

AD A 069703

ARI TECHNICAL REPORT  
TR-78-A32



**LEVEL**

**Training for Calibration**

by

Sarah Lichtenstein and Baruch Fischhoff  
DECISION RESEARCH  
1201 Oak Street  
Eugene, Oregon 97401  
A Branch of PERCEPTRONICS, Inc.

**NOVEMBER 1978**



**DDC FILE COPY**

Contract DAHC 19-77-C-0019

Monitored technically by S. M. Halpin and R. H. Phelps  
Human Factors Technical Area, ARI  
Edgar M. Johnson, Chief

Prepared for



U.S. ARMY RESEARCH INSTITUTE  
for the BEHAVIORAL and SOCIAL SCIENCES  
5001 Eisenhower Avenue  
Alexandria, Virginia 22333

Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR-78-A32	2. GOVT ACCESSION NO. 17 ARI	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) TRAINING FOR CALIBRATION	5. TYPE OF REPORT & PERIOD COVERED Final report	6. PERFORMING ORG. REPORT NUMBER PTR-1047-78-5
7. AUTHOR(s) Sarah/Lichtenstein & Baruch/Fischhoff	8. CONTRACT OR GRANT NUMBER(s) DAHC 19-77-C-0019	9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Research 1201 Oak Street Eugene, Oregon 97401 <i>New</i>
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F	11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333	12. REPORT DATE Nov 1978
13. NUMBER OF PAGES 51	14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) ---	15. SECURITY CLASS. (of this report) Unclassified
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.	17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) ---	18. SUPPLEMENTARY NOTES Technically monitored by S. M. Halpin and R. H. Phelps, Human Factors Technical Area, ARI. Decision Research is a branch of Perceptronics, Inc.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) training probability assessment calibration decision making	20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Two experiments attempted to improve the quality of people's probability assessments through intensive training. The first involved 11 sessions of 200 assessments each followed by comprehensive feedback. It showed considerable learning, almost all of which was accomplished after receipt of the first feedback. There was modest generalization to several related probability assessment tasks, no generalization at all to two others. The second experiment reduced the training to three sessions. It revealed the same pattern of learning and limited generalization. The implications of these results are discussed.	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

## FOREWORD

The Human Factors Technical Area of the Army Research Institute (ARI) is concerned with the demands of increasingly complex battlefield systems which are used to acquire, transmit, process, disseminate, and utilize information. This increased complexity places greater demands upon the operator interacting with the machine system. Research in this area is focused on human performance problems related to interactions within command and control centers as well as issues of system development. It is concerned with such areas as software development, topographic products and procedures, tactical symbology, user-oriented systems, information management, staff operations and procedures, decision support, and sensor systems integration and utilization.

An issue of special concern within the area of decision support has been to improve the use of information in forming intelligence estimates and making subsequent tactical decisions. Previous ARI research has confirmed the intuitive impressions of military personnel that serious deficiencies exist in current methods for integrating and evaluating information (ARI Technical Papers 200, 250 and 260). Once problem areas are localized, two general approaches can be taken: (1) judgment and decision-making aids can be developed; or (2) procedures for training analysts to overcome the difficulties can be developed.

The current effort concentrated on developing training techniques to help analysts to more accurately use numerical probabilities to indicate their degree of confidence in their decisions. For the intelligence analyst, the probability could represent how certain the analyst feels the information or subsequent decision will, in fact, turn out to be true. A computerized training procedure was developed and experimentally shown to be effective in substantially improving the accuracy of the probability estimates within a relatively short period of time.

Research in the area of decision support is conducted as an ARI in-house effort augmented by contracts with organizations selected for their specialized capabilities and unique facilities. The present research was conducted by personnel from Decision Research, a Branch of Perceptronics, under contract DAHC 19-77-C-0019. Research in this area is responsive to general requirements of Army project 2Q162722A765, and to special requirements of the US Army Combined Army Combat Development Activities, Ft. Leavenworth, Kans., and the US Army Intelligence Center and School, Ft. Huachuca, Ariz. This special effort was conducted under Army project 2Q161102B74F as basic research responsive to the above requirements.

Accession For	□
DTIC GARD	
Unannounced	
Justification	
By	
Date	
Approved	
Disposal	
	<b>A</b>

*Joseph Zeidner*  
JOSEPH ZEIDNER  
Technical Director

## Table of Contents

	<u>Page</u>
Summary	1
Introduction	3
Measurement of Calibration	4
The Brier partitions	6
The Brier partitions for four-alternative items	7
Over/underconfidence	8
Two-parameter model	9
Experiment 1	9
Method	9
Design	9
Instructions	11
Pre-test items	11
Training items	13
Feedback	14
Participants	17
Results	17
Training sessions	17
Participants' insights	24
Generalization tasks	25
Discussion	27
Experiment 2	27
Method	27
Participants	27
Results	28
Training sessions	28
Generalization tasks	28

	<u>Page</u>
Discussion	
References	29
Footnotes	31
Appendix	33
	35

## Summary

### Introduction

An intensive training program was developed to improve the use of probability judgments. It proved to be moderately effective with almost all participants. In addition, the learning generalized somewhat to some, but not all, of a number of related tasks. Drastically abbreviating the training period had no effect on performance, indicating that the present procedure may achieve measurable improvement with but a modest investment of resources.

### Background

A vital part of much decision making in military and intelligence contexts is assessing the probability that a particular piece of information is correct. Earlier research has shown that people have considerable difficulty making appropriate probability assessments. Their judgments are often sufficiently in error to prejudice the validity of decisions based upon them. In addition, the nature of the error in probability assessments varies from context to context so that it is difficult to improve them by simply applying an error correction factor. As consequence, training that generalizes to various contexts seems the only way to improve probability assessments. Several investigators have attempted such training with mixed results and modest generalization to other tasks (if generalization was tested at all).

### Approach

The present investigations were an attempt to provide a definitive answer to the question "Can people be trained to use probabilities more appropriately?" An intensive training procedure was developed in which participants went through 11 sessions requiring 200 probability assessments each. After each session they received detailed verbal and quantitative feedback on their performance and how to improve it. The 11 training sessions were preceded and followed by six other probability assessment tasks that differed from the training tasks in content and/or response mode. These pretests and posttests were designed to test the generalizability of whatever was learned in the training sessions.

### Findings and Implications

The 11-session training program proved to be quite effective. At its conclusion, all participants who were not using probabilities appropriately to begin with showed marked improvement. Somewhat surprisingly, almost all improvement came after the first round of detailed feedback. This improvement generalized to some, but not all, of the training tasks. A second study reduced the training program from 11 to 3 sessions. Similar results were observed: some improvement wherever possible, all improvement after the first session, considerable generalization.

These results are reason for optimism that people can be taught to use probabilities more effectively in a cost-effective manner. The fact that generalization was not universal suggests that some caution be used before assuming that people who have been trained in one context will retain their improvement in others.

Aside from being an intensive training study, this investigation also provided one of the first in-depth studies of individual differences in the use of probabilities. An unanticipated subsidiary result was the discovery that some people used probabilities well to begin with (and were unaffected by the training program). This reflects some natural ability and/or the careful instruction in the meaning of probabilities given to all participants before they began the various tasks.

## Introduction

According to the subjectivist, or Bayesian, position, all probability assessments are expressions of confidence in the state of one's knowledge (deFinetti, 1937; Phillips, 1973). All may be cast in the form "The probability that proposition A is true is .XX." While all such probability statements are expressions of an internal state--a degree of belief--they can be evaluated by external measures of goodness. For example, sets of probabilities must conform with the axioms of probability theory. The aspect of goodness examined in the present report is the correspondence, across a set of probability statements, between the probabilities and the truth of the propositions. If probability assessments are appropriate reflections of how much one knows, higher probabilities should be associated with correct propositions more often than should low probabilities. The formalization of this property is called "calibration." An assessor is considered to be well calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true is equal to the probability assigned. Thus, across all the occasions that the assessor assigns the probability .7, just 70% should be true; for all propositions to which .8 has been assigned, 80% should be true, and so forth.

A great deal of empirical research (reviewed by Lichtenstein, Fischhoff & Phillips, 1977) has shown that people's calibration is usually poor. Typically, people are overconfident: they believe they know more than they actually know. Thus it is not unusual for only about 60% of all the propositions to which a probability of .8 was assigned to be true in fact. However, people are not always overconfident in their probability assessments. When given very easy tasks, they are underconfident: the proportions true tend to be larger than the assessed probabilities. In short, severely biased calibration has been frequently observed. The direction of this bias depends primarily on the difficulty of the task (Lichtenstein & Fischhoff, 1977).

Since probability assessments so often play a key role in important decisions, eliminating the strong and systematic calibration bias is necessary. One approach to correcting the bias is for someone to adjust the probabilities after the assessor has produced them. This solution will not be generally applicable. Since the direction of the bias is a function of the difficulty of the task, one needs to know difficulty before knowing what correction to apply. But in real decision situations (as opposed to laboratory studies of decision making), there is usually no way to establish the difficulty of the task facing the assessor.

Since correction of probabilities is usually impossible, one would like to have probability assessors whose probabilities are unbiased to begin with. Since untrained people seem to be quite badly calibrated, the only reliable way to identify well-calibrated individuals may be to have people specifically trained for that skill. The National

Weather Service has been training and eliciting probability forecasts for over 10 years with excellent results (Murphy & Winkler, 1977a, 1977b) for rain and temperature predictions, and somewhat poorer results for tornado forecasts (Murphy & Winkler, 1977c). These were highly trained (and talented) individuals making assessments for what, in some senses, were quite homogeneous tasks. In addition, weather forecasters typically have extensive background data on the base rates of various weather phenomena; often they have access to computer-generated forecasts upon which to base their own predictions.

The results of laboratory training research are, however, sketchy, inconsistent and much less promising. Adams and Adams (1958) showed modest improvement in calibration after five training sessions, and in a later study (1961) they found some generalization of improvement in calibration across several different types of items. Pickhardt and Wallace (1974) reported slight improvement over five or six sessions of one task, but in another experiment employing a more realistic game setting, they found no improvement in calibration over 17 sessions. Choo (Note 1), using only one training session of 75 items, found little learning and no generalization.

No one has yet performed a thorough training study in which assessors receive intensive instruction and are trained for many sessions, with enough responses per session to ensure accurate feedback, and with sufficient tests of generalization. The present report describes two such training experiments. These experiments were an attempt to train "ordinary" individuals to be well calibrated. The focus of the training program was computerized feedback provided after each session of 200 assessments. Such long sessions were used in order to obtain relatively stable estimates of people's calibration, and thereby to avoid providing false or misleading feedback (for example, an individual whose assessments actually tend to be too high might in a small sample provide assessments that are too low). A large number of sessions (11) was used in Experiment 1 in order to rule out the possibility that the modest or negligible success of earlier training experiments was due to insufficient intensity of training. In order to evaluate the generalizability of whatever skills might be acquired in the training sessions, subjects completed identical pre-test and post-test tasks that differed from the training tasks in content and/or response mode.

#### Measurement of Calibration

One of the easiest ways to conceptualize good calibration is to plot it. Suppose you have been given a set of two-alternative factual items, such as "which is longer, the Suez Canal or the Panama Canal?" Your task is twofold: first, state which of the two alternatives is, in your belief, the correct answer, and second, state the probability that you are right. Your responses, then, are constrained to the range  $.5 \leq r_i \leq 1.0$ . Measuring the calibration of your probability assessment first involves comparing each of your answers with the truth. For example, given that you said ".8" for a large number of items, what proportion of those times did you pick the correct alternative? These proportions are plotted as in Figure 1.

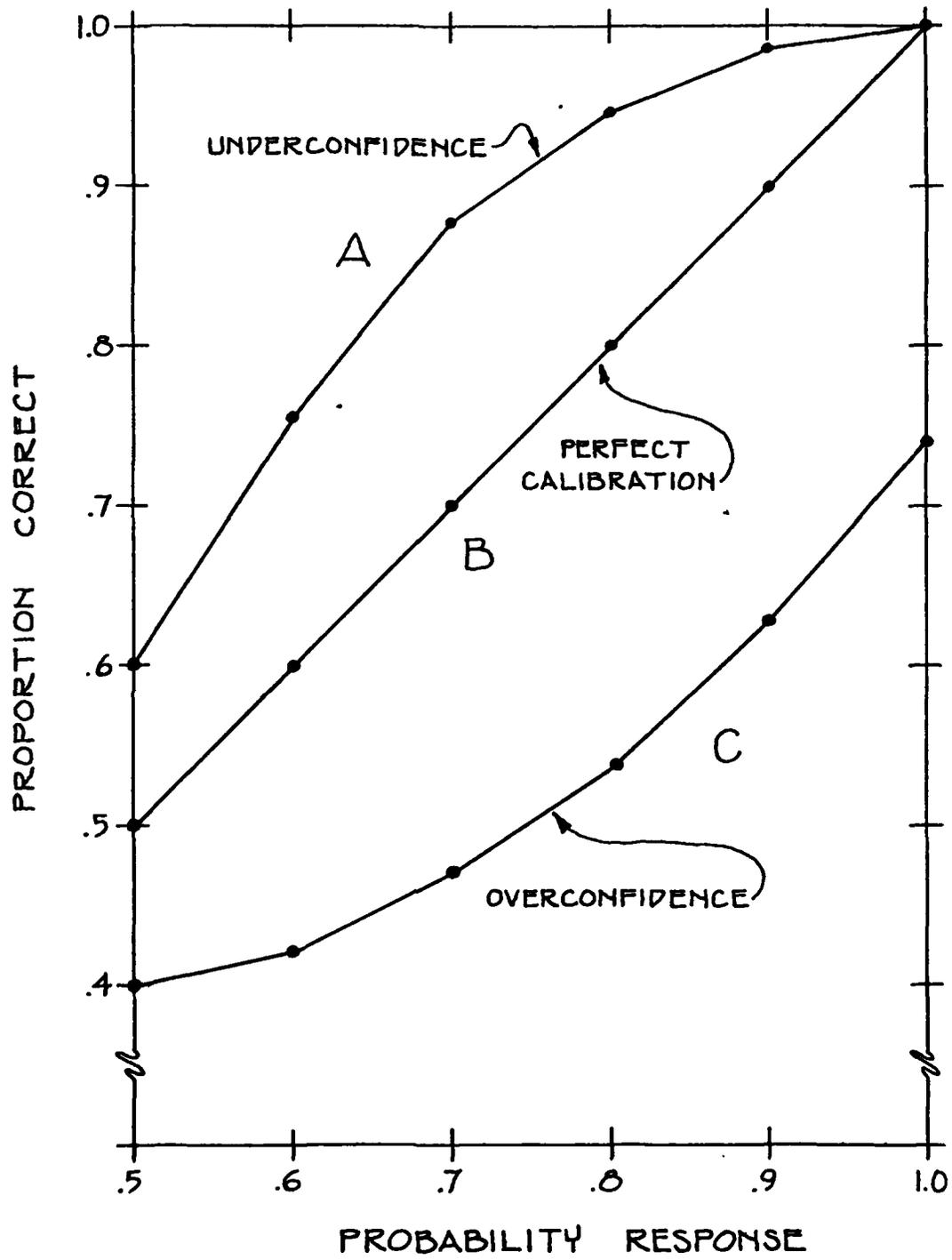


Figure 1

Exemplary Calibration Curves

If you are perfectly calibrated, the chosen alternative will be correct for 80% of the times you said .8, for 70% of your .7 assessments, and so on. Thus, your data will fall along the diagonal. Data below the diagonal indicate overconfidence; for example, of all the times you said .9, you were right on only 80% of them. Data above the diagonal indicate underconfidence; you knew more than your responses indicated.

Numerous measures of the adequacy of calibration have been suggested; for a review see Lichtenstein, Fischhoff and Phillips (1977). The measures chosen for the present study are the partitions of the Brier score proposed by Murphy (1973), a simple measure of over- or underconfidence, and a two-parameter model fitted to the data.

The Brier partitions. The Brier score (Brier, 1950) is a proper scoring rule, i.e., it reaches its best value only when the assessor responds with his or her true beliefs. It runs from 0 to 1 with 0 being the best. It can be calculated for either a single response or a set of responses. For the latter case, Murphy (1973) has shown that the Brier score can be partitioned into three additive parts. In the two-alternative task described above, Murphy's partition (which he calls the "special scalar partition," 1972a) is as follows:

$$\frac{1}{N} \sum_{i=1}^N (r_i - c_i)^2 = \bar{c} (1 - \bar{c}) + \frac{1}{N} \sum_{t=1}^T n_t (r_t - \bar{c}_t)^2 - \frac{1}{N} \sum_{t=1}^T n_t (\bar{c}_t - \bar{c})^2 \quad (1)$$

To calculate all of this, one takes responses to N two-alternative items for which the correct answer is known, and sorts them into T categories, such that all the numerically equal responses are in the same category. Then N is the total number of responses,

$r_i$  is any particular response (there are N of them),

$c_i$  is 1 if the response  $r_i$  was given to the correct alternative,

0 otherwise,

$\bar{c}$  is the proportion of the N responses that were attached to the correct alternative,

T is the number of categories into which one sorts the responses;

1, ... t, ... T,

$n_t$  is the number of responses in the t'th category,

$r_t$  is the numerical value of the responses in the t'th category, and

$\bar{c}_t$  is the proportion of responses in the t'th category which were attached to the correct alternative.

The term on the left of Equation 1 is the total Brier score, which is a general measure of goodness of probability assessment. The first term on the right is a measure of knowledge. If the correct alternative were always chosen, this component would have a value of zero. The maximum (worst) value is for total ignorance (or random guessing) .25.

The second term on the right of Equation 1 measures calibration (which Murphy calls "reliability"). Essentially, it is the weighted mean squared distance between the data points in Figure 1 and the

diagonal line indicating perfect calibration. Thus its ideal value is 0. The third term is called resolution. It measures the assessor's ability to sort the items into categories whose proportions of correct answers are maximally different. It is the variance of the  $\bar{c}_t$ 's across T. This term is subtracted from the others; it should be as large as possible.

As a proper scoring rule, the Brier rewards people for responding with their own true beliefs. Individuals who try to respond strategically so as to improve their scores often encounter a tradeoff between calibration and resolution. Some simple strategies to improve the calibration score (e.g., always respond with  $\bar{c}$ ) will degrade the resolution score, and vice versa. Murphy (1974) has shown that the "sample skill score," the resolution score minus the calibration score, is itself a proper scoring rule. This score can be used to test the hypothesis that advantages gained in training people to be well calibrated are offset by losses in resolution.

The partitions of the Brier score are sensitive to sample size. At the extreme, it would be impossible to have perfect calibration if one used a particular response, say .8, only once. As the sample size increases, variability due to chance decreases, but our impression from using the partitions is that even with a sample of 200 items chance variations can be large relative to the size of the scores.

In addition, when the number of items remains constant, the number of categories, T, will affect the size of the calibration and resolution components, particularly when, as often happens, the assessor uses one-digit responses (.5, .6, ... , 1.0) for the vast majority of responses, but uses two-digit responses (.55, .75, .99, etc.) occasionally. These occasional responses artificially inflate both the calibration and resolution scores since  $\bar{c}_t$  for these rarely used categories suffers greatly from random variation. Accordingly, in this report all data were converted to six grouped categories: .50 to .59, .60 to .69, ... , .90 to .99, and 1.0 before partition scores were calculated. The mean response in each grouped category was taken as  $r_t$ , and the proportion correct across the whole category was used for  $\bar{c}_t$ . This reduced the noise in some subjects' data.

In practice, variations in the knowledge component tend to be larger than variations in calibration or resolution. Thus the total Brier score is relatively insensitive to the latter two. For example, an increase of just .05 in the proportion of correct responses can improve the Brier score by more than most of our subjects could have accomplished by becoming perfectly calibrated. Since the difficulty of the test sets was not of primary concern in the present study, the focus of most of our analyses was on the calibration and resolution scores.

The Brier partitions for four-alternative items. When the assessor has not two but four mutually exclusive and exhaustive alternatives to assess, there are two ways of computing the Brier

partition scores, a vector method and a scalar method, as has been discussed by Murphy (1972a, b). In the scalar method, the probabilities assigned to the four alternatives are tallied separately, as if each had been given in response to a one-alternative true/false item, disregarding the fact that the four responses are constrained to sum to 1.00. In the vector method the four probabilities are kept together and retained as a vector throughout the calculations. These vectors are sorted into categories, with the vector (.1, .2, .3, .4) considered as different from (.4, .3, .2, .1). As Murphy (1972b) points out, the Brier score itself is numerically the same under the two methods, but the partition scores differ.

One of the tasks in the present study involved general-knowledge questions with four possible answers. The participants were asked to state the probability that each alternative was correct. For this task, there are two reasons for using the scalar rather than the vector method.

(a) The advantage of using the vector method is that it differentiates the response vector (.1, .2, .3, .4) from (.4, .3, .2, .1). This distinction is useful whenever the ordering of alternatives is consistent and meaningful, e.g., {rain, snow, sleet, no precipitation} in a weather forecasting context or {stock price goes up, stays the same, goes down} in a stock-market forecasting context. However, there was no such ordering for the items used here. Each of our items had entirely different alternatives. Thus there would be no valid distinction between the vectors (.1, .2, .3, .4) and (.4, .3, .2, .1).

(b) The vector method requires more data for stable measurement of the partition scores, because the data are separated into a much larger number of categories, so that the effect of random variation in each category is greatly increased. If subjects limit themselves to using only the 11 responses 0, .1, . . . , .9, 1.0, there are 286 different categories for the vector method but only 11 for the scalar method (see Murphy, 1972b, p. 1184).

For these reasons the scalar method was used to analyze the four-alternative data in this report, despite its disregard of the interdependencies in the data. Only the calibration and resolution components were computed. These scores are not comparable to calibration and resolution scores calculated on sets of data from two-alternative items. The knowledge score is not reported, because it is a constant when the scalar method is used.

Over/underconfidence. Since the calibration score, a squared measure, is insensitive to whether an individual is over- or underconfident, a measure of overall over/underconfidence was also calculated: the mean response minus the proportion correct over all items. A positive difference indicates overconfidence; a negative difference means underconfidence.

Two-parameter model. A quite different approach to the measurement of calibration is to fit a smooth curve to the data as represented in Figure 1, and to use the parameters of the curve as an indication of the assessor's calibration. Shuford and Brown (1975) took this approach using a straight-line model with least squares estimates of the slope and intercept for each subject. The model chosen for the present study is a straight line if the data, both  $r_t$  and  $\bar{c}_t$ , are transformed to log odds:

$$\log \frac{\bar{c}_t}{1-\bar{c}_t} = \log A + B \log \frac{r_t}{1-r_t} . \quad (2)$$

This model is an expansion (with the addition of the parameter B) of a model proposed by Schlaifer (1971). It will indicate perfect calibration when  $A = B = 1$ .

The parameter A indicates where the curve crosses  $r_t = .5$  in the  $r_t \times \bar{c}_t$  plot. When  $r_t = .5$ ,  $\bar{c}_{.5} = A/(1+A)$ . Thus values of A greater than 1 indicate that  $\bar{c}_{.5}$  is greater than .5. B is a curvature parameter; the curve is convex when B is greater than 1.0 and is concave when B is less than 1.0. Thus when both A and B are greater than 1.0, underconfidence across the entire scale is indicated; when A and B are both less than 1.0, overconfidence is indicated. When A and B are on opposite sides of 1.0, the curve will be partly overconfident and partly underconfident. Examples of these curves are shown in Figure 2.

The best-fitting parameters, A and B, to each set of data were found using Bayes' Theorem with flat priors on A and B to compute the most likely values of A and B given the data. The computer program searched 2500 different (A, B) pairs, with A ranging from .25 to 4.00 and B ranging from .01 to 5.00. Since the analysis was based on the raw  $r_i$ ,  $c_i$  data, no grouping was necessary. Since the response of 1.0 cannot be converted into odds, these responses were excluded from the calculations.

### Experiment 1

#### Method

Design. The participants attended 23 sessions, each lasting approximately one hour, spread over a four to five week period, with no more than two sessions on any one day. The first six sessions were pre-test tasks designed to test the generalizability of the training. No feedback was given for any of these pre-test tasks. All pre-test tasks were paper and pencil tasks. The first session began with approximately 45 minutes of instruction about probability assessment in general and calibration in particular.

The next 11 sessions were training sessions, each involving 200 two-alternative general-knowledge items. Training sessions 1 and 11 used the same items, in the same order; otherwise no items were

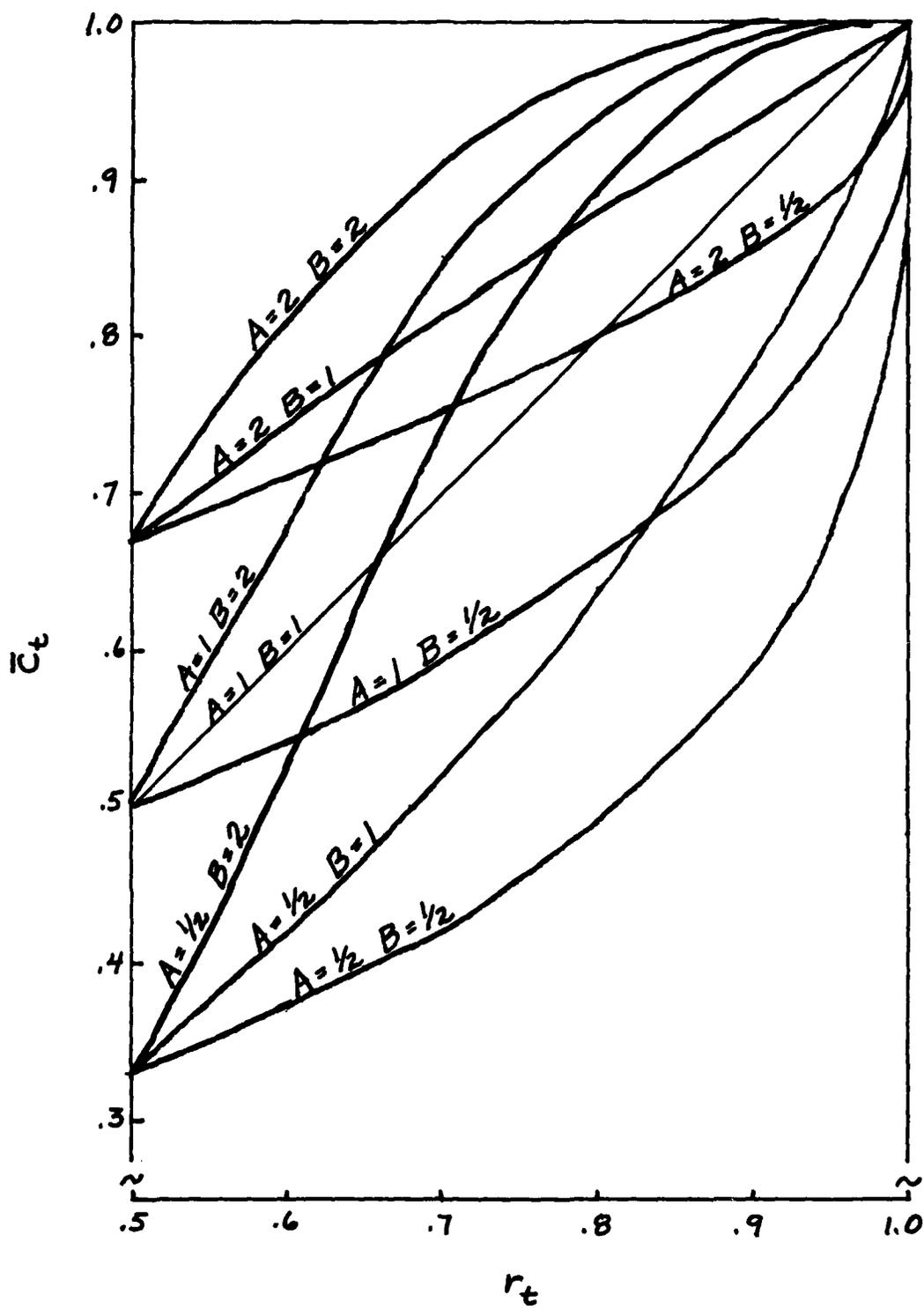


Figure 2

Exemplars of the Fitted Curves

repeated. The participants worked at a computer terminal which presented each item; they typed their responses on the keyboard. They were asked (a) to indicate by typing either "A" or "B" which alternative was the one they believed more likely to be correct and (b) to assign a probability from .5 to 1.0 that the chosen alternative was in fact correct. The participants were allowed to use as many decimal digits as they wished; none ever used more than two (e.g., ".99"). After a participant completed a session of 200 items, the computer printed performance summaries; these summaries were discussed with the participant in a tape-recorded debriefing.

In the last six sessions, the pre-test tasks were repeated as post-tests. The items were identical to those used in the pre-tests. Again, the participants used paper and pencil to respond, and were given no feedback.

Never during the experiment were the subjects told the correct answer to any item. Table 1 summarizes this design.

Instructions. Full instructions are given in the Appendix. The general introduction was read to three groups of four participants. It emphasized the properties that well-calibrated judgments should have, both in general and through extended examples of a well-calibrated and a poorly-calibrated assessor. No hint was given of the type of calibration typically observed.

Instructions for the various parts of the experiment were given individually. In particular, the fractile method for pre-test 6, uncertain quantities, was presented with individual tutoring.

Pre-test items. Examples of each task are given in the Appendix. These tasks, in order of completion, were:

Task 1: Handwriting samples. The participants were given 100 different cards each bearing the handwritten inscription "Mensa mea bona est." Their task was to determine whether each specimen had been written by an American or a European and then to assess the probability of their answer being correct. Thus, although the content of this task was homogeneous and different from that of the training sessions, it did use the same two-alternative format, with probabilities ranging from .5 to 1.0.

Task 2: Shapes. The participants saw 200 21.6 x 27.9 cm (8½ x 11") pieces of paper each of which presented two very irregularly shaped polygons. Their task was to determine which was larger by visual inspection (that is, without physically measuring). This set, too, used the two-alternative, half-range response mode.

Task 3: General-knowledge items. These items were similar to the items used for the training sessions (described below). On the basis of earlier administrations, the first 100 items were known to be especially hard while the second 100 were known to be easy.

Table 1  
Experimental Design  
Tasks Shown in Order of Presentation

Type of Item	Response Mode	No. of Items	Used in Exp II?	Special Features	Feedback
Pretest Tasks					
1 Handwriting	2 Alternatives	100	No		Paper and pencil tasks with no feedback
2 Shapes	"	200	Yes		
3 General Knowledge	"	200	Yes	100 hard then 100 easy	
4 " "	4 Alternatives	199	Yes		
5 " "	2 Alternatives	200	Yes	100 easy then 100 hard	
6 " "	Uncertain Quantities	77	No		
Training Sessions					
1 General Knowledge	2 Alternatives	200	Yes		On-line computer tasks with feedback
2 " "	"	"	Yes		
3 " "	"	"	No		
4 " "	"	"	No		
5 " "	"	"	No	All easy items	
6 " "	"	"	No	All hard items	
7 " "	"	"	No		
8 " "	"	"	No		
9 " "	"	"	No		
10 " "	"	"	No		
11 " "	"	"	Yes	Items identical to Training Session 1	
Posttest Tasks					
6 General Knowledge	Uncertain Quantities	77	No		Paper and pencil tasks with no feedback
1 Handwriting	2 Alternatives	100	No		
2 Shapes	"	200	Yes		
3 General Knowledge	"	200	Yes	All items identical to Pretest Tasks	
4 " "	4 Alternatives	199	Yes		
5 " "	2 Alternatives	200	Yes		

Task 4: Four-alternative items. These items were drawn from the same universe of content as the training items, but used a different response mode. Two additional possible answers were appended to each of 199 two-alternative general knowledge questions. The participants' task was to assess the probability (from .00 to 1.00) that each of these four alternatives was the correct one. They were promised that one and only one was correct and were constrained to make their four assessments total 1.00.

Task 5: General-knowledge items. The first 100 were easy and the second 100 were hard. The items were similar to those used in Task 3 and in the training sessions, but no items were duplicated.

Task 6: Uncertain quantities. The participants received 77 questions having numerical answers (e.g., How many miles long is the Nile River? How many vertebrae does an adult human have?). They used the fractile method to represent their confidence that the true answer lay in various ranges of possible values. Specifically, they assessed five fractile values: (a)  $P_{01}$ , a value such that there was .01 chance that the true value was lower than the value they specified; (b)  $P_{25}$ , a value such that there was a .25 chance that the true value was lower; (c)  $P_{50}$ ; (d)  $P_{75}$  and (e)  $P_{99}$ . The last three were values such that there was a .50, .25 or .01 chance of the true value being higher. For a well-calibrated individual 1% of all true answers should fall below  $P_{01}$ , 25% below  $P_{25}$ , etc. One common way of scoring calibration with such assessments is to look at the proportion of correct answers falling between  $P_{01}$  and  $P_{99}$ . If one were well calibrated, 2% of all answers should fall outside this range. The "surprise index," the percentage actually lying outside has ranged from 7% to 50% in a variety of experimental tests, with 35% being a representative value (Lichtenstein, Fischhoff & Phillips, 1977). Such a large surprise index indicates considerable overconfidence, that is, the belief that one can set much narrower confidence intervals than one should. A second measure is the interquartile index or percentage of true answer falling between  $P_{25}$  and  $P_{75}$ . As might be expected from the surprise index results, less than 50% of the true answers typically fall in the center of the distribution. Both understanding and performing this task appeared to be quite difficult relative to the other tasks.

All the pre-test tasks were repeated in the post-test. Task 6, uncertain quantities, was given first in the post-test because the participants disliked it (assigning fractiles was unfamiliar and difficult) and we did not want to end the experiment with an unpleasant task.

Training items. All training items were two-alternative questions requiring a probability response between .5 and 1.0. This particular format was chosen because the properties of responses to such items are fairly well known (Fischhoff, Slovic & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff & Phillips, 1977). A pool of over 2000 such items was created with the help of reference books ranging over a variety of content areas including geography, history,

literature, science and music. Several hundred of this set had been used in earlier experiments that provided estimates of item difficulty (i.e., the percentage of individuals answering the item correctly). Using these estimates, one set of 400 unusually hard items and one set of 400 unusually easy items were constructed. The mean percentages correct for these sets were approximately 10% lower and higher than for the remaining sets (for which the means were roughly 65%). Half of the hard set and half of the easy set were used in pre-test and post-test tasks 3 and 5. The remaining 200 easy items were designated as training set 5; the remaining 200 hard items became set 7. The other 1600 items were divided randomly into 8 additional sets of 200 to be used in training.

Feedback. Immediately after the participant responded to the two hundredth item of any training session, the computer printed out summary information about the session. For example, Figure 3 shows the feedback shown to Participant 6 after completing the fourth training session. This individual was moderately overconfident.

The first kind of feedback information was the number of correct and incorrect responses (HITS and MISSES) and corresponding percentages correct (P HIT) for each probability response used by the participant. This person used an unusually large number of different responses. In order to provide reasonably stable feedback, responses were grouped into the 6 categories appearing in the second table of Figure 3. The second column reports the mean of all probability responses grouped in each interval. Following the categorized data are a number of summary statistics: (a) the overall proportion correct (here, .685), (b) the overall degree of over- or underconfidence, equal to the mean probability minus the proportion correct (+.054), (c) the calibration score according to Equation 1 (.012), (d) knowledge, calculated according to Equation 1 (.216), (e) resolution, according to Equation 1 (.027), and (f) the Brier score, according to Equation 1 (.201).

The next line gives the most likely values of A and B (see Equation 2). The two following lines give information for the programmer. The participants were instructed to disregard these three lines of output.

The plot in Figure 3 shows the calibration curve for the categorized data. The number of responses was handwritten by each point to keep both participant and experimenter from overinterpreting unstable estimates of percentage correct. The smoothed curve was fit to the data according to the model given in Equation 2.

The results were discussed for five to twenty minutes with each participant; these discussions were loosely structured and tape recorded. Emphasis was placed on making the graphed data fit as closely as possible to the diagonal line of perfect calibration and getting all the 1.0 responses correct. The participants were warned not to overinterpret discrepant points based on few data.

Figure 3. Exemplary feedback for training sessions

SUBJECT 6 DAY 4  
ITEM FILE: SIX.

SCORE FILE: 506D04

NUMBER OF PROBABILITY CATEGORIES USED = 11				
PROB	P HIT	HITS	MISSES	TOTAL
0.50	0.538	7	6	13
0.55	0.517	15	14	29
0.60	0.611	11	7	18
0.65	0.687	22	10	32
0.70	0.471	8	9	17
0.75	0.650	13	7	20
0.80	0.667	2	1	3
0.85	0.833	10	2	12
0.90	0.571	4	3	7
0.95	0.810	17	4	21
1.00	1.000	28	0	28
TOTALS				
0.739	0.685	137	63	200

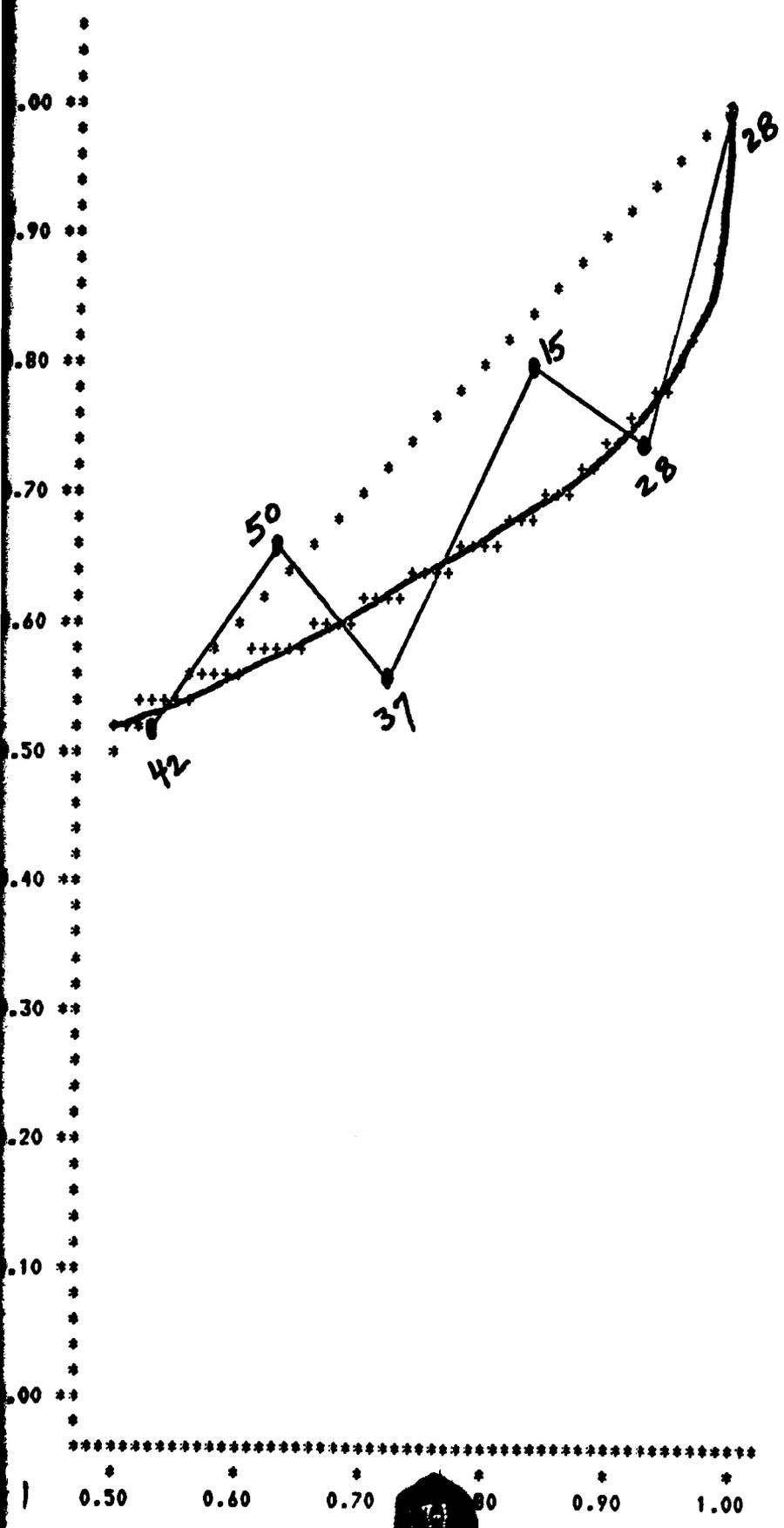
SUMMARY BY PROBABILITY RESPONSE INTERVALS:

PROBABILITY ESTIMATE INTERVAL	AV EST PROB OF CORRECT RESPONSE	PROPORTION CORRECT IN INTERVAL	NUMBER OF HITS	NUMBER OF MISSES	NUMBER OF RESPS
0.50-0.59	0.535	0.524	22	20	42
0.60-0.69	0.632	0.660	33	17	50
0.70-0.79	0.727	0.568	21	16	37
0.80-0.89	0.840	0.800	12	3	15
0.90-0.99	0.937	0.750	21	7	28
1.00-1.00	1.000	1.000	28	0	28
0.50-1.00	0.739	0.685	137	63	200

SUBJ #	PROP CRRCT	OVER/ UNDER	CALIB	KNDW	RESOL	BRIER
6	0.685	0.054	0.012	0.216	0.027	0.201

A\* = 1.1500    B\* = 0.4167  
 FIRST LOGP = -188.6    LOGP AT MODE = -110.5    PADD USED = 135.  
 ACCEPTABLE PADD RANGE FOR THIS DATA: 101.6 TO 197.5

UNIT  
RATE



Participants. Twelve individuals, 7 females and 5 males, were recruited by personal contact. They were paid \$2.35 per hour for participating, with a \$1 per hour bonus if they completed the full experiment, lasting 20 to 25 hours.

The participants' ages ranged from 17 to 33. Two were high school students known to the experimenters to be exceptionally bright (#5, #6). Seven were college students (#1, #2, #3, #4, #7, #11, #12). The other three participants were not in school. One had completed high school (#6), one had finished one year of college (#8), and one had a Ph.D. in experimental psychology (#9).

### Results

Training sessions. Table 2 shows the calibration scores for each of the eleven training sessions.<sup>1</sup> For ease of reading, each calibration score in Table 2 is multiplied by 1000, e.g., "7" represents ".007." Since the calibration score is not a familiar measure, we show in Figure 4 examples of actual data for several different values of the measure. Each line in Figure 4 is the calibration of one participant for one 200-item training session. All these examples contain at least 14 responses at every point (often our participants used a particular response less than 10 times in 200 trials; such curves tended to show greater irregularity). Our intuitive feeling, after seeing hundreds of calibration scores computed on real data, is that when a calibration score based on 200 responses is .007 or less, one should not reject the hypothesis that the assessor is perfectly calibrated.

Table 2  
Calibration Scores (x 1000)  
Experiment 1

Subject #	4	12	1	11	8	10	3	2	5	7	9	6	Mean
Session													
1	57	27	22	20	14	12	10	8	5	4	4	2	15.4
2	12	5	4	1	10	3	2	2	8	7	5	7	5.5
3	11	2	2	14	3	3	4	2	20	6	1	10	6.5
4	9	12	7	7	3	3	2	3	3	2	4	10	5.4
5	9	4	2	5	12	1	6	8	6	5	16	7	6.8
6	15	4	7	8	2	4	4	6	1	4	6	3	5.3
7	52	4	8	19	3	3	4	25	4	8	7	4	11.8
8	6	7	1	8	4	2	6	14	5	2	6	9	5.8
9	4	14	4	4	2	4	1	6	4	7	6	1	4.8
10	7	2	3	5	2	6	1	9	6	4	5	4	4.5
11	18	3	4	10	4	7	3	7	2	2	3	2	5.4

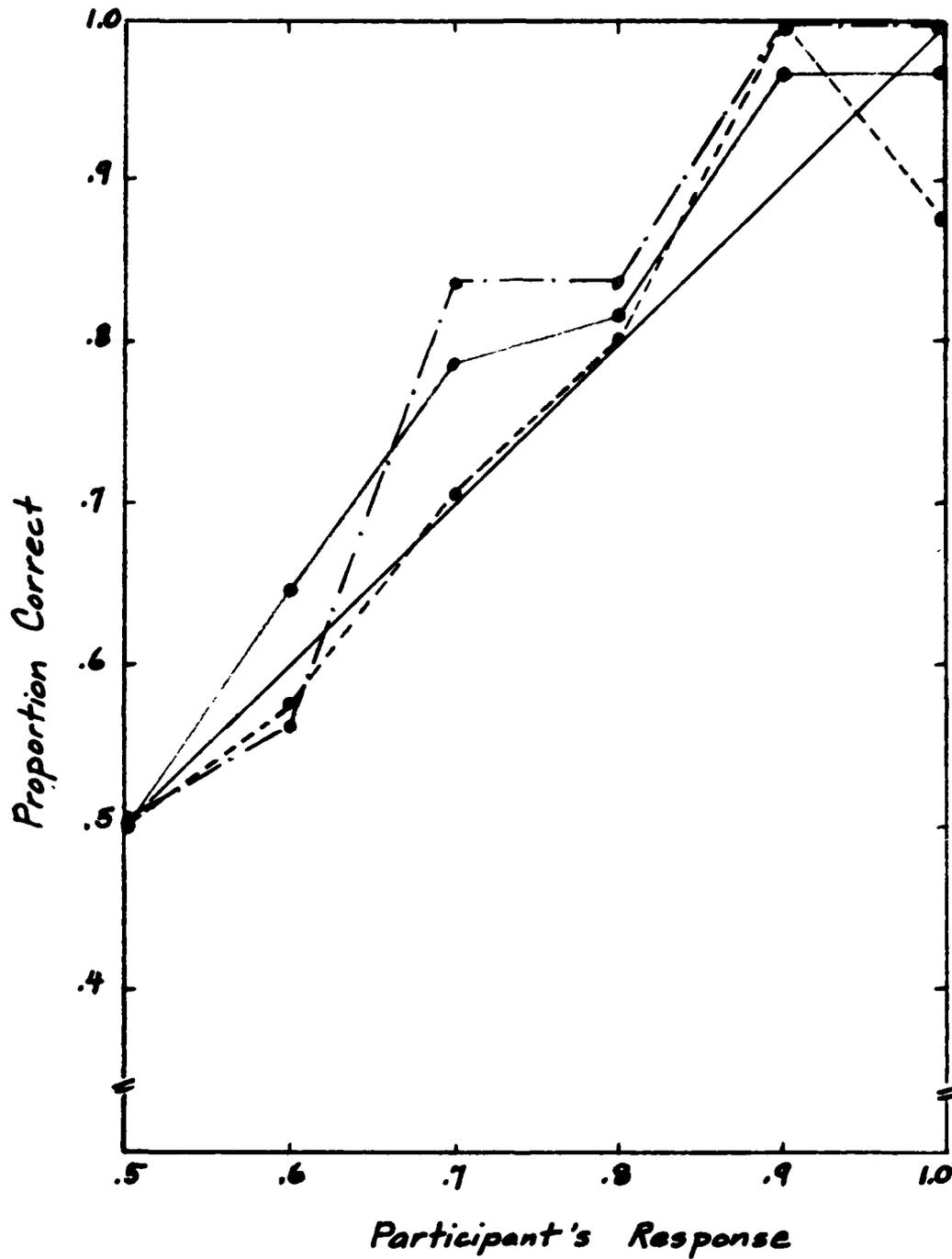


Figure 4a

Three Examples of Training Data  
with .002 Calibration Scores

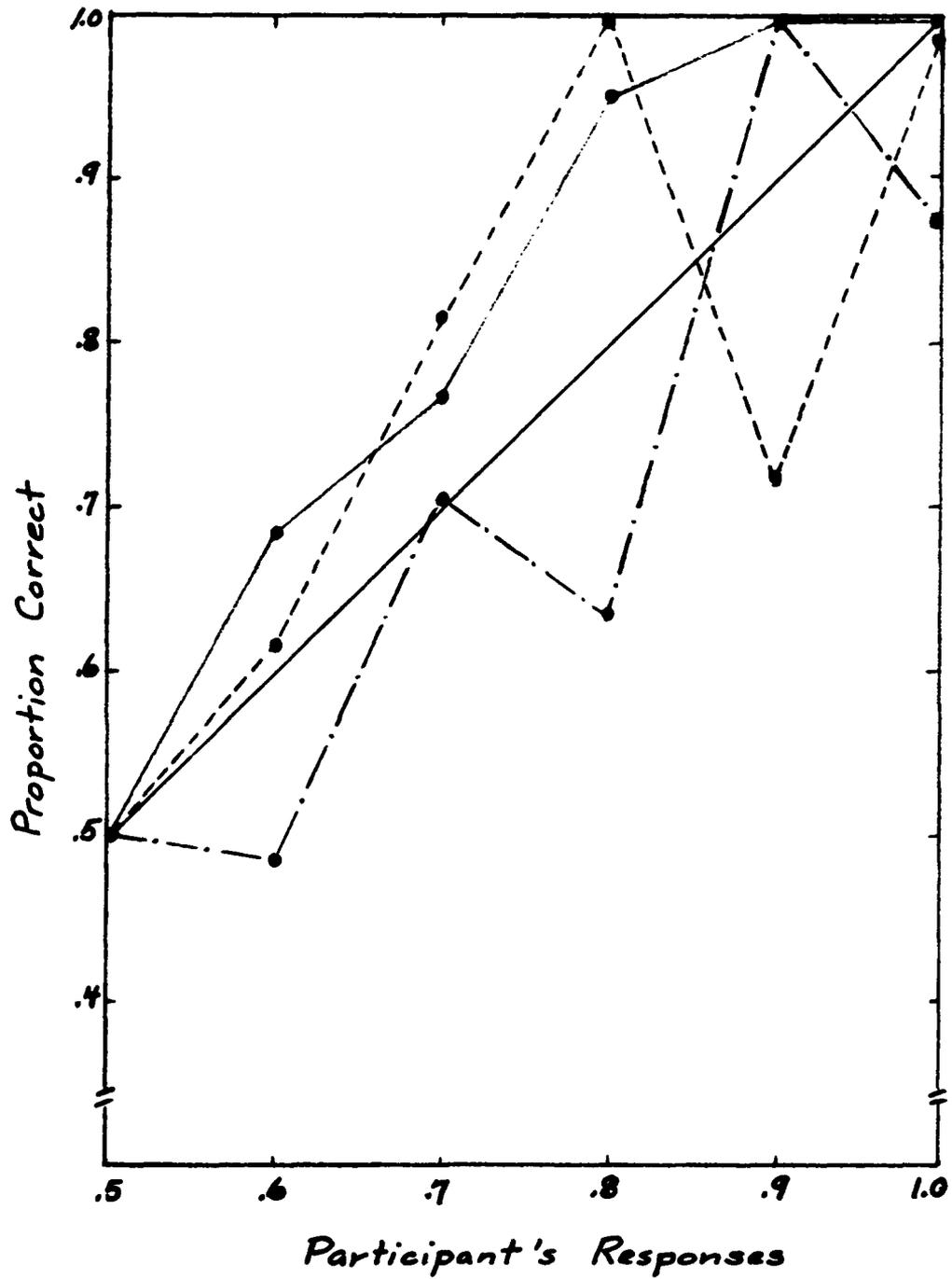


Figure 4b

Three Examples of Training Data  
with .005 Calibration Scores

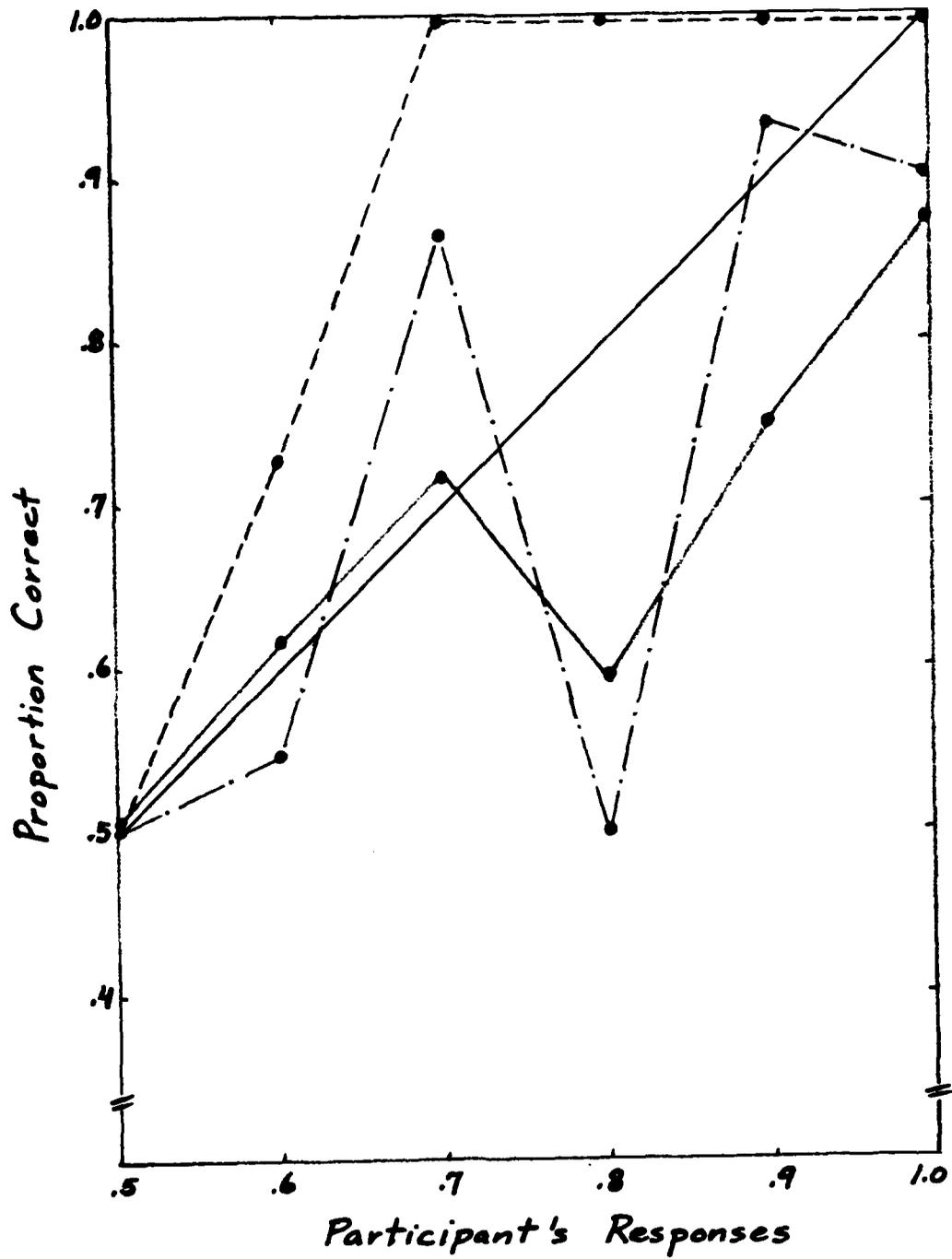


Figure 4c

Three Examples of Training Data  
with .010 Calibration Scores

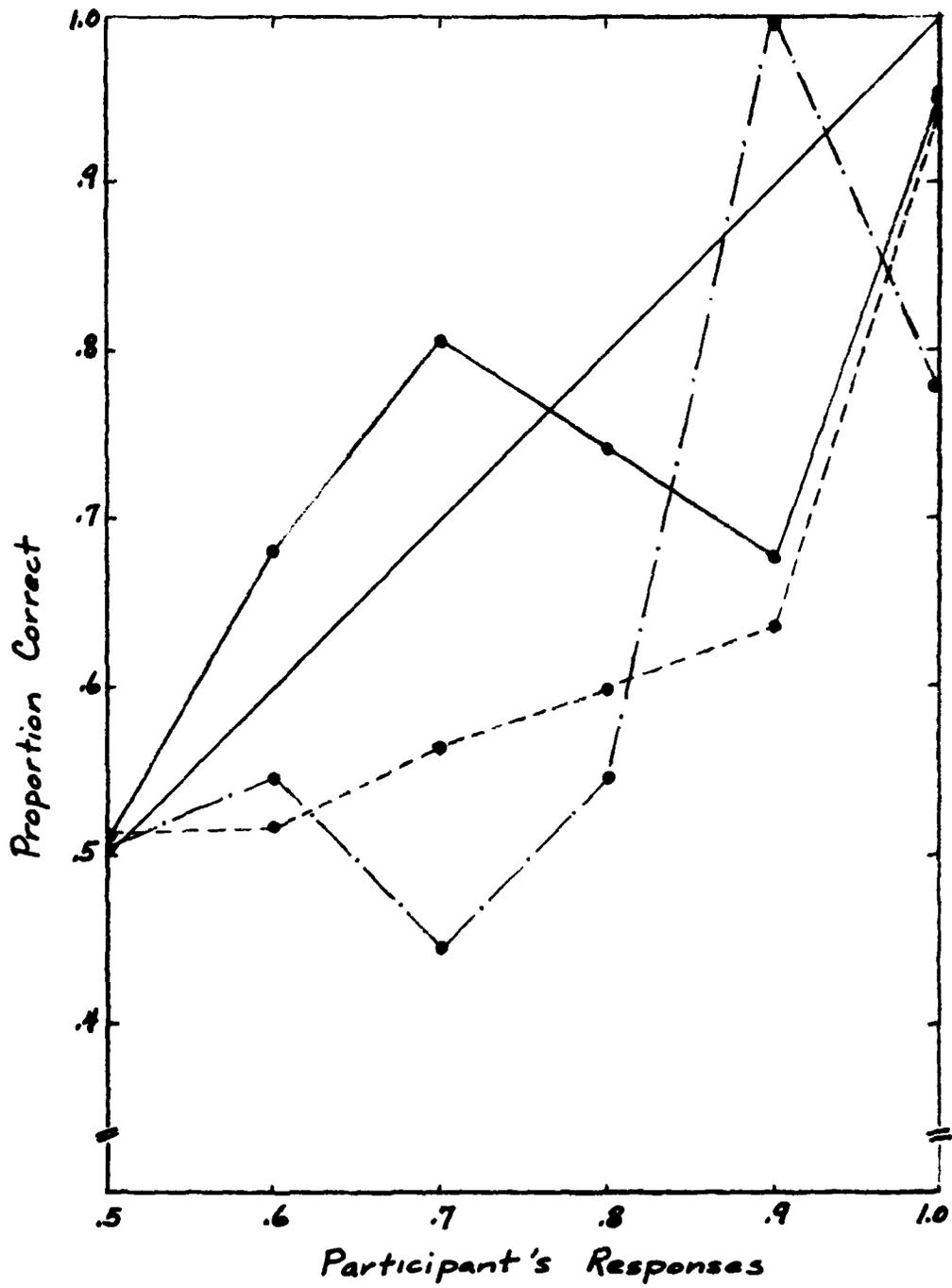
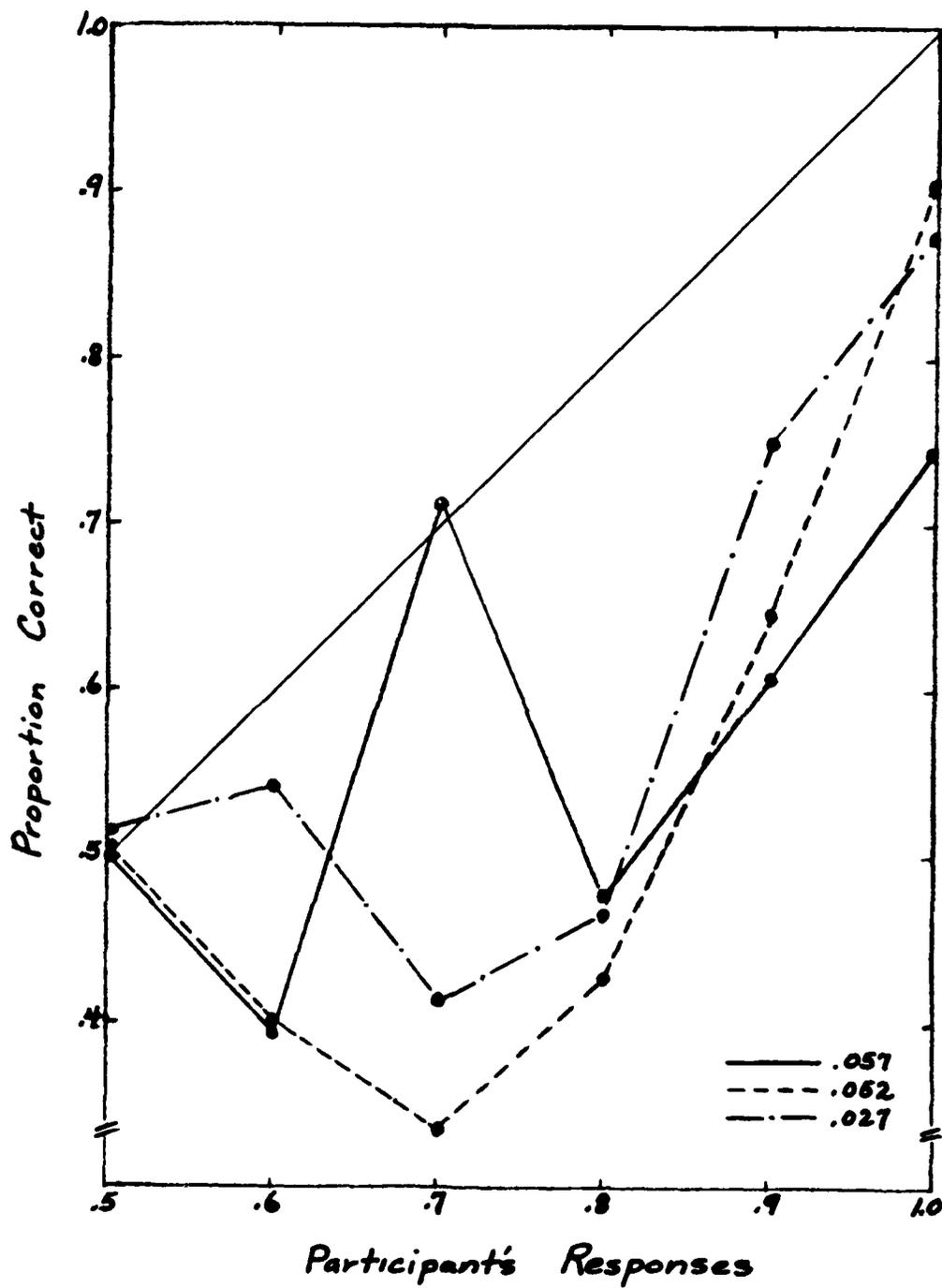


Figure 4d

Three Examples of Training Data  
with .020 Calibration Scores



Figures 4e

Three Examples of Training Data  
with Calibration Scores Exceeding .025

The data in Table 2 show improvement, with the mean calibration score shifting from .015 for Training Session 1 to .005 for Training Session 11. Calibration on Session 11 was superior to that on Session 1 for 11 participants and equal for the twelfth.

All the improvement came between the first and the second round of feedback. If an assessor improves, one would expect calibration scores on subsequent sessions to be better (lower). This is true for the first session: we calculated, for each participant, the proportion of sessions after the first session on which the calibration was better than the first session. Across the 12 participants, this proportion was .72. Indeed, the seven participants who scored .010 or more never did as badly again, so for them the proportion was 1.00. But this was not true for the second session. The proportion of later scores better than the participant's second score was .49; for the seven participants who had scored .010 or worse on the first session, this "improvement proportion" after the second session was .44. Similar "improvement proportions" of around .50 were found for all other sessions except Session 7, discussed below.

The interpolation of a particularly easy set of items (Session 5) had no effect on calibration. The mean percentage correct on Sessions 1-4 was 65.5; on Session 5 it was 75.0; this set was the easiest set for all 12 participants. Although the mean calibration score shifted slightly upward (i.e., worsened) on Session 5, individual participants were as likely to have higher scores as lower scores on Session 5 compared either with Session 4 or with Sessions 2 through 4. The interpolation of a particularly hard set (Session 7) did produce a decrement in calibration. The mean percentage correct on Session 7 was 57.8%; this round was the hardest for all 12 participants. Calibration was worse on Session 7 than on Session 6 for ten participants and equal for the other two participants.

One participant (#4) was terribly calibrated to begin with, made considerable improvement after Session 1, but remained worse than the rest of the group. The next six participants shown in Table 2 learned all they were going to learn during the first session. The other five subjects appeared to be well calibrated to begin with and contributed only random variation to the group results.

Nine of the 12 participants were overconfident on the first session, with overconfidence scores ranging from +.03 to +.21 (the latter was Participant 4), while 3 were slightly underconfident (-.01 to -.03). Mean overconfidence was +.063. Mean over/underconfidence across the 12 participants was close to zero on all other sessions except for Session 5, the easy session (mean, -.026; 9 participants underconfident, ranging from -.01 to -.10; 3 participants overconfident, +.01 to +.03) and Session 7, the hard session (mean, +.038; 8 participants overconfident, ranging from +.02 to +.16; 3 participants underconfident from -.01 to -.03). Participant 4 was overconfident on all sessions (range +.03 to +.21), while Participant 7 was underconfident on all sessions, but only slightly (range -.01 to -.04).

The resolution scores showed no trends across sessions. There was definitely no tendency for resolution to go down (worsen) as calibration went down (improved). The overall mean resolution score was .028. The session with the highest resolution was the easy session, 5 (mean, .034; 5 participants earned their best resolution on this session; none earned their worst), while the hard session, 7, had the lowest resolution (mean .018; 7 participants earned their worst resolution on this session; none earned their best).

One aspect of the participants' task that was emphasized by the experimenters was that they should get all of their 1.0 responses correct. Participants were unable to attain this goal consistently, although all attained it at least once in the eleven sessions, and one participant attained it seven times. By reducing the number of 1.0 responses from 22% of all responses on Session 1 (range: 9% to 48%) to 10% of all responses on Session 11 (range: 0% to 18%), the participants effected a modest increase in percentage of 1.0 responses correct from 91% (range: 74% to 98%) in Session 1 to 97% (range: 91% to 100%) in Session 11. Again, it appeared that most of the change occurred between the first and second sessions.

Participants' insights. During the feedback and discussion after every training session, we questioned the participants about what they were learning. They were rarely able to describe their cognitions; the most insightful was this comment:

"I don't know how to verbalize it, but there's some kind of a compartmentalization trip that's happening in my head about those categories. I'm beginning to feel the categories more than I did before, rather than just a blur from .5 on."

Often participants reported what they concentrated on during that session. For example, "I was more careful on my 1.0's and .9's" or "I tried to pay attention to my .6's and .7's."

Difficulties in using the intermediate responses of .6, .7 and .8 were frequently remarked upon:

"I think I'm still kind of unsure on how to use .7 and how to use .8 because I either feel I don't know it or I know it, and it's sort of hard to say how much I don't know something,"

"I think I'm judging a little bit better between my .5's and .6's--just what is a .6--but I don't think I really know what a .7 is yet," and

"I think probably the most significant thing to me that I did was kind of experimenting with the .7's and .8's and seeing that I really did need them, that I really had a use for them, that I wasn't just sticking a number in there when I felt it."

Occasionally a participant summarized what the different responses meant to them. For example:

".7's and .8's; I wonder why I even used them when I think about it because .5's to me represent no knowledge at all, one way or another, .6's represent a preference, and then .9's and 1.0's represent knowledge. .7's and .8's, I don't know why I'm using them," and

".5's, I had no idea, okay? I kind of had a little feeling but I decided that I really . . . didn't know after all. .6's I had a little bit stronger feeling so I thought well, maybe I'll go with it. .7's and .8's are kind of wishy-washy, or . . . I can't talk about why I did that, but I just felt 'kind of' but I absolutely wasn't certain. A lot of my .99's I was really certain but I was blown away yesterday when I got two of them wrong as 1.0's, so I figured I don't know everything; I'll use a .99 instead of a 1.0."

Generalization tasks. Did the participants' improvement on the training tasks lead to improvement on the post-test tasks? Table 3 contrasts pre-test and post-test statistics on the four two-alternative, half-range tasks: shapes, handwriting samples and the general knowledge items in Tasks 3 and 5. There was marked improvement on the shapes and general knowledge tasks, but none on the handwriting task. Note that as calibration scores improved (decreased) from pre-test to post-test, resolution scores did not fall. The gain from decreased calibration scores was not offset by a decrease in resolution scores.

Table 3  
Pre- and Post-Test Results for Two-Alternative Items  
Experiment 1

	Pre- test	Post- test	#Ss im- proved	Pre- test	Post- test	#Ss im- proved
	Task 1 Handwriting			Task 2 Shapes		
Mean Calibration	.023	.026	7	.013	.005	8
Mean Resolution	.011	.013	7	.017	.020	7
Mean Brier	.252	.254	7	.223	.217	6
Proportion Correct	.576	.556	5	.648	.634	5
Mean Response	.629	.587	-	.715	.627	-
Over/underconfidence	.054	.032	7	.066	-.007	8
	Task 3 General Knowledge			Task 5 General Knowledge		
Mean Calibration	.025	.010	10	.023	.010	11
Mean Resolution	.024	.029	7	.027	.032	7
Mean Brier	.232	.209	10	.224	.209	9
Proportion Correct	.636	.639	7	.642	.629	5
Mean Response	.744	.671	-	.732	.665	-
Over/underconfidence	.107	.032	10	.090	.036	8

Table 4 contrasts pre-test and post-test for the four-alternative, general-knowledge questions, Task 4. Substantial improvement in calibration is observed here. This improvement was attained by decreasing the use of the extreme responses of 0.0 and 1.0, as can be seen in the distributions presented at the bottom of the table.

Table 4  
Four-Alternative Questions  
Experiment 1

	Pre-test	Post-test	# Ss Improved
Mean Calibration:	.017	.007	11
Mean Resolution:	.029	.033	9

Distribution of Responses						
	Response Category					
	0	.01-.19	.20-.29	.30-.49	.50-.99	1.00
Pre-test	.33	.09	.28	.13	.11	.06
Post-test	.24	.09	.39	.14	.09	.05
# of Subjects decreasing usage (of 12 Ss)	9	5	3	5	9	10

Table 5 contrasts pre-test and post-test results on the uncertain quantities task. Although this task also involved general knowledge items, there was no improvement by any measure. The surprise index remained at a discouragingly high 40% (instead of the appropriate 2%). This percentage is typical of previous findings (Lichtenstein, Fischhoff & Phillips, 1977). At both times, the interquartile range included only about 35% of responses. Nor was there any general shift of values upward or downward from pre-test to post-test. The participants tended to suppose that the true answer was smaller than it really was, so that the majority of surprise answers fell above the .99 fractile.

Table 5  
Uncertain Quantities:  
Percentages of Responses

	Pre-test	Post-test
Surprises	41.1	40.0
Within Interquartile Range	32.3	36.6
True higher than .99 fractile:		
As percentage of all responses	27.4	25.3
As percentage of surprises	66.6	63.2

### Discussion

The training sessions improved the calibration of most participants on subsequent sessions; the only participants who showed no learning were those who were apparently well calibrated to start with. However, all measurable improvement came with the first round of training. Training also improved performance on some of the generalization tasks, although none was noted on the uncertain quantities or handwriting tasks.

The dissimilarity of response modes between the uncertain quantities task and the training task coupled with the absence of any discussion of the relationship between calibration in the two contexts may account for this failure to generalize. The failure to generalize with the handwriting samples is more difficult to explain, both because it used the same two-alternative, half-range format and because generalization was found with the (perhaps more dissimilar) shapes task.

Although effective, the present training procedure is both arduous and expensive. The fact that all measurable improvement came after the first round suggests that it may also be unduly long. Experiment 2 tested this hypothesis by using an abbreviated test procedure that deleted Training Sessions 3-10. The participants were given the first set of items, the second set and then the first set again, along with pre-tests and post-tests. It need not be a foregone conclusion that a shortened sequence will prove equally successful. Although Sessions 3-11 showed no noticeable improvement, they may have provided valuable practice and the opportunity to explore subjective feelings of certainty. The interpolated easy and hard sets (Sessions 5 and 7, respectively) may have brought home principles of probability assessment that might not have been otherwise understood.

### Experiment 2

#### Method

Experiment 2 differed from Experiment 1 only in the deletion of training Sessions 3-10 and the uncertain quantities and handwriting analysis tasks. Since the participants showed no improvement on these two generalization tasks with the extended training series of Experiment 1, there was no reason to expect any improvement in Experiment 2.

All instructions were the same as in Experiment 1. The feedback after the training sessions was mostly the same, but the discussion of the Brier score and its partitions was abbreviated because the participants would not have sufficient opportunity to see how these scores varied over time; we were concerned that they might place undue emphasis on what were really random variations of these scores. Most of the discussion focused on the calibration plot and the need to get all data points falling on the diagonal.

Participants. Twelve participants were recruited as before. Their ages ranged from 17 to 37. Three were exceptionally bright high school

students (#13, #14, #18), 4 were college students (#15, #17, #21, #23), and 3 were graduate students (#19, #20, #22). The two participants who were not in school had a high school education (#24) and a masters degree (#16) respectively.

### Results

Training sessions. Table 6 gives the calibration scores for the training sessions. All seven participants scoring .007 or higher on the first session improved their scores when they repeated this set of items as Session 3. The mean calibration scores across all 12 participants for the three sessions were .010, .005 and .007. Again, it appears that all learning took place between Sessions 1 and 2.

Table 6  
Calibration Scores (x 1000)  
Experiment 2

Subject #	18	15	21	13	23	24	17	16	14	19	20	22	Mean
Session													
1	26	19	18	12	11	8	7	6	5	3	3	3	10.1
2	5	6	2	5	5	4	4	5	3	10	5	6	5.0
3	6	9	12	3	2	4	6	17	6	7	6	5	6.9

All 12 participants were overconfident on Session 1 (mean = +.05; range +.01 to +.13), while 9 were overconfident on Session 3 (mean across 12 participants was +.03, range -.01 to +.09).

Resolution scores did not significantly change during the three training sessions. The mean resolution for Session 1 was .031 (range .018 to .042); for Session 3 it was .035 (range .027 to .053).

Improvement via using fewer 1.0 responses was modest. In Session 1, 24% of all responses were 1.0 (range 9% to 48%); in Session 3, 18% were 1.0 (range 6% to 52%). The overall percentage correct for 1.0 responses was .94 (range .89 to 1.00) on Session 1 and .97 for both Sessions 2 (range .90 to 1.00) and 3 (range .91 to 1.00). Ten participants used fewer 1.0's on the last session than the first; 8 participants got a higher percentage correct.

Generalization tasks. Table 7 shows pre-test and post-test statistics for the three two-alternative tasks: shapes and the general knowledge items in Tasks 3 and 5 (these participants were not given the handwriting task). These data are highly similar to the parallel data from Experiment 1 shown in Table 1. Generalized improvement in calibration did occur, with no decrement in resolution. Table 8 shows performance on the four-alternative task. The results are again highly similar to Experiment 1 (see Table 4).

Table 7  
Pre- and Post-Test Results for Two-Alternative Items  
Experiment 2

	Task 2 Shapes			Task 3 General Knowledge			Task 5 General Knowledge		
	Pre- test	Post- test	#Ss im- proved	Pre- test	Post- test	#Ss im- proved	Pre- test	Post- test	#Ss im- proved
Mean Calibration	.019	.005	8	.020	.007	10	.014	.006	9
Mean Resolution	.015	.018	9	.023	.026	7	.028	.030	6
Mean Brier	.230	.215	10	.216	.198	12	.207	.195	10
Proportion Correct	.650	.633	5	.660	.666	7	.651	.659	7
Mean Response	.703	.635	-	.748	.696	-	.716	.675	-
Over/underconfidence	.053	.002	8	.088	.031	11	.065	.016	9

Table 8  
Four-Alternative Questions  
Experiment 2

	Pre-test	Post-test	# Ss Improved
Mean Calibration:	.017	.008	11
Mean Resolution:	.037	.040	11

Distribution of Responses						
	Response Category					
	0	.01-.19	.20-.29	.30-.49	.50-.99	1.00
Pre-test	.36	.08	.27	.12	.10	.07
Post-test	.30	.08	.36	.12	.08	.06
# of Subjects decreasing usage (of 12 Ss)	11	7	3	7	9	10

### Discussion

These two experiments have shown that people who are not well calibrated to begin with can be taught to be well calibrated with intensive performance feedback after a single session of 200 items. This improvement occurred without the participants ever learning the true answers to any items.

However, this training resulted in improvement on only some of the tasks on which the participants were not trained. Generalization of training failed completely for the most dissimilar task, the assessment of probability distributions for a series of uncertain numerical quantities. Generalization of training also failed for a task we had supposed was not terribly different from the training task: discriminating European from American handwriting. This failure should serve to warn calibration trainers that generalization of training cannot be assumed.

Almost half of our 24 participants appeared to be well calibrated on their first training session. We know of no other experiment in which individual differences in calibration have been studied, so we cannot say whether this was due to the extensive instruction and experience (albeit without feedback) that these participants received before the first training session, whether this unexpectedly high proportion of good probability assessors exists in the population at large, or whether the individuals in this study were unusual.

We did not explore the use of item-by-item feedback using a proper scoring rule, as recommended by Shuford and Brown (1975). While this technique might improve the efficiency of the training, the additional information provided (specifically, knowledge of the correct answer) might only confuse the participant and retard learning. Further research can resolve this issue.

Before recommending adoption of the present training procedure, two questions need answers. One is: To what extent will learning generalize from the items used in the training sessions to the questions encountered in the trainee's professional activities? On the one hand, we are moderately encouraged by the substantial generalization found here. On the other hand, we do not fully understand the reasons for the utter failure of generalization with handwriting samples or uncertain quantities. What other tasks would encounter similar difficulties?

The second question is whether the training program can be abbreviated further, making it less arduous and more cost effective. For example, one could explore the following possibility: Have trainees complete one session in a non-computerized group administration. Instead of scoring them individually, present the typical Session 1 result from the many similar assessors who have completed it earlier, saying "This is a very good guess at how your curve will look and this is what you have to do to improve." Will such short training be effective or will the feedback be rejected with the claim "I'm not like that"?

## References

- Adams, J. K. & Adams, P. A. Realism of confidence judgments. Psychological Review, 1961, 68, 33-45.
- Adams, P. A. & Adams, J. K. Training in confidence judgments. American Journal of Psychology, 1958, 71, 747-751.
- Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 1950, 75, 1-3.
- Choo, G. T. G. Training and generalization in assessing probabilities for discrete events. Brunel Institute of Organisation and Social Studies, Technical Report 76-5, September, 1976.
- de Finetti, B. La prevision: Ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincare, 1937, 7, 1-68. English translation in: H. E. Kyburg, Jr. & H. E. Smokler (eds.), Studies in subjective probability. New York: Wiley, 1964.
- Lichtenstein, S. & Fischhoff, B. Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 1977, 20, 159-183.
- Lichtenstein, S., Fischhoff, B. & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. deZeeuw (eds.), Decision making and change in human affairs. Dordrecht, Holland: D. Reidel, 1977.
- Murphy, A. H. Scalar and vector partitions of the probability score: Part I. Two-state situation. Journal of Applied Meteorology, 1972a, 11, 2, 273-282.
- Murphy, A. H. Scalar and vector partitions of the probability score: Part II. N-state situation. Journal of Applied Meteorology, 1972b, 11, 8, 1183-1192.
- Murphy, A. H. A new vector partition of the probability score. Journal of Applied Meteorology, 1973, 12, 4, 595-600.
- Murphy, A. H. A sample skill score for probability forecasts. Monthly Weather Review, 1974, 102, 1, 48-55.
- Murphy, A. H. & Winkler, R. L. Can weather forecasters formulate reliable probability forecasts of precipitation and temperatures? National Weather Digest, 1977a, 2, 2, 2-9.
- Murphy, A. H. & Winkler, R. L. Reliability of subjective probability forecasts of precipitation and temperature. Journal of the Royal Statistical Society Series C (Applied Statistics), 1977b, 26, 1, 41-47.
- Murphy, A. H. & Winkler, R. L. Probabilistic tornado forecasts: Some experimental results. Tenth Conference on Severe Local Storms, October 18-21, 1977, Omaha, Nebraska. Published by American Meteorological Society, Boston, Massachusetts, 1977c.
- Phillips, L. D. Bayesian statistics for social scientists. London: Nelson & Sons, Inc., 1977.
- Pickhardt, R. C. & Wallace, J. B. A study of the performance of subjective probability assessors. Decision Sciences, 1974, 5, 347-363.
- Schlaifer, R. Computer programs for elementary decision analysis. Boston: Harvard University Press, 1971.
- Shuford, E. & Brown, T. A. Elicitation of personal probabilities and their assessment. Instructional Science, 1975, 4, 137-188.

## Footnotes

Our deepest thanks to Gerry Hanson for conducting this experiment, to Barbara Combs and Peggy Roecker for compiling the enormous item pool needed, and to Ruth Phelps, Stanley Halpin, Edgar Johnson and Paul Slovic for their comments on this project.

1. These calibration scores, and all subsequent analyses, are based on data grouped into six categories (.5-.59, .6-.69, ... , .9-.99, 1.00). In addition, all responses of .5 were rescored so that exactly half of them were correct. Many of our subjects chose the correct alternative randomly when they assigned a probability of .5. Since there was often a substantial number of such responses, the proportion correct had a large effect on the overall calibration score. When subjects chose their responses randomly, variations in this proportion correct reflected random fluctuations. In order to keep such fluctuations from exerting too large an influence on results, .5 responses were treated as being half correct in all reported analyses. This has the effect of improving calibration scores in almost all cases. Analyses done with the unaltered data produce the same general conclusions as stated in the text, but are somewhat messier.

## Appendix

## Instructions and Sample Items for Experiment 1

Preliminary Instructions

I am Gerry Hanson and I will be the experimenter during this experiment.

To start with, I'm going to explain why you're here and what we are trying to do in this experiment. Then I'll explain the task in greater detail, and finally today you'll be trying your hand at it. Feel free to interrupt me at any time for questions.

The overall goal of the experiment is this: We are going to train you, to see if you can become more accurate in your probability judgments, and, in the process, become "well calibrated." I will explain what that means later on in this session. Your task won't involve any mathematical skills; we want to train you to express your own intuitions and judgments in assessing probabilities.

There are three main parts to this experiment. In part one, you will be making probability judgments without any training. We want to get a measure of how well you do without training. This will be a paper and pencil task, so we can schedule more than one person at a time. In part two, you will be trained in probability judgments. We'll give you feedback about how you're doing, and try to get you to do better. This task will be done on the computer terminal, so we can schedule only one person at a time. In part three, you will be tested again without training to see if you have learned what we've tried to teach you. Again, this will be a paper and pencil task for which we can schedule more than one person at a time.

The length of time for the whole experiment will be approximately 11 sessions of about one hour each, depending on your individual speed. At first you will no doubt be somewhat slow, but as time goes on, you will find that your speed will increase. We'd like you to schedule a session every day, or almost every day.

It is extremely important to us that you complete the experiment. Otherwise, we can't use your data. It may become very tedious work, so if you have any doubts, it would be better for us if you drop out early rather than late in the experiment. Quite frankly, we need subjects who are careful and hard working.

Payment will be made once a week at \$2.65 per hour. All those who complete the experiment will have their pay re-computed at \$3.65 per hour as a bonus for finishing the whole experiment. Thus for every hour you work we'll put \$1 "into escrow" for you, to be given to you if you complete the experiment. The longer you stay in, the more you stand to lose by quitting.

Individual appointments will be set up for subsequent sessions and you will be given an appointment card like this (show card) to help

remind you of the day and time of your next appointment. Feel free to call me to reschedule appointments whenever necessary.

We will ask you to sign a consent form, agreeing to participate in this experiment. Also, we may decide to tape record some of the sessions, in which we ask you for your reactions to the task. We'll want to remember what questions you had, and check to make sure we said the same thing to every person. These tapes, and indeed, all of your data, will be held in confidence. When we publish a report of the experiment, your names will not be included. Do you have any objection to being tape recorded?

I assure you there are no tricks in this experiment. It will all unfold as it happens, so that you will know all by the end of the experiment. If you wish, you will be able to get a report on the experiment, but it may take a year before it's finished. Do ask questions whenever you want. When you ask questions and we sound vague, it is because we don't know yet, or we know and we don't want to tell you yet. In that case, we will be honest and tell you just that. We ask that you please limit your talking with others about the experiment, especially fellow subjects. Also, do not look things up in reference books of any kind, because we want your true "gut feeling" in response to the questions.

Speaking of questions, I'll stop now to see if you have any up to this point.

During the middle part of the experiment, you'll be seeing a huge number of "items"--questions with two possible answers, like:

- Crater Lake was formed by
- (a) the impact of a meteorite
  - (b) a volcanic eruption

One of the two possible answers is always correct; the other is always wrong. Your task is to decide which answer is correct and state the probability that you have chosen the right answer.

In the first and last parts of the experiment, you'll see some variation--slightly different tasks. We'll explain these when we come to them.

Now I would like to explain what probability assessment is. Because we have found these judgments are not always easy to make, I'd first like to spend some time discussing the concepts of probability judgment with you. Then I'll explain what you have to do to be "well calibrated."

Probabilities are numbers between 0 and 1 that express uncertainty. Let's take the above example: Crater Lake was formed by (a) the impact of a meteorite or (b) a volcanic eruption. Suppose you are not sure of the answer. You think the answer you have chosen is correct, but you are not certain. The question is, how certain are you? You will have to make a probability assessment that expresses your degree of certainty. If you give a probability of .8, you're saying there are about 8 chances out of 10 that your answer is correct. If you give a probability of 1.00,

you know for certain you have chosen the correct answer; if you say .5, that indicates you're completely unsure whether your answer was correct. The more certain you are that you are right, the larger the number you should choose. But what number should you choose? This is the nub of the problem. We are asking you to do a very difficult task. We want you to examine your own "gut feelings" of certainty and uncertainty and translate those feelings into a probability number.

During the main part of the experiment, on the computer terminal, you will be shown a number of statement of fact with two possible answers, one of which is right and one wrong. First, you will decide which alternative you think is correct (please select one answer even when you are completely unsure which is correct; the computer is fussy about these things). Next, decide what the probability is that your answer is correct. This probability can be any number from .5 to 1. It is your degree of certainty about the correctness of your answer.

Why are we forbidding you to use a probability less than .5? Because your first task was to choose the alternative from the two given that is most likely to be correct. If, after doing that, you assign a probability of, say, .3 to your chosen alternative, that would logically imply that you believe there's a .7 chance the one you didn't choose is the correct one. That means that one, the one you didn't choose, is more likely to be the correct answer. So a probability of less than .5 suggests that you goofed the first step, by not choosing the alternative which is most likely correct.

Technically speaking, you can use any number you want, like .703 or .832319 (providing it is in the range .5 to 1.0), but you will find out very soon that you are not capable of making subtle discriminations such as deciding whether to give a .703 or a .704. You probably won't want to use numbers with a lot of fancy extra digits.

If you respond that the probability is .6, it means that you believe that there are about 6 chances out of 10 that your answer is correct. A response of 1.00 means that you are absolutely certain that your answer is correct. A response of .5 means that your best guess is as likely to be wrong as right. And how do you decide whether to say .6 or .7? You have to review all the information you have in your head about the item in question, and gauge how confident you are about the correctness of your choice.

The key thing we want you to learn in this experiment, which we call being "well calibrated," is to learn how to translate your own internal feelings of certainty, uncertainty, and partial certainty into the precise language of probability numbers. We want you to be well calibrated in the same sense that a thermometer is well calibrated. When a calibrated instrument says 32° F, it means the same thing every time, and it means something very specific: The temperature at which water freezes.

Likewise, you should mean the same thing every time you say .5. That means (a) I'm completely uncertain between the two possible answers

and (b) on average, I have a 50% chance of getting this one right.

We can find out whether you are well calibrated. Suppose we have a subject, Paul, who responds to 100 different items. Over that set of 100 responses, he said ".5" 30 times, and ".6" 10 times, and so forth, as shown below:

<u>Paul said</u>	<u>How many times</u>
.5	30
.6	10
.7	10
.75	20
.8	0
.9	10
1.0	20
	—
Total	100

(That .8, which Paul never used, was thrown in to remind you that you don't have to use all the one-digit numbers if you don't want to. That .75 was included to remind you that you don't have to limit your responses to one-digit numbers. You can use .54 or .99 if you want to.)

Now suppose we look at how many times that Paul said ".5" and was right (that is, selected the correct answer), and how many times Paul said ".5" and was wrong, and so forth for each of the different probabilities he used, as shown below:

<u>Paul Said</u>	<u>How Many Times</u>	<u>Times Right</u>	<u>Times Wrong</u>	<u>Percent Correct</u>
.5	30	15	15	50
.6	10	6	4	60
.7	10	7	3	70
.75	20	15	5	75
.9	10	9	1	90
1.0	20	20	0	100
Totals	100	72	28	72%

We can calculate, as shown above, the percent correct for each different response. If Paul's data looks like this, he is perfectly calibrated, because his response is always equal to the percent correct. For exactly 70% of all the times he said ".7," he was right, and 30% of the time, he was wrong. He got half of his ".5" responses right, and all of his "1.0" responses right, and so on.

Now let's look at another subject, Baruch, who, by incredible coincidence, gave the same number of different responses as Paul did. But Baruch's calibration data looks like this:

Baruch <u>Said</u>	How Many <u>Times</u>	Times <u>Right</u>	Times <u>Wrong</u>	Percent <u>Correct</u>
.5	30	18	12	60
.6	10	8	2	80
.7	10	8	2	80
.75	20	13	7	65
.9	10	9	1	90
1.0	20	16	4	80
Totals	100	72	28	72%

Baruch was not well calibrated. For only one class of his responses was he "right on": he did get exactly 90% of his ".9" responses correct. But otherwise, he didn't use the probabilities the way he should have. Across the 30 times he said ".5," he got 60% of them right, instead of the desired 50%. This is a kind of underconfidence; he knew more than he thought he knew. At the other extreme, he was wrong too often when he said "1.0"--he got only 80% right (to be perfectly calibrated, you can never be wrong when you say "1.0"). This is overconfidence; he knew less than he thought he knew.

Notice that Paul and Baruch both got, overall, 72% of their answers correct. They both have the same degree of knowledge. But knowledge is independent of calibration. So don't worry about how much you know and don't know in this experiment--we don't care much about that. The items we have chosen for this experiment will surely include some items you know very well, and some items you just don't know at all. We hope we've selected a lot of items you're not completely sure about, because those are the items on which you'll get a chance to practice your skill at assigning probabilities.

By the way, I ought to warn you that we will never at any time during the experiment be telling you the right answer to any item. I hope that won't frustrate you. After you've completed the experiment, you're free to find that out. In all these sessions, you'll be shown at least 3,000 different items, so perhaps after a while, you'll get used to not learning the correct answers to the items as you respond to them.

Don't worry if you don't know the answers to some items. We're not so much interested in how much you know as we are interested in how well you can express your own feelings of knowing or not knowing. Complete every item; try not to miss any. If you have a change of heart, you can, and should, go back and change an answer. This is about all I have to tell you about the experiment. So we'll stop again to answer questions and sign the consent form (on the next page). Then the session will end with you trying your hand at the first 100 items.

#### Consent Form

Experiment: Calibration Study I  
 Experimenters: Gerry Hanson and Sarah Lichtenstein

I have been informed of the nature of this experiment and have agreed to participate. I understand that, if I so desire, I can leave the experiment at any time. I will be paid \$2.65 per hour at the end of each week. If I complete the experiment, my wages will be re-computed at \$3.65 per hour. I also agree to be tape recorded.

Signed \_\_\_\_\_  
Date \_\_\_\_\_

### Task 1

#### Task 1 Instructions

In this task you will see 100 samples of a simple handwritten sentence (in Latin):

MENSA MEA BONA EST

Your job is to judge whether each sentence was written by an American or a European.

First indicate whether you think the person who wrote the sample is an American or a European, by circling A or E on the answer sheet.

Then give a probability response from .5 to 1.0 that expresses your degree of belief in the correctness of your answer. Please try to be well calibrated. Feel free to refer to the instructions we have just gone through if you wish. Have you any questions? If not, go ahead and start. Since we have only one copy of each sample, you'll have to pass them around. Please BE SURE you match the item number in the upper right corner with the item number on the answer sheet.

#### Task 1 Sample Items

1. *Mensa mea bona est*
2. *Mens a mea Bona Est*
3. *Mensa Mea Bona Est*

#### Task 1 Answer Sheet [correct answers for the three sample items are shown]

Circle E for European or A for American and write the probability that you are correct in the space provided.

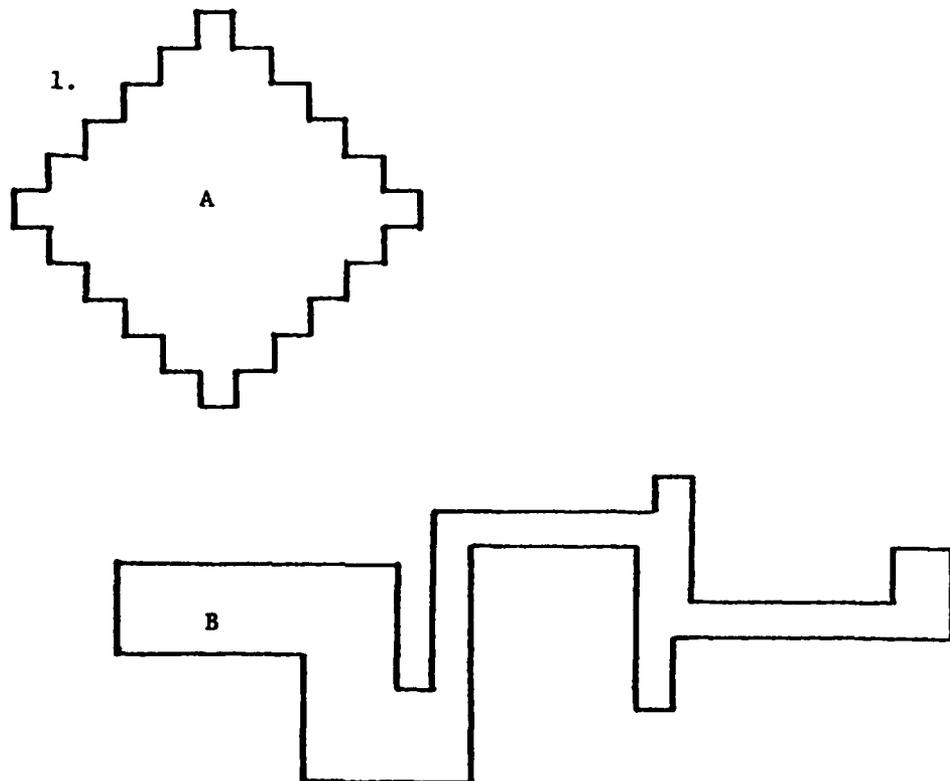
- |                                       |                                    |       |      |   |       |      |   |       |
|---------------------------------------|------------------------------------|-------|------|---|-------|------|---|-------|
| 1. <input checked="" type="radio"/> E | A                                  | _____ | 4. E | A | _____ | 7. E | A | _____ |
| 2. <input checked="" type="radio"/> E | A                                  | _____ | 5. E | A | _____ | 8. E | A | _____ |
| 3. E                                  | <input checked="" type="radio"/> A | _____ | 6. E | A | _____ | 9. E | A | _____ |

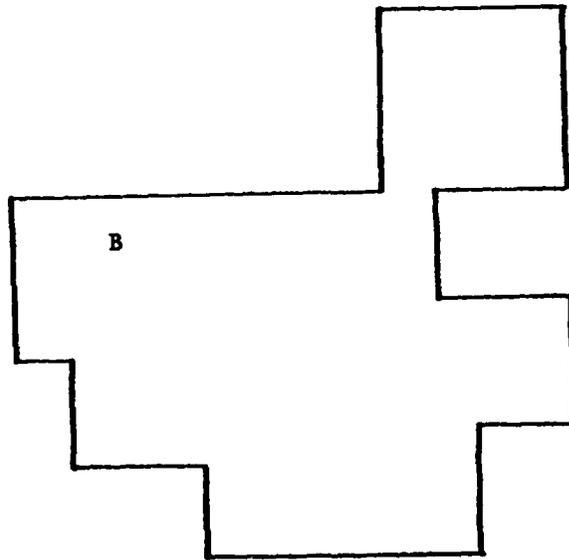
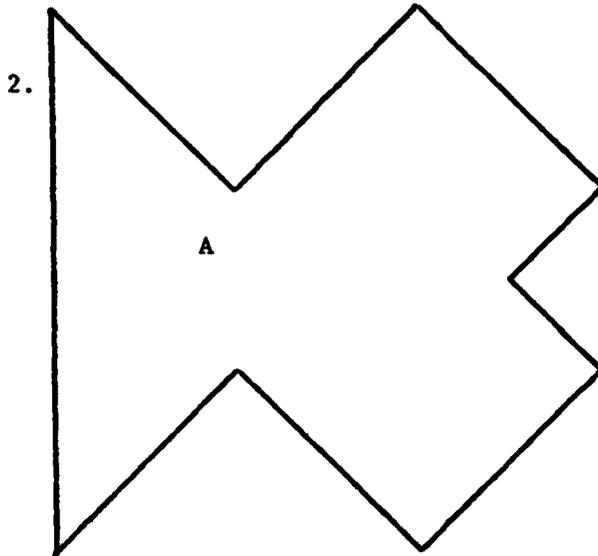
## Task 2

Task 2 Instructions

This task is composed of 200 items. Each item shows two irregular shapes on the same page, labeled A and B. One of the shapes is larger than the other. First, decide which item is LARGEST and mark your answer on the answer sheet. Second, decide what the probability is that your answer is correct. This probability can be any number from .50 to 1.00. It can be interpreted as your degree of certainty about the correctness of your answer. For example, if you respond that the probability is .60, it means that you believe that there are about 6 chances out of 10 that your answer is correct. A response of 1.00 means that you are absolutely certain that your answer is correct. A response of .50 means that your best guess is as likely to be right as wrong. Don't estimate any probability below .50, because you should always be picking the alternative that you think is more likely to be correct. Write your probability in the space provided on the answer sheet.

Try to be well calibrated in your responses.

Task 2 Sample Items



Task 2 Answer Sheet [answers to the two sample items are shown]

Please decide which figure (A) or (B) has the greater area, and then mark the probability (between .5 and 1.00) that you are correct.

- |    |                                    |   |       |    |   |   |       |    |   |   |       |
|----|------------------------------------|---|-------|----|---|---|-------|----|---|---|-------|
| 1. | <input checked="" type="radio"/> A | B | _____ | 4. | A | B | _____ | 7. | A | B | _____ |
| 2. | <input checked="" type="radio"/> A | B | _____ | 5. | A | B | _____ | 8. | A | B | _____ |
| 3. | A                                  | B | _____ | 6. | A | B | _____ | 9. | A | B | _____ |

## Task 3

Task 3 Instructions

This task is composed of 200 items. Each item is a brief phrase followed by two alternatives, labeled A and B. Only one of the alternatives is correct. Read each item and the two alternatives carefully. First, decide which alternative you think is correct, and mark your answer on the answer sheet. Please indicate an answer, either A or B, even when you are completely unsure which is correct. Second, decide what the probability is that your answer is correct. This probability can be any number from .50 to 1.00. It can be interpreted as your degree of certainty about the correctness of your answer. For example, if you respond that the probability is .60, it means that you believe that there are about 6 chances out of 10 that your answer is correct. A response of .50 means that your best guess is as likely to be right as wrong. Don't estimate any probability below .50, because you should always be picking the alternative that you think is more likely to be correct. Write your probability in the space provided on the answer sheet.

Try to be well calibrated in your responses.

Task 3 Sample Items

1. The only bachelor United States president was
  - A. James Madison
  - B. James Buchanan
2. A rudder is located on an airplane's
  - A. Tail
  - B. Wings

Task 3 Answer Sheet [answers to the two sample items are shown]

Please circle the answer (A or B) you think is correct and write the probability that you are right in the space provided.

- |   |              |              |
|---|--------------|--------------|
| 1. <input checked="" type="radio"/> A B _____ | 4. A B _____ | 7. A B _____ |
| 2. <input checked="" type="radio"/> A B _____ | 5. A B _____ | 8. A B _____ |
| 3. A B _____                                  | 6. A B _____ | 9. A B _____ |

## Task 4

Task 4 Instructions

In this task you will find 200 statements, each with four possible answers. This time, we want you to give a probability response expressing your sense of certainty or uncertainty for all four alternatives.

You may use any number from .0 to 1.00 but the total of all four responses must sum 1.00. If you are completely sure one of the alternatives is correct, you would use 1.00 for that alternative and 0 for the other three. If you are completely unsure, your response should be .25 for each of the four answers. Assign a 0 to any alternative you are sure is wrong. If you can surely eliminate one alternative but feel entirely uncertain about all the other three, assign 0,  $.33\frac{1}{3}$ ,  $.33\frac{1}{3}$  and  $.33\frac{1}{3}$ .

Here's an example:

Which is the greatest distance from Chicago?

	<u>Paul Said</u>	<u>Baruch Said</u>
a. Melbourne	<u>.45</u>	<u>.8</u>
b. Mexico City	<u>0</u>	<u>0</u>
c. Capetown	<u>.2</u>	<u>.05</u>
d. Singapore	<u>.35</u>	<u>.15</u>
	1.00	1.00

Paul felt certain the answer is not Mexico City. Capetown seemed unlikely, Melbourne and Singapore seemed like the best bets, with Melbourne getting a slight edge. Baruch also eliminated Mexico City and thought Melbourne was the most likely to be correct. He differed from Paul in that he felt more sure about Melbourne being right.

Again, try to be well calibrated. All the alternatives to which you assign 0 should be wrong answers, and all your 1.0's should be right answers. Twenty percent of your .2's should be right, and 80% wrong . . . and so on.

#### Task 4 Sample Items

- Mammoths died out about
  - 10,000 years ago
  - 75,000 years ago
  - 5,000,000 years ago
  - 150,000 years ago
- About how tall (at the shoulder) is an adult male Afghan hound?
  - 27 inches
  - 20 inches
  - 17 inches
  - 30 inches

#### Task 4 Answer Sheet [answers to the two sample items are shown]

State the probability that each of the four alternatives is the correct answer. Make sure that the four probabilities you give sum

to 1.

1. a   /    
 b   O    
 c   O    
 d   O  

2. a   /    
 b   O    
 c   O    
 d   O  

3. a         
 b         
 c         
 d       

### Task 5

#### Task 5 Instructions

The Instructions for Task 5 were identical to those for Task 3.

#### Task 5 Sample Items

The items for Task 5 were similar to those for Task 3, although no items were repeated.

#### Task 5 Answer Sheet

The answer sheet for Task 5 was identical to that for Task 3.

### Task 6

#### Task 6 Instructions

This task is somewhat different from the others. Instead of responding with a probability, you're going to respond with a guess (actually, 5 guesses) at the answer.

Each item asks about a quantity, a number. For example, suppose I show you a bottle full of beans, and ask how many beans are in this bottle? You won't know exactly how many, but you can make a guess.

Now, here's where probabilities come in. We want you to give an estimate of how many beans, such that the probability is .50 that the true number of beans is above your guess, and .50 that the true number of beans is below your guess. We'll call this first estimate the "50th percentile."

Next, state a number such that the probability that the true number is smaller than your estimate is just .25, while the probability that the true number is larger than your estimate is .75. This second response you make is called the "25th percentile."

Next, state a number such that the probability that the true number is smaller than your estimate is .75, while the probability that it's larger is .25. This is the "75th percentile."

Next, the 1st percentile: This is a low number, such that there's just a 1% chance the true number is lower than the number you state.

Last, the 99th percentile: This is a high number, such that there's a 99% chance the true number is lower than the number you state.

The answer sheet for this uncertain quantities task will provide spaces for your answers in ascending order:

1st percentile \_\_\_\_\_  
 25th percentile \_\_\_\_\_  
 50th percentile \_\_\_\_\_  
 75th percentile \_\_\_\_\_  
 99th percentile \_\_\_\_\_

We don't really care in what order you fill in the five percentiles. But, of course, the 1st percentile should be the smallest number, with each answer larger than the one before, and the 99th percentile should be the largest number.

The responses you give need not be evenly spaced or symmetric. For example, the following is a perfectly acceptable set of answers to the number of beans:

1st percentile     200  
 25th percentile   400  
 50th percentile   500  
 75th percentile   1,000  
 99th percentile   1,500

Notice that there's a 300 bean difference between the 1st and 50th percentiles, and a 1,000 bean difference between the 50th and 99th percentiles. That's okay.

The answerer, Sarah, is almost (98%) certain the number of beans lies between 200 and 1,500, and she'd be willing to bet even money that number is over 500. Her subjective odds are 3 to 1 that the number is over 400 (75% chance versus 25% chance), and 3 : 1 that it is under 1,000.

An explanation of how the concept of calibration applies to this task may help you perform the task. For every item, we know the true answer. We will make a tally for each item in one of six categories, depending on the value of the true answer compared with the values you assigned to the five percentiles. The categories are:

1. Lower than the 1st percentile
2. Between the 1st and 25th percentiles
3. Between the 25th and 50th percentiles
4. Between the 50th and 75th percentiles
5. Between the 75th and 99th percentiles
6. Higher than the 99th percentile

So if we were tallying Sarah's responses, if the number of beans is really 157, we'd put a tally in category #1. If it's really 327, we'd put a tally in category #2, and so on. Each item gets just one tally.

In this task calibration means: over many such items, exactly 1% of the tallies should fall in the first category (i.e., below the 1st percentile), exactly 24% in the second category (between the 1st and 25th percentiles), and so forth:

Perfect Calibration:

<u>Category</u>	<u>% of Tallies</u>
1	1
2	24
3	25
4	25
5	24
6	1

These percentages follow directly from the definitions of the percentiles. The 99th percentile is a number such that there's a .99 probability that the true answer falls below the number you state. That means, in the long run, just 1% of the true answers will fall in category #6.

This whole task is another way of expressing your own subjective feelings of uncertainty. If you are very uncertain about the true answer, spread your estimates over a wide range. If you are able to narrow the answer down quite precisely, your estimates should be close together.

For each item, after you have written your answers, review them with the following bets in mind:

First, if you had a \$1 bet on the true answer being either above or below the 50th percentile, would you care which side of the percentile paid off for you? If you have a preference, it means your 50th percentile is wrong. Adjust it until you are indifferent between betting on the interval above the 50th percentile and betting on the interval below the 50th percentile.

Second, the 25th, 50th, and 75th percentiles should split the whole range into four equally likely segments. Would you rather bet on one of the segments rather than any of the others? If so, something's wrong. Adjust your answers until you are indifferent to which one of the following four bets you might play:

1. Win \$1 if true answer is below the 25th percentile, otherwise win nothing.
2. Win \$1 if true answer is between the 25th and 50th percentiles, otherwise win nothing.
3. Win \$1 if true answer is between the 50th and 75th percentiles, otherwise win nothing.
4. Win \$1 if true answer is above the 75th percentile, otherwise win nothing.

Third, check your 1st percentile and 99th percentile answers. Try to think about being well calibrated on these: 1% of the true answers should fall below your lowest number, and 1% above your highest number.

We can't tell you exactly how to pick your estimates--that all depends on how much you know about each uncertain quantity. But we can warn you about two (contradictory!) pitfalls. On the one hand, please don't give us ludicrously high or low estimates. For example, if we ask you how many Polaroid cameras were sold last year in the United States, please don't give as your 1st percentile answer "6"--you know that's a silly answer. And don't give as your 99th percentile answer "2 billion"--that's almost ten cameras for every man, woman and child in the United States.

On the other hand, try to avoid being too sure of your knowledge, thus making your estimates too close to each other. Remember, there's supposed to be only a 2% chance that the true answer will fall outside the range you give us--one percent below the lowest, and one percent above the highest.

One more warning: Some of the questions are questions about percents. Don't confuse your assessments, which will all be percents, with the percentiles. For example, what is the percent of U.S. citizens who are Roman Catholics? For an item like this, you may be giving a low percent as an answer to a high percentile. Try this one below:

1st percentile \_\_\_\_\_  
 25th percentile \_\_\_\_\_  
 50th percentile \_\_\_\_\_  
 75th percentile \_\_\_\_\_  
 99th percentile \_\_\_\_\_

We apologize for that confusing word, "percentile."

Okay, now try another example. How old is the man in the attached picture?

1st percentile \_\_\_\_\_  
 25th percentile \_\_\_\_\_  
 50th percentile \_\_\_\_\_  
 75th percentile \_\_\_\_\_  
 99th percentile \_\_\_\_\_

Since this task takes longer per item than the others, we are giving you only 77 items. Take your time on each item. Your goal is to be well calibrated.

As usual, feel free to ask any questions you wish.

Task 6 Sample Items [with answers]

1. In what year did the Peoples' Republic of China join the UN? [1971]
2. How many cubic inches are there in a liquid quart? [57.75 cu. in.]



Task 6 Answer Sheet

1. 1st percentile \_\_\_\_\_  
 25th percentile \_\_\_\_\_  
 50th percentile \_\_\_\_\_  
 75th percentile \_\_\_\_\_  
 99th percentile \_\_\_\_\_

2. 1st percentile \_\_\_\_\_  
 25th percentile \_\_\_\_\_  
 50th percentile \_\_\_\_\_  
 75th percentile \_\_\_\_\_  
 99th percentile \_\_\_\_\_

## Training Sessions

Instructions

No written instructions were provided for the computerized training sessions. Each participant was shown how to use the computer: Type the letter A or the letter B, followed by either a decimal point (period) and a number, or the number 1 without a decimal point. Finish by typing "Return." Participants were shown how to correct an error, either before or after typing "Return."

Sample Items

The items appeared in the following format (underlined letters and numbers indicate the participant's responses):

1. WHICH EVENT HAPPENED FIRST?  
 A. STALIN'S FIRST FIVE-YEAR PLAN  
 B. WILL ROGERS DIES

[correct answer: A]

?

B.6

2. THE POTSDAM CONFERENCE WAS HELD AT THE END OF  
 A. WORLD WAR II  
 B. WORLD WAR I

[correct answer: A]

?

A.55

3. WHICH IS THE CORRECT SPELLING?  
 A. BOOMERANG  
 B. BOOMARANG

[correct answer: A]

?

A1

Feedback

No written instructions were provided for the discussions between experimenters and participants that occurred after every training session.

### Post-Test Instructions

Now that you have finished the training portion of the experiment, we're going to ask you to do six post-tests, the same tasks you did as pre-tests.

Since most subjects found the Uncertain Quantities task (the one where you gave the .01, .25, .50, .75 and .99 percentiles) most noxious, we are going to have you do it first. That way, when you finish it, you'll know you're over the worst.

Please RE-READ the instructions for each task.

Your goal in this post-testing section is to apply whatever you learned in the training to all these other tasks on which you've gotten no feedback. TRY TO BE PERFECTLY CALIBRATED on each task. Don't concentrate on remembering what you answered to these items before. Instead, try to approach these items anew, to show us that you have learned to be well calibrated.