DOCUMENT RESUME

ED 041 459 95 EM 008 130

AUTHOR Donaldson, Wayne; Glathe, Herta
TITLE Signal Detection Analysis of Recall and Recognition Memory.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C. National Center for Educational Research and Development.
REPORT NO WP-48
PUB DATE Jul 69
NOTE 32p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.70
DESCRIPTORS Discriminant Analysis, Memory, Paired Associate Learning, *Recall (Psychological), *Recognition

ABSTRACT

Three paired-associate learning studies were run to compare signal detection analysis of recall and recognition memory performance. Experiment I showed that (a) rates of recall and recognition discriminability are substantially different in later trials and (b) a previously suggested correction for guessing does not transform the data to theoretical expectations. Experiment II showed that subjects' guessing rates change systematically over trials and further supported the inappropriateness of a guessing correction. Experiment III attempted to hold constant the probability of guessing correctly. It was suggested that for the purposes of comparing recall and recognition, the most profitable transformation of the recall data is in the nature of the "one-of-M-orthogonal" signals paradigm. A reference list is appended. (Author)

UNIVERSITY OF PITTSBURGH - LEARNING R & D CENTER

WORKING PAPER 48

SIGNAL DETECTION ANALYSIS OF RECALL AND RECOGNITION MEMORY

WAYNE DONALDSON AND HERTA GLATHE

SIGNAL DETECTION ANALYSIS OF RECALL AND RECOGNITION MEMORY

Wayne Donaldson

and

Herta Glathe

Learning Research and Development Center

University of Pittsburgh

July 1969

Published by the Learning Research and Development Center supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

SIGNAL DETECTION ANALYSIS OF RECALL AND RECOGNITION MEMORY

Wayne Donaldson and Herta Glathe

Learning Research and Development Center

University of Pittsburgh

Abstract

Three paired-associate learning studies were run to compare signal detection analysis of recall and recognition memory performance. Experiment I showed that (a) recall and recognition <u>d</u>'s are substantially different in later trials and (b) a previously suggested correction for guessing does not transform the data to theoretical expectations. Experiment II showed that <u>S</u>'s guessing rates change systematically over trials and further supported the inappropriateness of a guessing correction. Experiment III attempted to hold constant the probability of guessing correctly. It was suggested that for purposes of comparing recognition and recall, an ROC analysis of recall data is inappropriate and that a better approach is the use of the "forced-choice" or "one-of-M-orthogonal" signals model. Finally, a possible interpretation of a recall ROC <u>d</u>' is suggested.

SIGNAL DETECTION ANALYSIS OF RECALL AND RECOGNITION MEMORY[1]

Wayne Donaldson[2] and Herta Glathe

Learning Research and Development Center

University of Pittsburgh

It has come to be almost standard procedure among many researchers to analyze recognition memory data in terms of the signal detection model (TSD) (Donaldson & Glathe, in press; Donaldson & Murdock, 1968; Egan, 1958; Murdock, 1965; Norman & Wickelgren, 1965; Parks, 1966). Attempts to conceptualize and analyze recall processes in similar terms have lagged behind. For at least two reasons this is not surprising. First, TSD was developed in a context (psychophysical detection) that is methodologically similar enough to recognition memory procedures (Donaldson & Murdock, 1968) to make adoption fairly easy. Second, in experiments where TSD analysis has been applied to recall, the interpretation of the data is at best unclear (Murdock, 1966; Bernbach, 1967). Bernbach (1967) has developed a signal-detection-oriented, finite state model of memory which appears to be capable of handling both recall and recognition memory data. Briefly, a $\underline{d}^*$ value is hypothesized which is to be interpreted as the maximum discriminability between underlying new item and old item distributions and which is not affected by memory factors. This degree of discriminability is presumably what is being measured in recall studies, which show the TSD sensitivity measure (d')[3] to be unaffected by typical memory variables such as number of interpolated items (Murdock, 1966) or number of trials (Bernbach, 1967). In recognition studies, the d' measure, according to Bernbach's model, reflects some changing fraction of d*, that fraction being (1- δ) where δ is the forgetting parameter. As such, d' from recognition procedures is directly affected by the independent variables which typically are shown to affect memory. Obviously this is not to say that forgetting occurs only in recognition tasks, but rather, that in recall tasks, d' simply does not measure forgetting. A complete description, including the mathematical development, can be found in Bernbach's (1967) paper.

At a first level of analysis, two strong predictions can be derived from Bernbach's model. First, as previously mentioned and supported, d' from recall studies should be independent of memory manipulations. Second, under comparable conditions, d' from recall procedures should always be higher than d' from recognition procedures, with the recognition d' limited from above but increasing towards the recall d' = d* as δ approaches zero.

Before proceeding, however, a further consideration has to be mentioned. In one of the few papers that specifically considers the nature of the two types of tasks, or more specifically, the two types of analysis applied to the tasks, Clarke, Birdsall, and Tanner (1959) have argued for a mathematical difference between the two procedures. Their mathematical development indicates that, for comparable situations, a Type II analysis (response conditional--as applied to recall data) will yield a lower d' than a Type I analysis (stimulus conditional--used in analyzing recognition data). This consideration must then alter the predictions in such a way as to permit the recognition d' to be higher than the recall d' (d*) measure.

The initially stated strong predictions are now no longer clear. In attempting to integrate these two positions it is difficult to ascertain whether the recall d' is predicted to be lower than the maximum value (d*) or the recognition d'. In other words, must the recall d' always fall below that for recognition? Or is the lower recall d' relative to the maximum value that is obtained from the recognition data (d*)?

The present series of studies was designed to examine the TSD analysis of recall and recognition procedures under comparable experimental conditions.

Experiment I

Method

Female undergraduates enrolled in Introductory Psychology at the University of Pittsburgh served as Ss in order to fulfill course requirements. Each S was tested for 12 study-test trials on a paired-associate list (PAL). Each of the A members of a 20-pair list was a single consonant letter. The B members were single digits from 0 to 9, each digit being paired with two different consonant letters. Lists were filmed and presented by means of a Dunning animatic projector at a study phase rate of 1 pair/second and a test phase rate of 5 seconds/test item. The test procedure was either recognition or recall. In the recognition test phase, Ss saw 20 letter-number pairs and for each pair were required to indicate their confidence as to whether the pairing was correct or incorrect. This was done by placing a short vertical line through a 10 cm. horizontal line labelled "positive-incorrect" on the left end and "positive-correct" on the right end. The line was subsequently divided into 10 equal intervals for purposes of analysis. Half of the pairs presented in each test phase were correct pairings, half were incorrect, and Ss were aware of this characteristic of the test lists. For the recall test phase, Ss were presented the single letters one at a time and were required to recall the number that they thought was paired with the letter (omissions were not permitted) and then to rate their confidence by placing a mark somewhere on a 10 cm. line labelled "guess" at the left end to "positive-correct" at the right end. The presentation order of study and test items was changed randomly for every phase, the only restrictions being that pairings occurring in the last four presentation positions were not tested in the first four test positions. Eighty Ss were randomly assigned to four groups of 20 each. One group received all 12 trials under recognition conditions, a second group received 12 recall trials, a third group received 5 trials of recall followed by 5 trials of recognition and concluded with 2 recall trials, and the fourth group received 5 recognition trials, 5 recall trials and then 2 final recognition trials.

Results

The confidence rating scale was divided into 10 equal segments and the tabulated confidence rating data were analyzed by the Ogilvie and Creelman (1968) EPCROC program. The d' sensitivity measure, as calculated from the point of intersection of the ROC with the negative diagonal is shown in Figure 1, for the four conditions as a function of trials. The

---

Insert Figure 1 about here

---

data for trials 1 through 5 have been combined for the two groups who started under recognition procedures and for the two groups who started under recall procedures, no major differences between the groups being apparent. As expected, recognition d' increased over trials. Recall d' started at the same level as recognition, followed it closely over at least the first four trials, and then proceeded to level off much more quickly than did recognition. Also, recall d' following recognition trials was consistently lower than d' in the all-recall condition whereas recognition measures following recall trials remained at the same level as those following recognition trials. Figure 1 also shows the d' measure for the group that received all 12 trials under recall after Bernbach's suggested correction for guessing is applied, where g is taken as 1/N, N being the number of alternatives, in this case, 10. According to Bernbach this correction is necessary to eliminate those items which are correctly recalled but are done so by chance and hence represent samples from the noise or null state distribution rather than from the signal or memory state distribution. The correction failed to eliminate either the substantial increase from trial 1 to trial 2 or the smaller increase over the last half of the trials. Only for trials 2 through 6 did the correction yield data that do not increase over trials.

In general, the ROC curves appeared to be well fit by straight lines when plotted on double probability graph paper. Of the 58 comparisons (12 trials by 4 conditions plus the combined data for recall and recognition over trials 1 through 5) 3 of the ROC curves provided a X^2 significant at beyond the .01 level. Concerning the slope of the ROCs, none of the 29 recognition ROCs deviated reliably from a slope of one. In recall conditions, the slope of the ROC was significantly below one on early trials, and gradually increased until by trial 6 and thereafter, it was not reliably different from unity.

Discussion

In general the data do not conform well to expectations. The "corrected-for-guessing" recall data, with the exception of the increase from trial 1 to trial 2, replicated Bernbach's findings over the first six trials (the limit of his data). However, over the final six trials even the corrected recall *d'* tended to increase. That the recall measures are lower than those for recognition in the later trials would be expected from the Clarke, Birdsall, and Tanner considerations. A point of concern with the Clarke, et al. formulation based on exponential distributions, however, is their prediction that the slope of a Type II ROC should be inversely related to recall *d'*. The present data indicate that both the uncorrected recall *d'* and the slope of the ROC increase as a function of trials. More extensive discussion of the data of Experiment I will be postponed until other considerations and more data have been presented.

There are two points of concern in the present data. First, the assumption that the guessing rate remains constant over trials is a questionable one in this situation. It implies that S is not able to acquire any information other than the correct pairing, i.e., that he cannot acquire negative information as to which alternatives are not correct and thereby reduce the number of alternatives from which he is selecting a response (Murdock, 1963). The second, and more serious, point of concern is the use of a high threshold guessing correction in the context of a technique that has cast serious doubt on the validity of the model underlying that very correction (Swets, 1961). This second point will be considered later. The following two studies were addressed to the first point. Experiment II constituted an attempt to demonstrate and then measure a changing guessing rate. Experiment III employed a situation in which the guessing rate presumably could not change.

Experiment II

Method

This study was actually carried out prior to the first one while equipment and films were being prepared. Forty-five female Ss were run for 10 study-test trials on a PAL recall task. The procedure was identical to that for the group in Experiment I that received only recall trials except that materials were presented on index cards rather than by film strip. The presentation rates were 1 second/pair during study and 5 seconds/item during test although the timing was, of course, less accurate than in Experiment I. The same letter-digit pair lists were used as in Experiment I.

To measure any changes that might occur in the guessing rate, the confidence rating scale was modified. Following recall of a number response, S was required to rate her confidence on a ten-point scale. She was presented with a row consisting of the numbers 1 through 10 and asked to circle the number that she felt best represented the number of alternatives from which she had selected her answer. For example, she was asked to circle the number 10 if she felt that any of the ten numbers (0 through 9) could have been correct, i.e., if all alternatives were equally likely to be correct. Circling the number 5 indicated that the number recorded as an answer was selected from among 5 alternatives, which five of course not being specified. A rating of 2 indicated a choice from among 2 alternatives, and a rating of 1 signified that the answer given was the only one considered, i.e., positively correct. This is the same as, but more extensive than, the confidence rating scale used by Phillips and Atkinson (1965).

Results

The confidence rating data were analyzed by the same program as that used in Experiment I. Figure 2 shows d' plotted as a function of trials with the all-recall group from Experiment I plotted for purposes

- - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - -

of comparison. Clearly there were no consistent differences between the two groups. Figure 2 also shows the data from this study as corrected by Bernbach's formula with $g = 1/N = 0.10$. The correction again served to eliminate the increase over trials including, in this case, the trial

1 to trial 2 increase found in the corrected data in Experiment I. (It is not clear why similar recall d' should be so differentially affected by the identical correction as it is in trial 1 performance for Experiments I and II, and the effect must be interpreted as evidence against the validity of the correction.) Here, in fact, the corrected trial 1 value was somewhat higher than those found for trials 2 through 6, an effect also apparent in the 4 and 16 response cases in the data reported by Bernbach.

Again, the ROC curves for trials are well fit by straight lines on double probability paper, one of the ten curves yielding a X^2 beyond the .01 level. Over trials as in Experiment I, the slope of the ROC increased from 0.44 on trial 1 to values not reliably lower than 1.0 on trials 7 through 10.

Figure 3 shows a plot of the inverse of the probability correct given rating j against rating j. If Ss were able to assess accurately the number of alternatives from which they were choosing, the points

----

Insert Figure 3 about here

----

should fall on a straight line with a slope of 1. Clearly, the good fit found by Phillips and Atkinson (1965) for up to four alternatives can be extended at least to five and probably to seven, actual performance beyond that point being better than Ss' ratings indicate.

Figure 4 is a plot of cumulative percent frequency of usage of the 10 rating categories, the parameter being trials. The graph indicates

- - - - - - - - - - - - - - -

Insert Figure 4 about here

- - - - - - - - - - - - - - -

very clearly that the frequency of usage of the ratings 1 and 2 increased markedly over trials (reflected in increasing slopes of lines from 0 to 1 [not shown] and from 1 to 2) and that frequency of usage of higher ratings declined. Having obtained a measure of the number of alternatives from which an S selected her answer as a function of trials, and showing that the measure decreased over trials, it should be possible to correct the data with a trial-dependent, changing, guessing rate. However, when one does correct, using as an estimate of g the inverse of the mean confidence rating for each trial, the corrected hit rate is greater than 1.00 for every trial. Using the inverse of the median confidence rating yields corrected hit rates greater than unity for all but trial 1. The only statistic from the confidence rating data that yields usable values of hit rate is the inverse of the mean confidence rating excluding the rating of one. In other words, one is calculating the mean number of alternatives from which S is choosing only for those items which she rated as being selected from two or more alternatives. Using this statistic, the corrected values are shown in Figure 2. Corrected in this way, d' clearly was not constant over trials. Rather, it decreased sharply over the first 3 trials and then increased over the remaining trials. The drop over the first 3 trials cannot be attributed to the higher incidence of the higher and hence less accurate ratings in the earlier trials. Had the higher ratings been completely accurate, the guessing rate statistic would have been larger and the corrected d' would have been higher.

Discussion

The data of Experiment I have been replicated, and a correction factor of 1/N eliminates the increase in recall d' over trials. The data indicate, however, that the assumption of a constant guessing rate is untenable in that Ss are capable of indicating quite accurately that they are selecting their responses from less than N alternatives and that the mean number of alternatives from which they do select decreases over trials. But the use of a correction factor based on either the mean or median confidence rating yields meaningless data, namely, a hit rate consistently above perfect performance. These data further call into question the tenability of the high threshold correction. An alternative approach to the problem is to provide a situation in which the number of alternatives from which S can select cannot change. Experiment III was designed with that purpose in mind.

Experiment III

Method

Forty female Ss from the same population were run for 9 trials on a 20-pair PAL list. The A members were the same 20 consonants as before; the B member of each pair was the number zero or one, each of the two numbers being paired with ten different letters. Presentation was by film strip using the Dunning animatic projector, the study and test presentation rates being 1 second and 5 seconds respectively, as in the earlier studies. Half of the Ss received recall trials, the other half received recognition trials. The procedure for the two groups was identical to that for the all-recall and all-recognition groups in Experiment I.

Results

The confidence rating data were analyzed by the Ogilvie and Creelman (1968) program, and d' as a function of trials is shown in Figure 5. The recognition curve increased over trials and was in fact

----

Insert Figure 5 about here

----

indistinguishable from the recognition curves obtained in Experiment I. The recall data, on the other hand, were very low and showed no systematic change over trials. Using a correction factor of $g = 0.50$ served to increase all points on the recall curve somewhat, but the data still remained very low and the lack of a change over trials remained.

Again, the ROC curves were all well fit by straight lines on normal-normal paper as none of the X^2 values were reliable at the .01 level. The slopes of the recognition curves were all around 1.0 while the slopes of the recall curves showed the same trend as in the earlier recall conditions, i.e., to start low and to increase toward 1.0 over trials.

Discussion

At this point it seems quite clear that the Signal Detection analysis of recall data, at least through the use of Type II ROC curves, does not provide much insight into the nature of the processes that are involved. At least two questions need to be answered. One question concerns a suitable measure of recall performance, particularly if one wishes to compare it with recognition. A second question is the meaning of ROC analysis of recall data. Consider the questions in order.

One way to think about the recall task in a TSD framework is to consider it as an N-alternative forced-choice situation. In other words, suppose that when a cue (presentation of an A member) is given for recall, S generates the N possible alternatives (a relatively simple task in the studies presented here, although retrievability would clearly become a factor in studies using material from less restricted or less integrated sets) and then selects and outputs that alternative with the highest likelihood (on some scale) of being correct. Given that assumption, the only statistic one needs from the recall data is the per cent correct on each trial. For purposes of comparison with recognition performance, or to compare recall data from experiments with different numbers of alternatives, the proportion correct can be transformed to a forced-choice d' value using Elliot's (1964) tables.

This type of transformation on data of this kind is not novel, being known in the speech communication literature as the detection model for "one-of-M-orthogonal" signals. Green and Birdsall (1964) analyzed the Miller, Heise, and Lichten (1951) data in such a framework and even though it is admitted that the assumption of orthogonality is not met, the analysis serves to transform the articulation scores from different vocabulary sizes to a single function.

This transformation was carried out on the recall data from the three studies reported here. For Experiments I and II the number of alternatives is 10 and the forced-choice d' values were obtained by linear interpolation between the 8 and 16 alternatives conditions in Elliot's tables. For Experiment III the number of alternatives is of course two. The results of this transformation of the recall data are shown in Figure 6 along with the ROC recognition results from Experiments I and III.

Insert Figure 6 about here

It is clear that treating the recall data in this way serves to eliminate the major differences between the recall and recognition curves. As always, there is an exception. In this case it is the data from Experiment II. A 10 alternative forced-choice transformation on the per cent correct data from that study not only does not move the curve up to the recognition curves but in fact does not move the curve away from that calculated on the basis of the confidence ratings. This adds further support to the idea that recall ROC analysis has little to contribute to an understanding of the processes involved. All comparisons of the recall data from Experiments I and II, except that based on the ROC, show performance from Experiment I to be superior to that in Experiment II. The lower performance in Experiment II may well be a function of the nature of the confidence rating scale used. Remember that in Experiment I S was required only to draw a vertical line through a horizontal scale whereas in Experiment II S was required to indicate her confidence by circling a number; the use of numbers in the confidence rating scale may well have interfered with the retention of the correct letter-number pairings in the retention task.

Having suggested then that for purposes of comparing recall and recognition, the most profitable transformation of the recall data is in the nature of the "one-of-M-orthogonal" signals paradigm, the question still remains as to the meaning of the recall ROC curves as typically obtained from the confidence rating data.

In attempting to interpret a recall d', two other pieces of data would appear to be relevant. The first is Pollack's (1959) message reception data which indicate the number of available responses rather than the total number of stimuli to be the crucial parameter. The second is Clarke's (1960) intelligibility test data which show recall d' to be inversely related to the number of alternatives. The argument one would wish to make then with respect to the recall data presented here is that in Experiments I and II, the recall d' increases as Ss manage to reduce the number of available responses through the elimination of those they "know" to be incorrect. Following this argument through, the recall d' in Experiment III does not increase because there are, in effect, no incorrect alternatives to eliminate without automatically yielding the correct response. A possible objection to this interpretation is that the recall d' curve for 2 alternatives would then be expected to be higher with the curve for 10 alternative situations increasing toward it as Ss reduced the effective number of alternatives from 10 to 2. To counteract this objection one might hypothesize the important factor to be the number of alternatives that can be eliminated rather than the number remaining. Clearly more work needs to be done to clarify the factors involved.

Finally, I should like to mention another piece of data from the present studies, not directly related to the major issues under consideration but important from another point of view. Creelman and Donaldson (1968) and Parks (1966) have suggested that S establishes a criterion on the basis of a probability matching model rather than on an ideal observer, maximum likelihood model. In the recall tasks used here, the proportion correct (in effect the a priori probability of an old item) obviously increases over trials. The Ss criterion becomes increasingly "less strict" over trials and it is possible to compare the proportion of items S identifies as old with proportion of old items. For Experiments I and III, the obvious cutoff point is the center of the confidence rating line, anything to the right of center being designated as a response of "old". In Experiment II there is no obvious demarcation so the cutoff that provided the best fit for the trial on which the per cent correct was closest to 0.50 was used (trial 5) for all trials, the split being between ratings 2 and 3. The 50 per cent correct trial was selected, as that is the point where the probability matching model and the maximum likelihood model are indistinguishable (Creelman and Donaldson, 1968). Figure 7

- - - - - - - - - - - - - - -

Insert Figure 7 about here

- - - - - - - - - - - - - - -

shows a plot of the proportion of items S called correct against the actual proportion correct. The data points all fall fairly well along the diagonal, adding further evidence in support of a probability matching model.

References

Bernbach, H. A. Decision processes in memory. Psychological Review, 1967, 74, 462-480.

Clarke, F. R. Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests. Journal of the Acoustical Society of America, 1960, 32, 35-46.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. Two types of ROC curves and definitions of parameters. Journal of the Acoustical Society of America, 1959, 31, 629-630.

Creelman, C. D., & Donaldson, W. ROC curves for discrimination of linear extent. Journal of Experimental Psychology, 1968, 77, 514-516.

Donaldson, W., & Glathe, H. Recognition memory for item and order information. Journal of Experimental Psychology, 1970. (in press)

Donaldson, W., & Murdock, B. B., Jr. Criterion change in continuous recognition memory. Journal of Experimental Psychology, 1968, 76, 325-330.

Egan, J. P. Recognition memory and the operating characteristic. Technical Note AFCRC-TN-58-51., Indiana University, Hearing and Communication Laboratory, 1958.

Elliot, P. B. Tables of d'. University of Michigan: Electronics Defense Group, 1959. (Technical Report No. 97). Published in J. A. Swets (Ed.), Signal detection and recognition by human observers. New York: Wiley, 1964.

Green, D. M., & Birdsall, T. G. The effect of vocabulary size on articulation score. University of Michigan: Electronic Defense Group, Technical Memorandum No. 81 and Technical Note AFCRC-TR-57-58, 1958. Published in J. A. Swets (Ed.), Signal detection and recognition by human observers. New York: Wiley, 1964.

Miller, G. A., Heise, G. A., & Lichten, W. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology, 1951, 41, 329-335.

Murdock, B. B., Jr. An analysis of the recognition process. In C. N. Cofer and B. S. Musgrave (Eds.), Verbal behavior and learning. New York: McGraw-Hill, 1963.

Murdock, B. B., Jr. Signal-detection theory and short-term memory. Journal of Experimental Psychology, 1965, 70, 443-447.

Murdock, B. B., Jr. The criterion problem in short-term memory. Journal of Experimental Psychology, 1966, 72, 317-324.

Norman, D. A., & Wickelgren, W. A. Short-term recognition memory for single digits and pairs of digits. Journal of Experimental Psychology, 1965, 70, 479-489.

Ogilvie, J. C., & Creelman, C. D. Maximum likelihood estimates of ROC curve parameters. Journal of Mathematical Psychology, 1968, 5, 377-391.

Parks, T. E. Signal-detectability theory of recognition-memory performance. Psychological Review, 1966, 73, 44-58.

Phillips, J. L., & Atkinson, R. C. The effects of display size on short-term memory. Technical Report 78, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1965.

Pollack, I. Message uncertainty and message reception. Journal of the Acoustical Society of America, 1959, 31, 1500-1508.

Swets, J. A. Is there a sensory threshold? Science, 1961, 134, 168-177.

Footnotes

1. The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare to the Learning Research and Development Center, University of Pittsburgh. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.

2. Requests for reprints should be sent to Wayne Donaldson, Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania, 15213.

3. Throughout the paper d' is to be interpreted simply as the distance of the ROC curve above the chance diagonal as measured at the point of intersection of the ROC with the negative diagonal. When the term is used it will always be clear, usually through pre-modification, whether it comes from recall or recognition procedures. For recognition studies this is the standard definition of d'. For recall data it is equivalent to what Clarke, Birdsall and Tanner (1959) label $\sqrt{d_e}$. In a later section of the paper d' is modified as being forced-choice, the definition being in the text. While the interpretation of d' may differ depending on whether the ROC is derived from recall or recognition data, it is just that question of interpretation that is being examined here and consequently d' will be used throughout only as a descriptive statistic.

Figure Captions

Figure 1 d' as a function of trials for all groups in Experiment I. Also plotted is d' as a function of trials for the group receiving all-recall trials after Bernbach's (1967) suggested "correction for guessing" has been applied to the data.

Figure 2 d' as a function of trials for Experiment II. Also plotted is d' as corrected by Bernbach's formula with (a) a constant guessing rate and (b) a trial-dependent, changing guessing rate derived from the confidence rating data. For purposes of comparison, the uncorrected d' for the group receiving all-recall trials in Experiment I is also shown.

Figure 3 The reciprocal of the probability correct, conditionalized on the confidence rating given by S, as a function of confidence rating (Experiment II).

Figure 4 Cumulative per cent frequency of use of different confidence ratings, with trials as the parameter (Experiment II).

Figure 5 d' as a function of trials for the recall and recognition groups of Experiment III. Also plotted is d' as a function of trials after Bernbach's suggested "correction for guessing" has been applied to recall data.

Figure 6 Forced-choice d' for the recall group of Experiments I, II and III and standard ROC d' for the recognition groups of Experiments I and III.

Figure 7 Proportion of items called "correct" as a function of actual proportion of correct items (recall conditions from all experiments). Individual points are for separate trials.

Figure 1

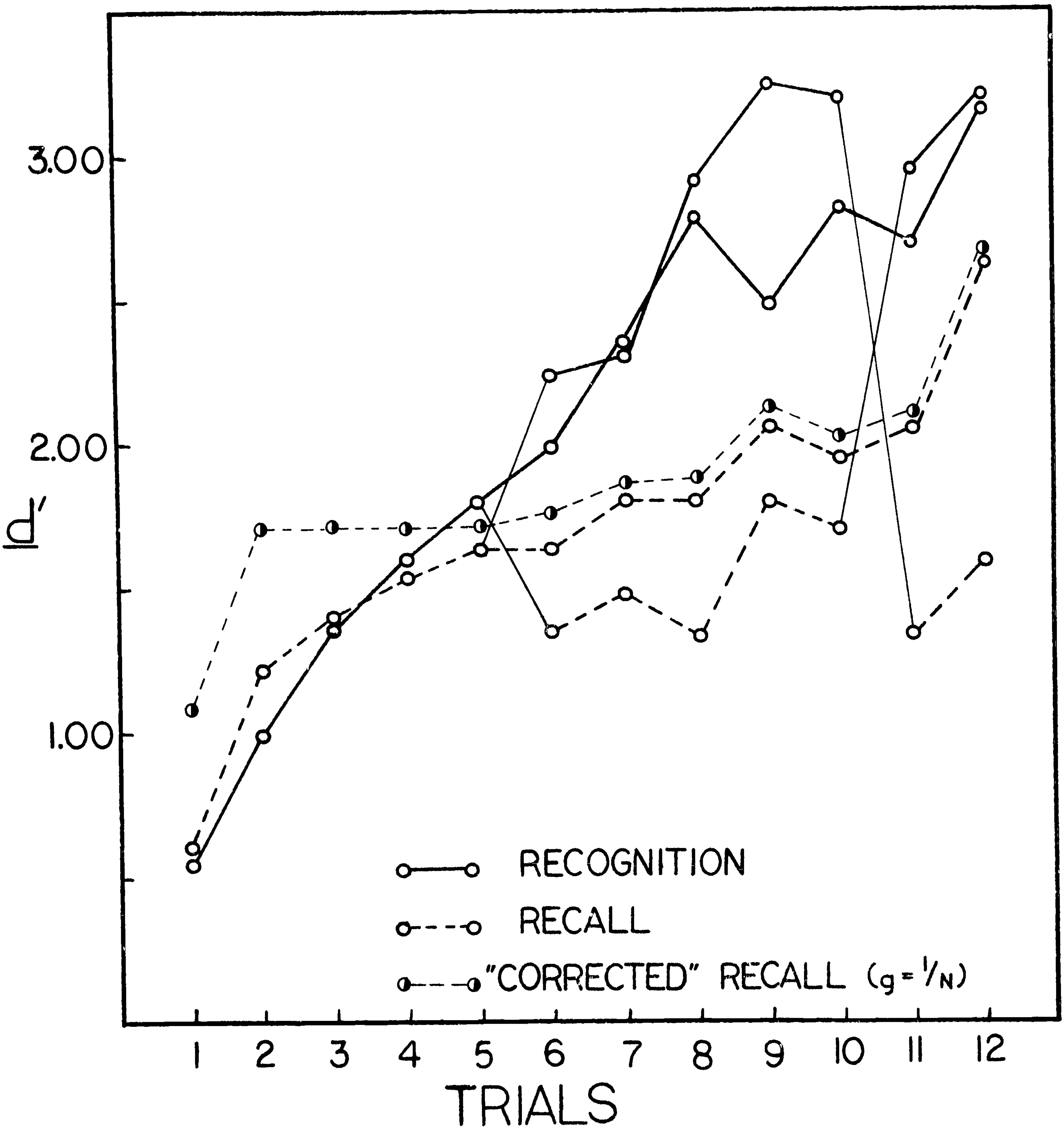

3.00
2.00
1.00
d'
RECOGNITION
RECALL
"CORRECTED" RECALL (g = 1/N)
1
2
3
4
5
6
7
8
9
10
11
12
TRIALS

Figure 2

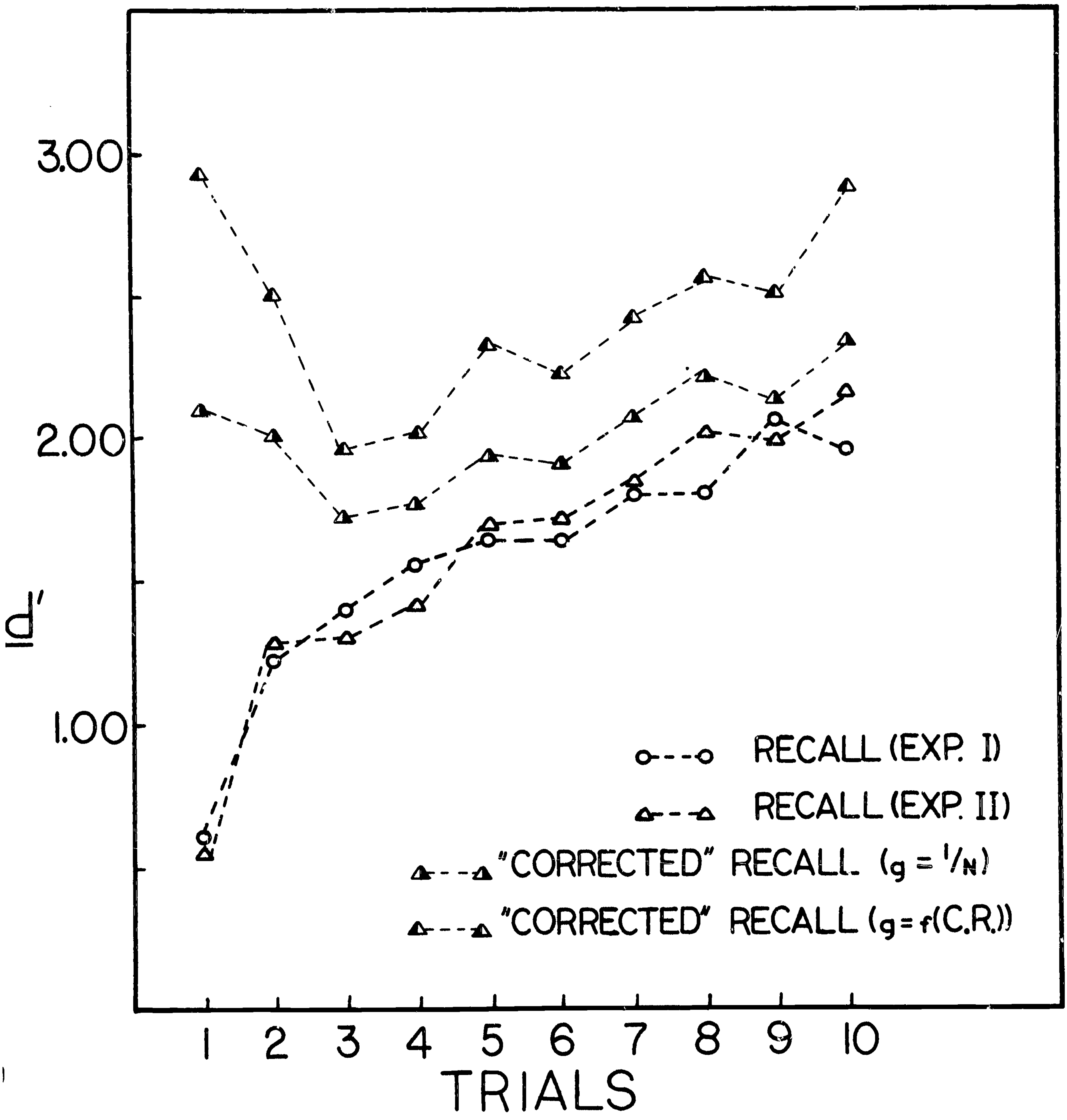

3.00
2.00
1.00
d'
RECALL (EXP. I)
RECALL (EXP. II)
"CORRECTED" RECALL (g = 1/N)
"CORRECTED" RECALL (g = f(C.R.))
1
2
3
4
5
6
7
8
9
10
TRIALS

Figure 3

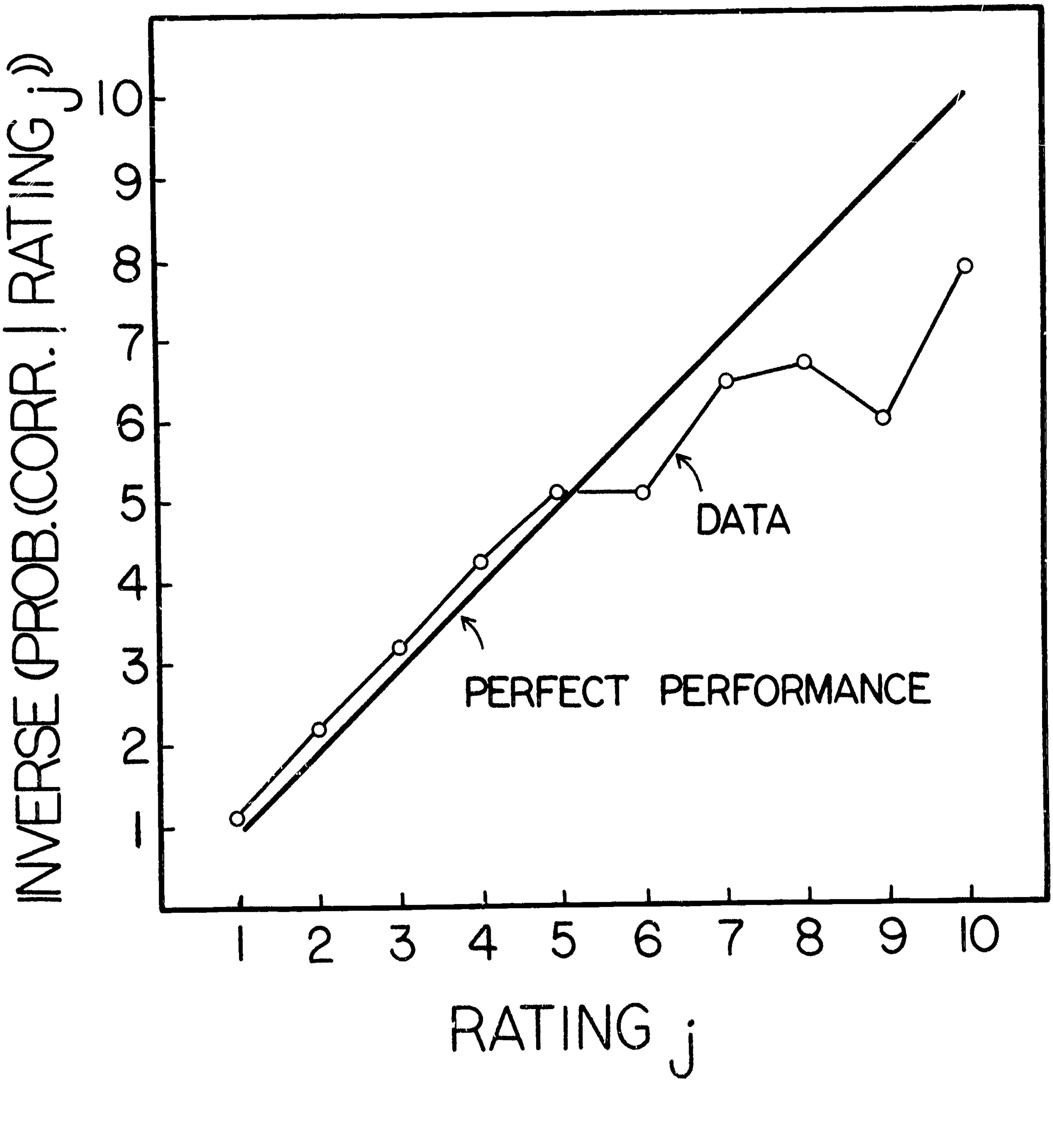

INVERSE (PROB.(CORR. | RATING j))
10
9
8
7
6
5
4
3
2
1
DATA
PERFECT PERFORMANCE
1
2
3
4
5
6
7
8
9
10
RATING j

Figure 4

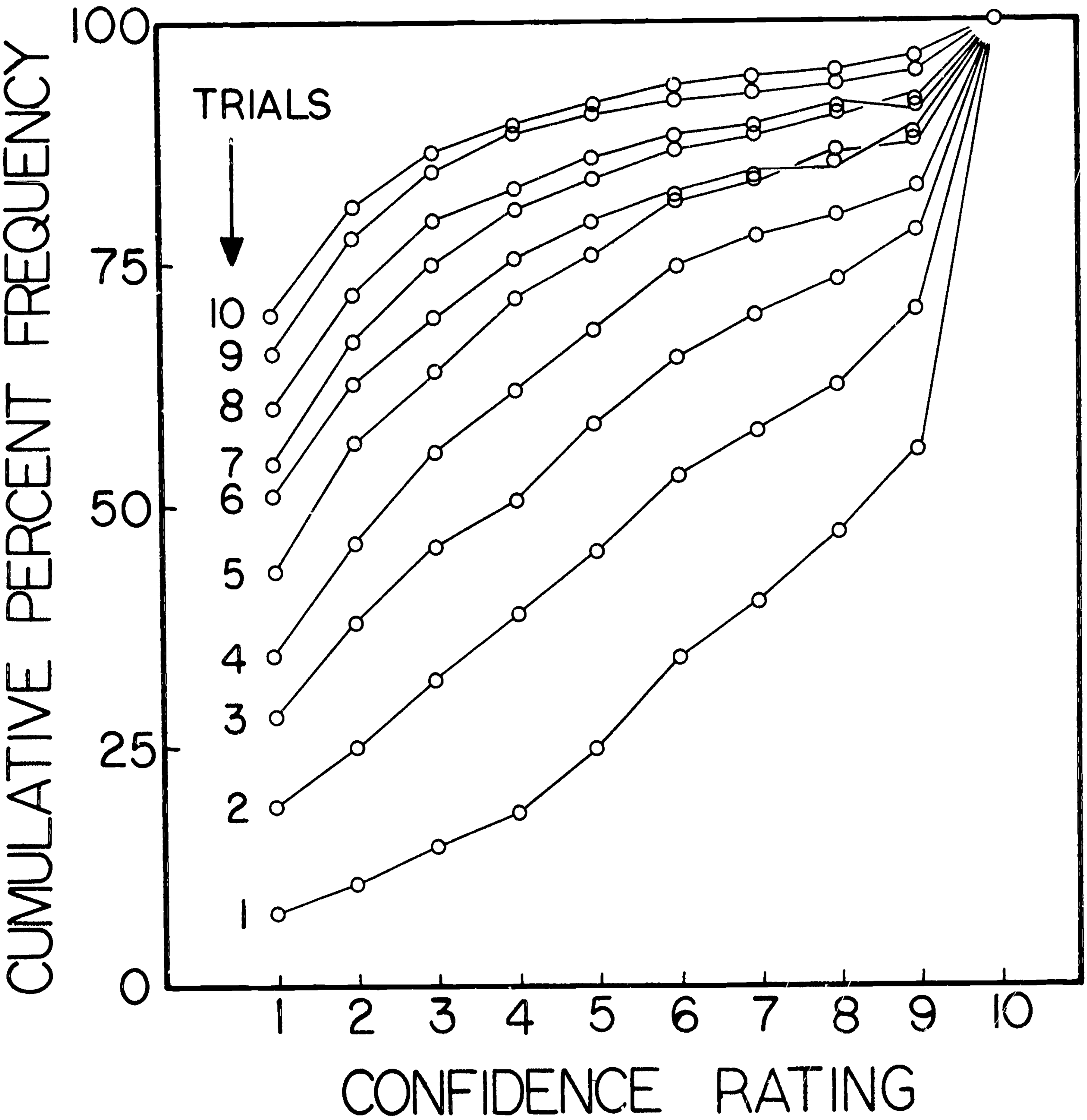

100
75
50
25
0
CUMULATIVE PERCENT FREQUENCY
TRIALS
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10
CONFIDENCE RATING

Figure 5

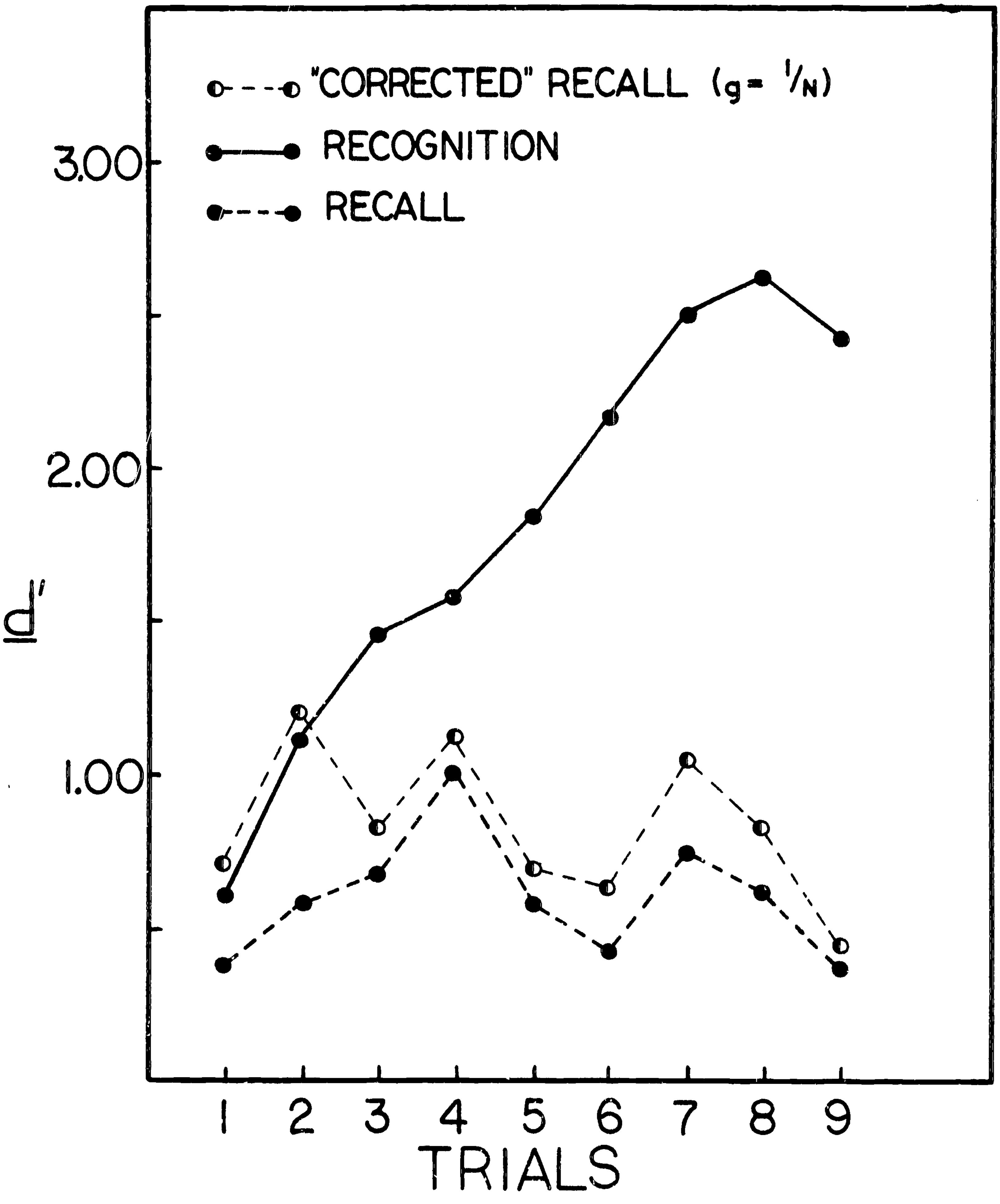
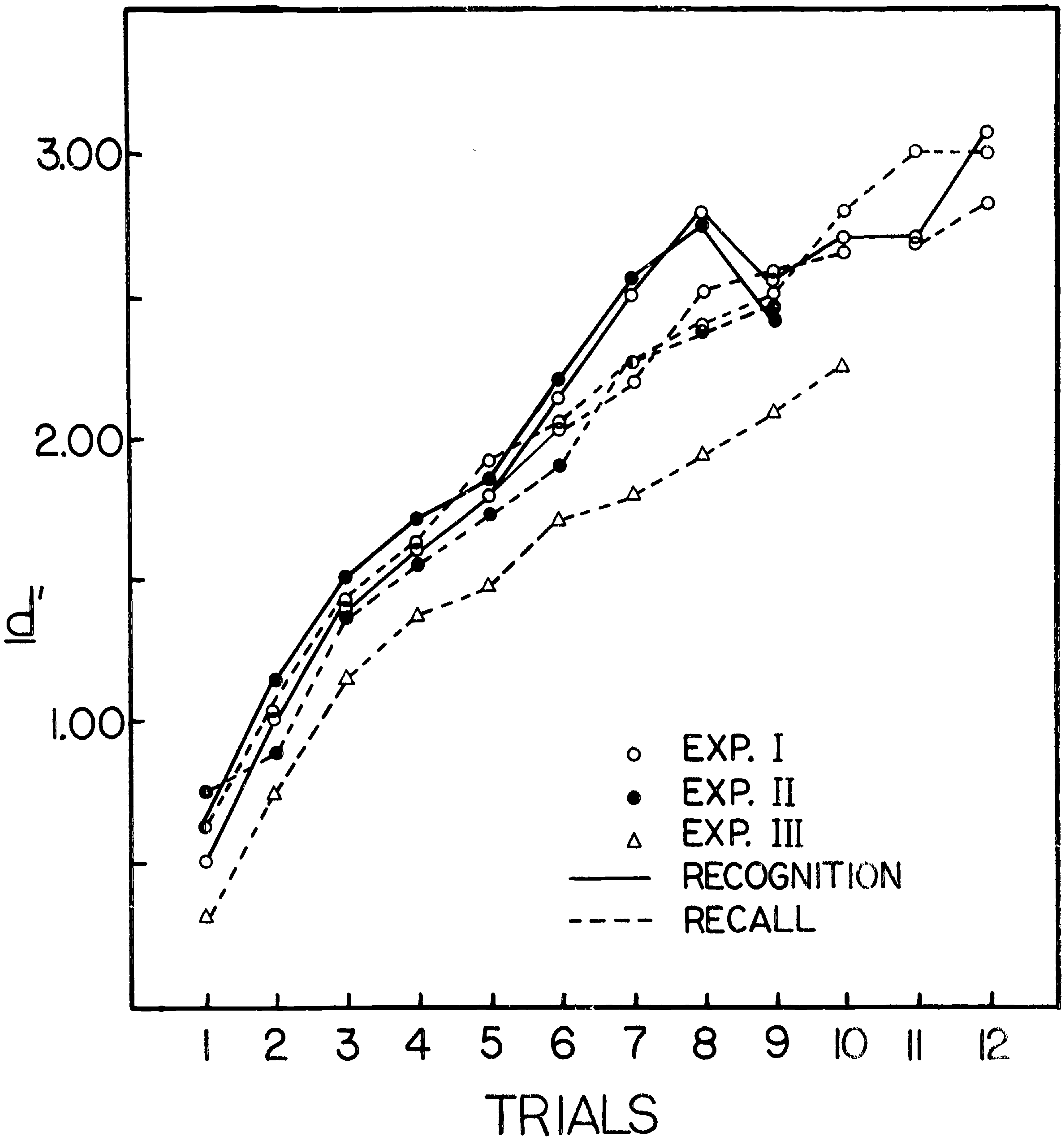
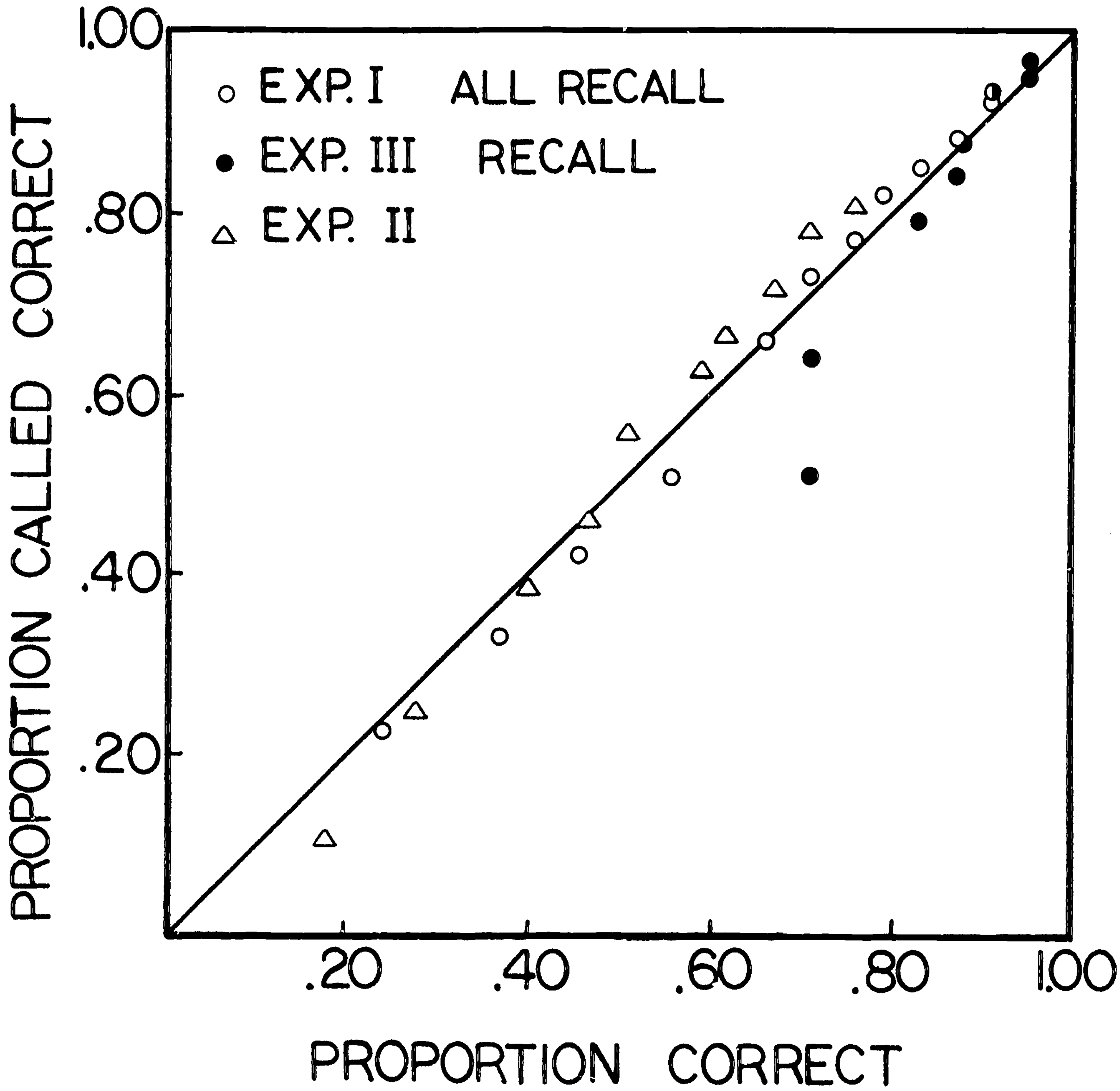

"CORRECTED" RECALL (g= 1/N)
RECOGNITION
RECALL
3.00
2.00
1.00
d'
1 2 3 4 5 6 7 8 9
TRIALS

Figure 6



3.00
2.00
1.00
d'
EXP. I
EXP. II
EXP. III
RECOGNITION
RECALL
1
2
3
4
5
6
7
8
9
10
11
12
TRIALS

Figure 7


1.00
.80
.60
.40
.20
EXP. I ALL RECALL
EXP. III RECALL
EXP. II
PROPORTION CALLED CORRECT
.20
.40
.60
.80
1.00
PROPORTION CORRECT