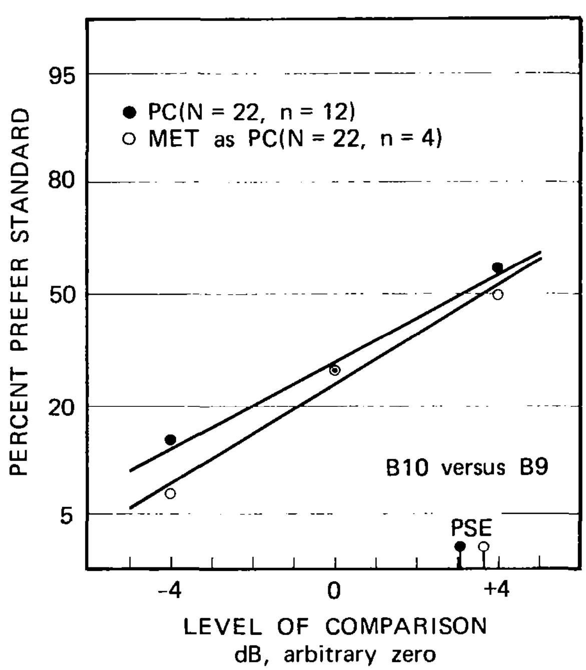

FIGURE 9*

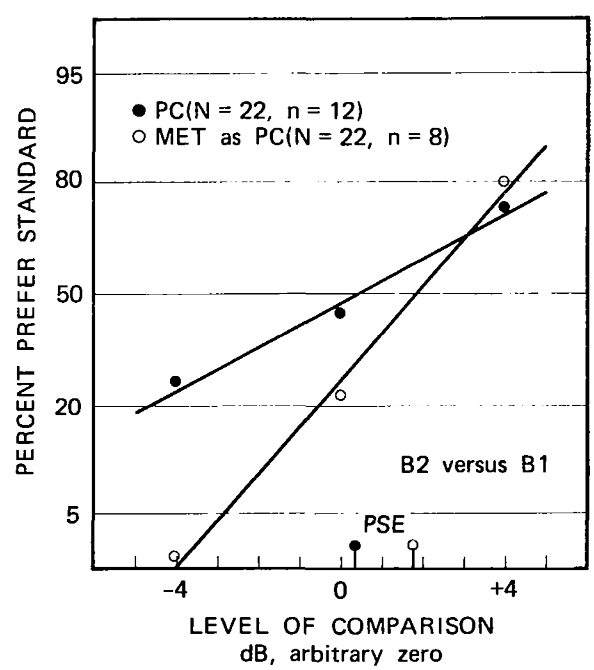

FIGURE 10*

*See page 24

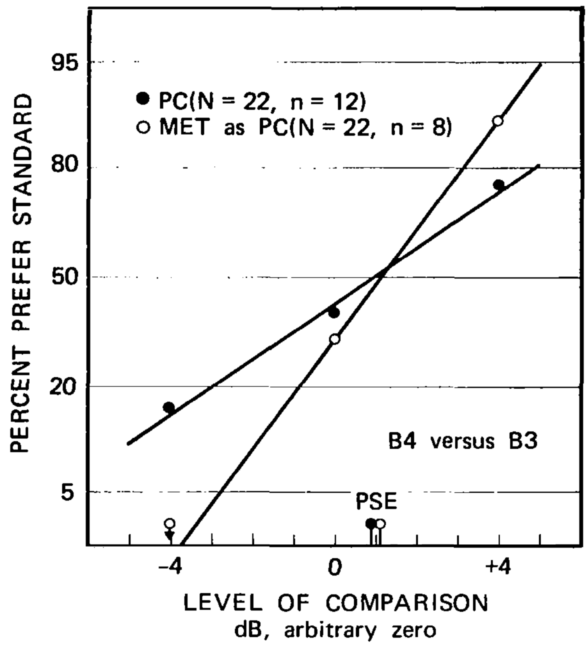

FIGURE 11*

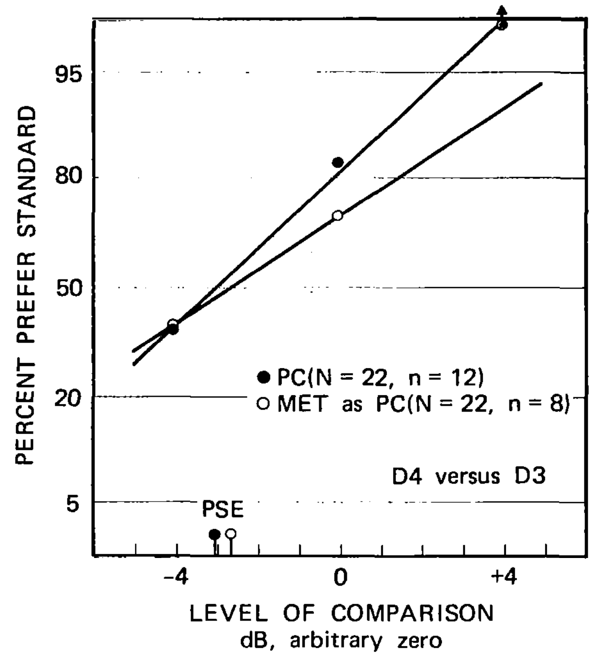

FIGURE 12*

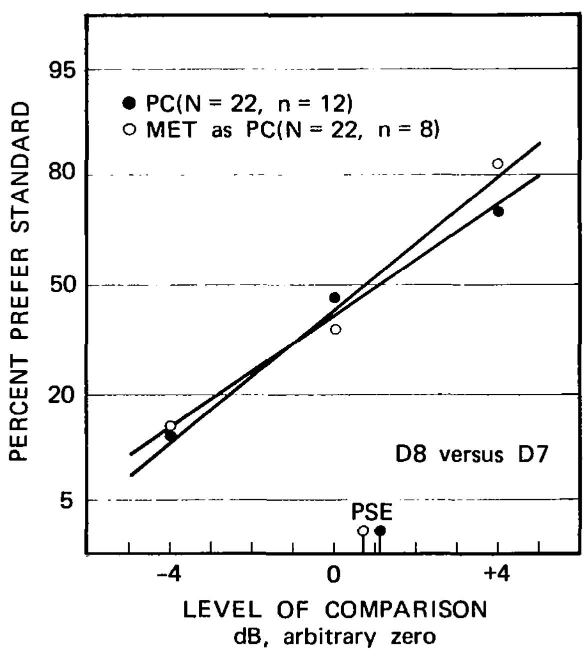

FIGURE 13*

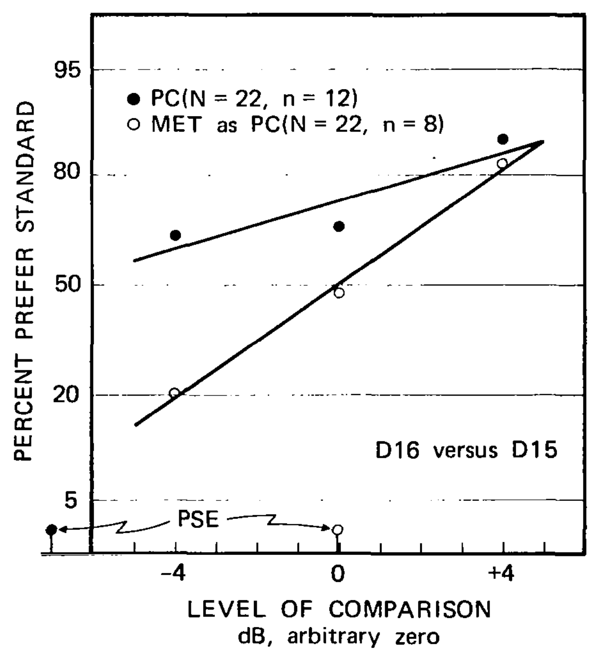

FIGURE 14*

*See page 24

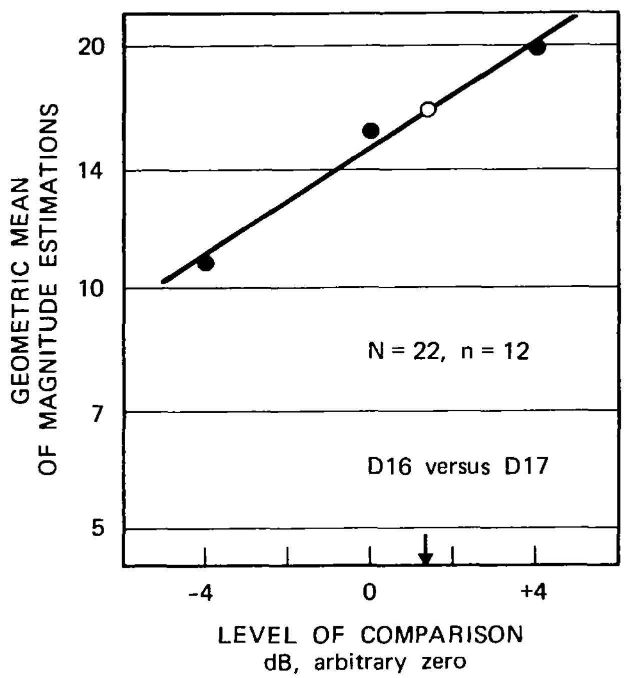

FIGURE 15*

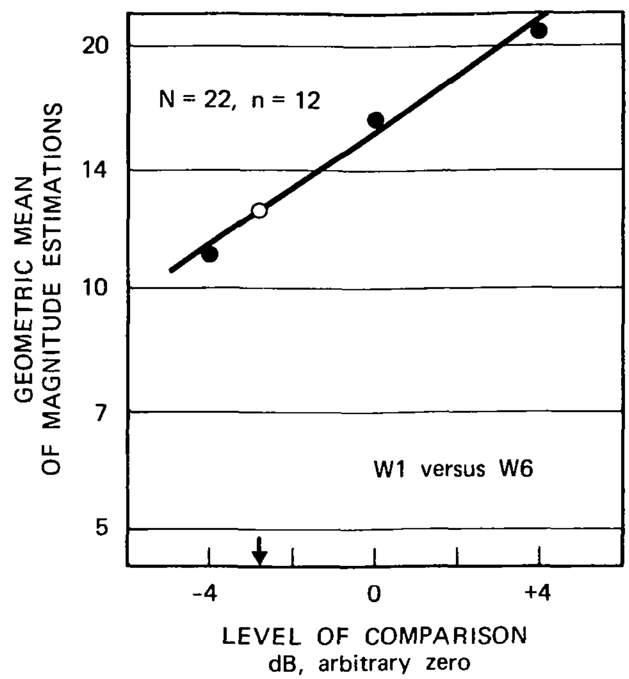

FIGURE 16*

*Power law functions showing the geometric mean of magnitude estimates of comparison stimuli as a function of the intensity level of the comparison stimulus. The solid line function provides a least square error fit to the solid data points. The open circle is the geometric mean of the magnitude estimate for the appropriate standard stimulus and is always plotted on the power law function to determine the point of subjective equality. The value of this point is also indicated on the abscissa.

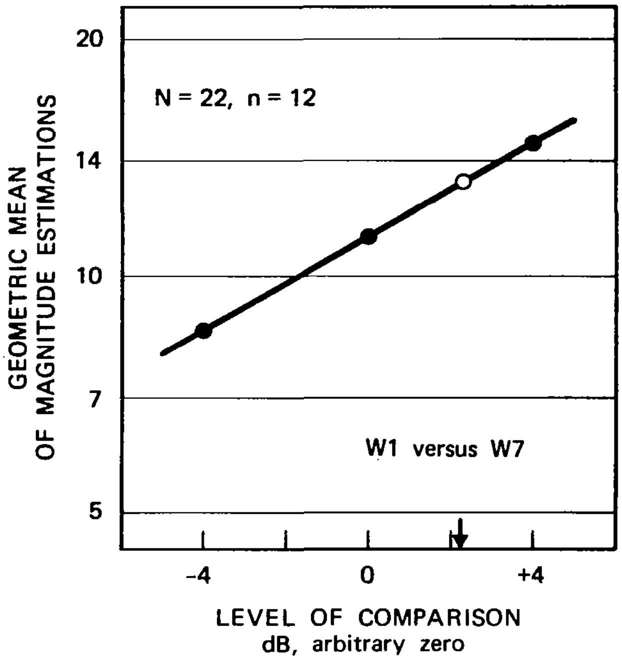

FIGURE 17*

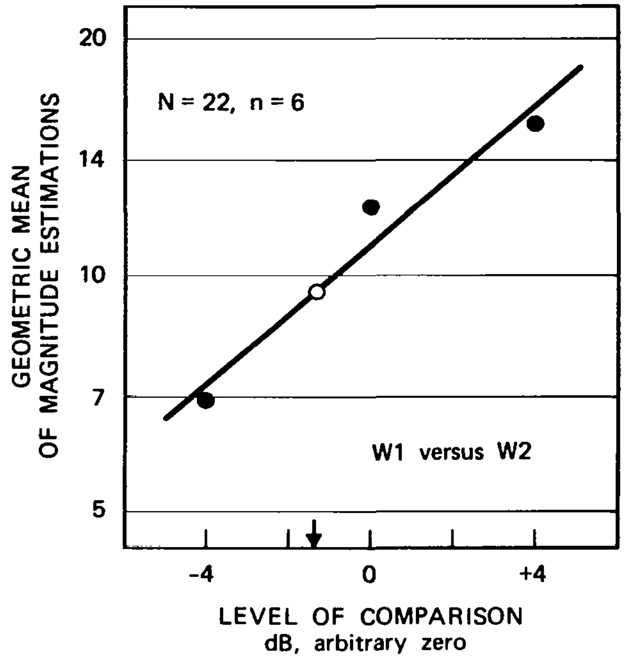

FIGURE 18*

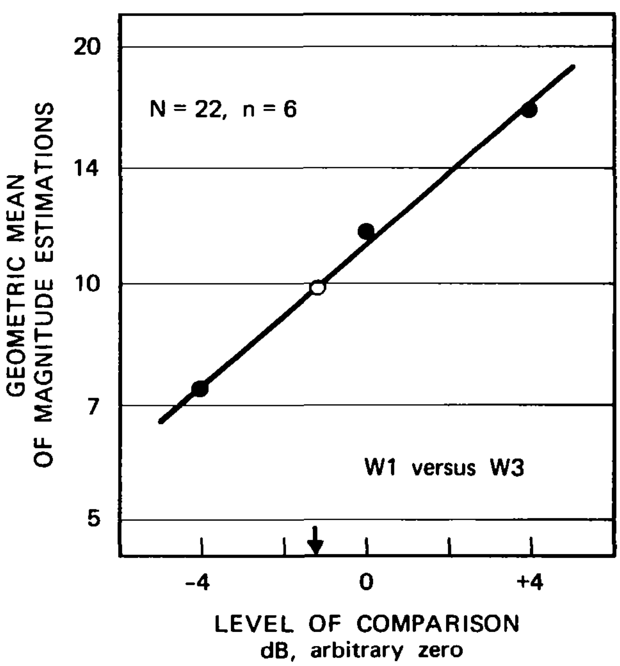

FIGURE 19*

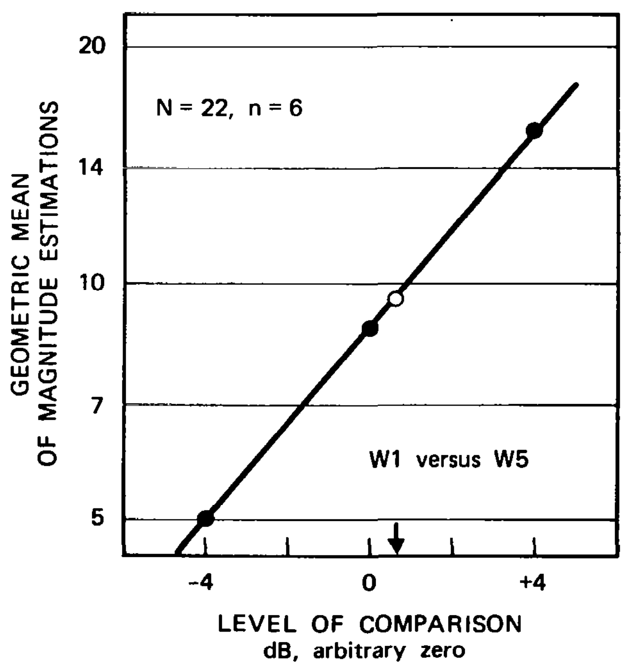

FIGURE 20*

*See page 28

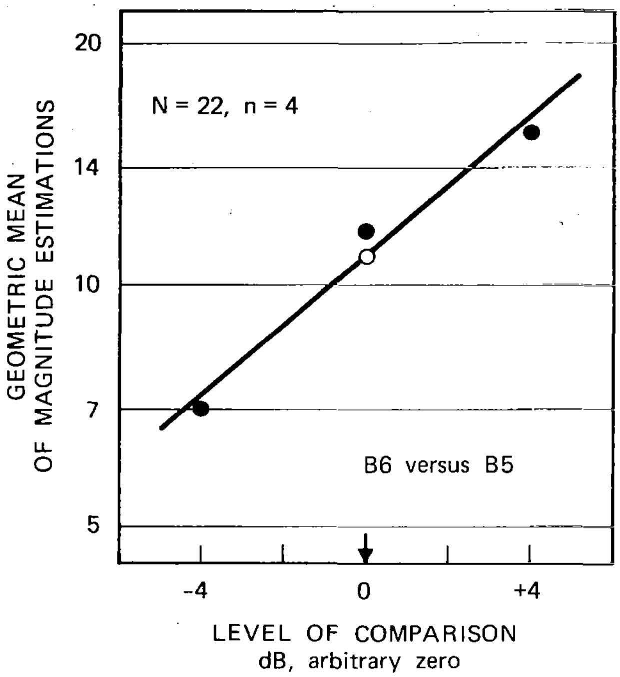

FIGURE 21*

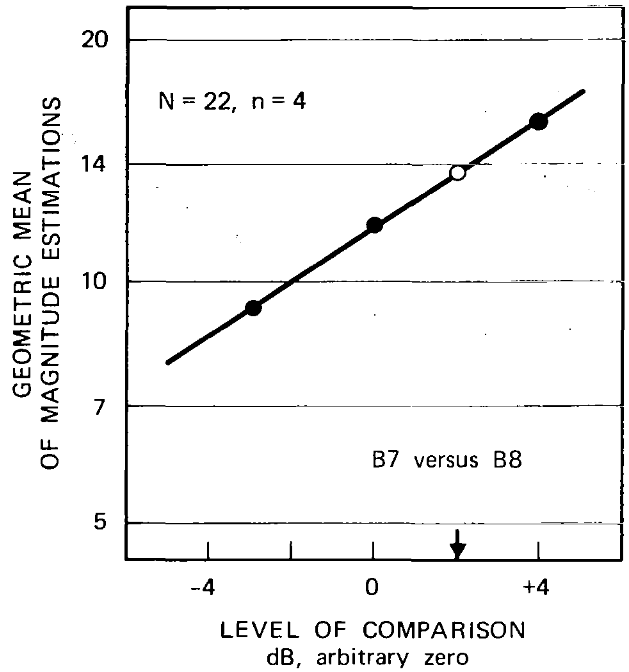

FIGURE 22*

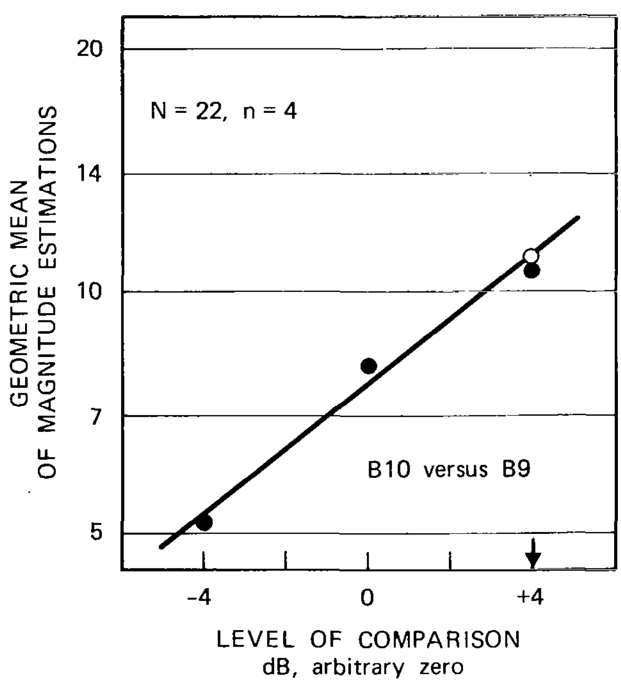

FIGURE 23*

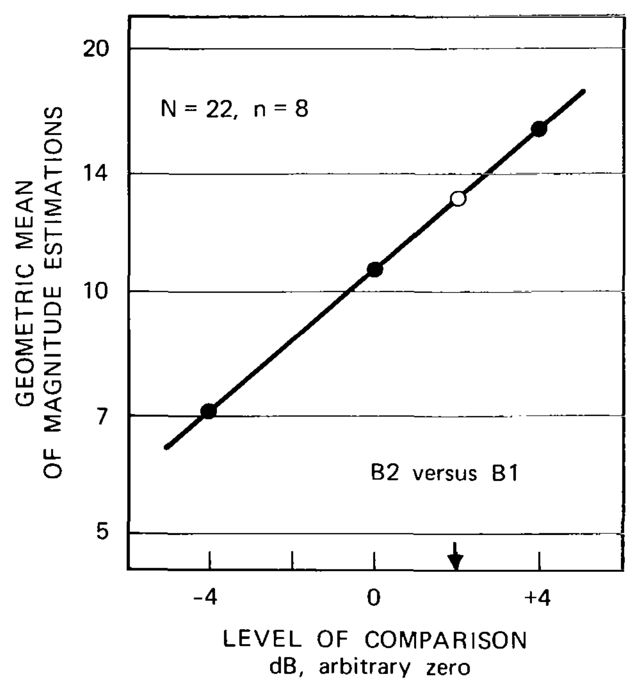

FIGURE 24*

*See page 28

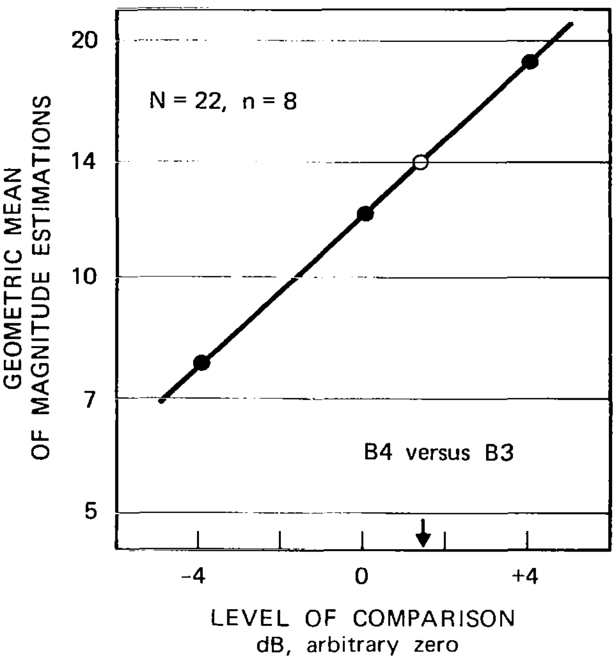

FIGURE 25*

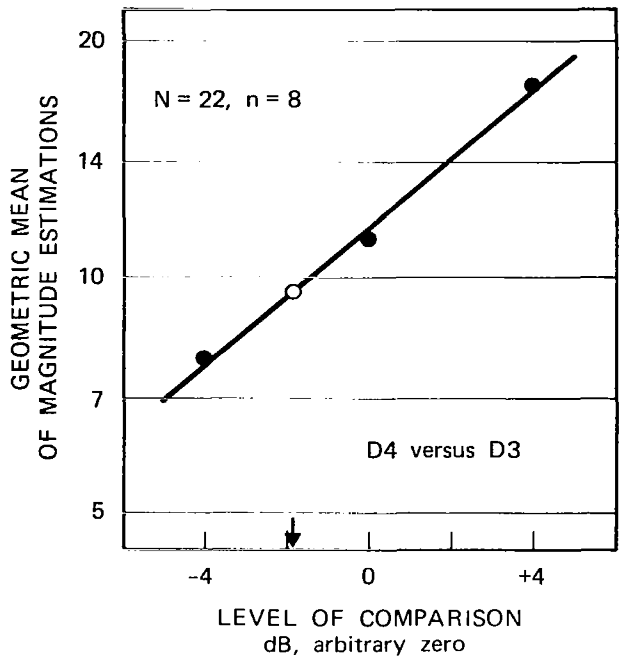

FIGURE 26*

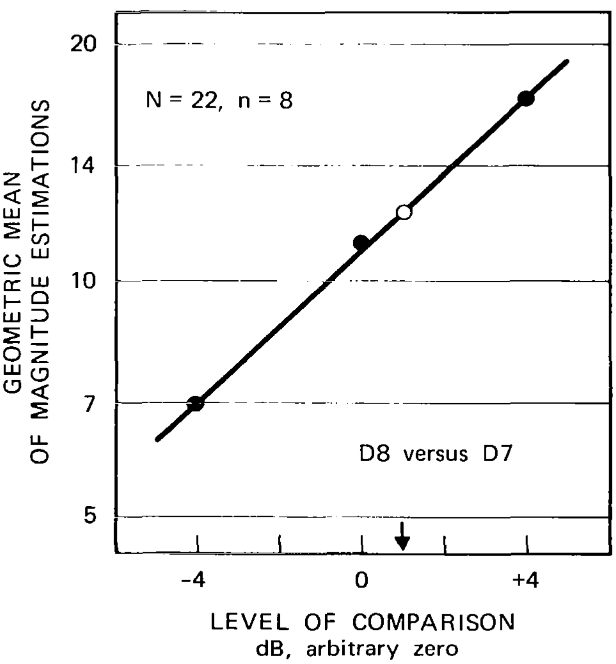

FIGURE 27*

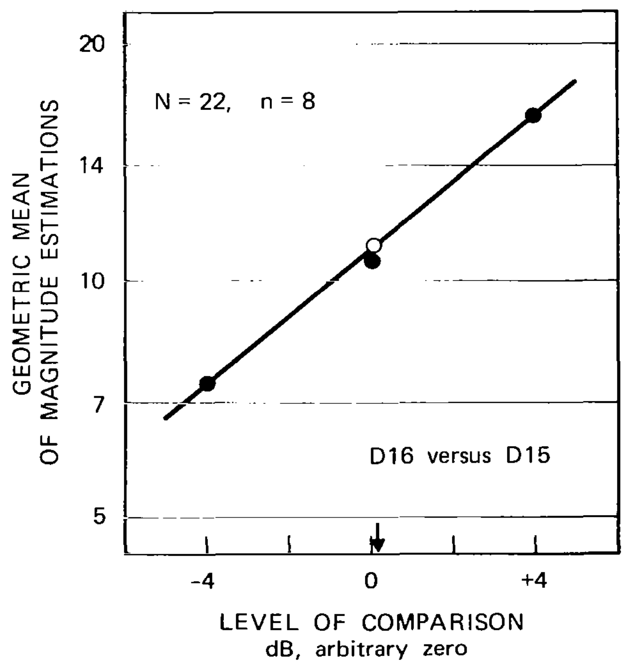

FIGURE 28*

*See page 28

Appendix A

INSTRUCTION TO LISTENERS

NAME	GROUP	TAPE	DATE

Circle A if first sound is more acceptable.
Circle B if second sound is more acceptable.

INSTRUCTIONS:

The primary purpose of the tests being conducted is to determine, if possible, how people feel about the relative acceptability of one type or level of aircraft noise when compared with a second type or level of aircraft noise.

You will hear a series of sounds from aircraft. The sounds will occur in "pairs" and your task is to judge which sound in each pair you think would be more acceptable to you if heard in or near your home during the day and/or evening when you are engaged in typical, awake activities.

After you have heard each pair of sounds, please quickly decide which of the two you feel would be more acceptable to you. If you think the second sound of a pair would be more acceptable, circle B for that particular pair. If you think the first sound in the pair would be more acceptable to you than the second, circle A.

Please concentrate on the judgment at hand and give an answer even though the two sounds may seem approximately equal in acceptability to you. If you feel that there is absolutely no real difference in terms of acceptability of the two sounds, please circle either A or B, giving the best guess you can, and put a question mark after that pair.

There are no "right" or "wrong" answers, nor do we expect people to agree with each other. We are interested in how you feel about the sounds and how people differ in their judgments of the acceptability of these aircraft sounds.

An announcement of the item number will be made before each pair of sounds is to occur. The sounds of a pair will be separated by a few seconds. During the test period, which will be approximately 15 minutes, please remain quiet and attentive. Give us your best judgment and imagine, if you will, that you are listening to these sounds in or near your own home.

1.	A	B
2.	A	B
3.	A	B
4.	A	B
5.	A	B
6.	A	B
7.	A	B
8.	A	B
9.	A	B
10.	A	B
11.	A	B
12.	A	B
13.	A	B
14.	A	B
15.	A	B
16.	A	B
17.	A	B
18.	A	B
19.	A	B
20.	A	B

ANSWER SHEET

Name ______________________________ Date ____________________

Sex ________ Age ______ Session _________ Listening Position ____________

INSTRUCTIONS

We are asking you to help us solve a problem concerned with noise: How annoying or disturbing are various kinds of sound when heard in your home? You will be asked to give a score to each sound.

First, we will produce a sound whose noisiness score is 10. Use that sound as a standard, and judge each succeeding sound in relation to that standard. For example, if a sound seems <u>twice</u> as noisy as the standard, you will write 20 in the appropriate box on the answer sheet. If it seems only <u>one-quarter</u> as noisy, write 2.5. If it seems <u>three</u> times as noisy, write 30, and so on.

Please try to judge each sound carefully, and give it a score that tells how strong the annoyance seems to you. There are no right or wrong answers. The important thing is to say how you rate each of the sounds.

1. ______	11. ______	21. ______	31. ______
2. ______	12. ______	22. ______	32. ______
3. ______	13. ______	23. ______	33. ______
4. ______	14. ______	24. ______	34. ______
5. ______	15. ______	25. ______	35. ______
6. ______	16. ______	26. ______	36. ______
7. ______	17. ______	27. ______	37. ______
8. ______	18. ______	28. ______	38. ______
9. ______	19. ______	29. ______	39. ______
10. ______	20. ______	30. ______	40. ______