DOCUMENT RESUME

ED 238 337 HE 016 865

AUTHOR Zechmeister, Eugene B.; And Others
TITLE Training College Students to Assess Accurately What They Know and Don't Know.
INSTITUTION Loyola Univ., Chicago, Ill.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [83]
GRANT NIE-G-81-0093
NOTE 74p.
AVAILABLE FROM Psychology Department, Attn: Zechmeister, Loyola University of Chicago, 6525 N. Sheridan Road, Chicago, IL 60626.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Behavioral Science Research; *Cognitive Measurement; *College Students; *Confidence Testing; *Decision Making; High Achievement; Higher Education; Knowledge Level; Low Achievement; Pretests Posttests; *Response Style (Tests)

ABSTRACT

A study was conducted to improve confidence judgment (CJ) accuracy of college students through training in discriminating known from unknown information. Both low- and high-achieving college students were given CJ tasks, consisting of general information questions, before and after a brief training session. In addition, as part of the initial CJ test, half the students in each achievement group were asked to provide reasons why they selected a particular answer. Training included personal feedback about each student's performance on the CJ pretest and discussions and written exercises directed toward teaching students to weigh carefully the evidence for why a particular answer was correct. Findings include the following: low achievers were more overconfident than were high achievers; the requirement to provide reasons for why an answer was correct reduced overconfidence for low, but not for high, achievers; and training led to significant improvement in CJ performance, although the effect was greater for low than for high achievers. It appears that high achievers were more likely to engage spontaneously in those cognitive activities that are important to making appropriate judgments about what is known. (Author/SW)

Training College Students to Assess Accurately What They Know and Don't Know

Eugene B. Zechmeister, Kathleen M. Rusch
and Kathryn A. Markell
Loyola University of Chicago

Running head: Judging What is Known

This research was sponsored by Grant NIE-G-81-0093 from the National Institute of Education. We would also like to acknowledge support given to K. M. Rusch by the Graduate School of Loyola University.

We wish to thank Michael Losoff and Maria Robles for their help in the data collection phase of this experiment, and Kathleen O'Keane and Elizabeth Smith for their aid in data analysis. Richard Bowen of Loyola University kindly allowed us to make several appeals for volunteer subjects in his introductory psychology class and also adjusted his examination procedure so that we could collect confidence judgment data. Fianlly, this project benefited significantly from suggestions given by Stanley E. Nyberg and John J. Shaughnessy, with whom this research was discussed on numerous occasions.

Requests for reprints should be sent to Eugene B. Zechmeister, Psychology Department, Loyola University of Chicago, 6525 N. Sheridan Road, Chicago, Illinois 60626.

Abstract

Both low and high achieving college students were asked to give confidence judgments for answers to general information questions before and after a brief training session aimed at improving confidence judgment (CJ) accuracy. In addition, as part of the initial CJ test half the students in each achievement group were asked to provide reasons why they selected a particular answer. Training included personal feedback as to each student's performance on the CJ pretest and discussions and written exercises directed toward teaching students to weigh carefully the evidence for why a particular answer was correct. The post-test was given approximately 2 weeks following training. Major findings were that: (a) low achievers were more overconfident than high achievers; (b) the requirement to provide reasons for why an answer was correct reduced overconfidence for low, but not high, achievers; and (c) training led to significant improvement in CJ performance, although the effect of training was greater for low than high achievers. It appears that high achievers are more likely than low achievers to engage spontaneously in those cognitive activities that are important in making appropriate judgments about what is known.

Training College Students to Assess Accurately What They Know and Don't Know

An important characteristic of the human learner is the ability to discriminate between known and unknown information. In fact, efficient learning and remembering would seem to depend on it. This metacognitive skill is the basis for decisions regarding the progress of learning (e.g., Bisanz, Vesonder & Voss, 1978), the current state of knowledge about an event (King, Zechmeister, & Shaughnessy, 1980), the likelihood of later retention of presently unrecallable facts (Hart, 1965; 1967), and the correctness of answers retrieved from long-term memory (Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein, Fischhoff, & Phillips, 1977).

Although most adult learners generally can distinguish what they know and what they do not know, such an ability may not be well developed in younger learners or in learners who are experiencing learning difficulties (Bransford, Stein, Vye, Franks, Auble, Mezynski, & Perfetto, 1982; Flavell, 1979). Furthermore, both anecdotal and experimental evidence suggest that discrimination between known and unknown information is far from perfect in many individuals. Few instructors, for instance, have not been confronted by a student who, having done poorly on an examination, laments that "I thought I really knew it!" Indeed, many of us have had the experience of claiming absolute

certainty in our knowledge of a particular fact, only to find out later that this degree of confidence was unwarranted.

Perhaps the most frequently cited studies of a memory monitoring ability are those carried out by Hart (1965; 1967). He showed that people can predict reliably later recognition of presently unrecallable facts. Using a recall-judgment-recognition task, college students were first asked to recall information from long-term memory. When recall was not successful students were asked to judge whether they would recognize the item when presented among several alternatives. Items rated as low on a "feeling of knowing" scale were less likely to be recognized than items rated high on this scale. Therefore, Hart's experiments provide evidence for an ability to discriminate what is known (but temporarily inaccessible) and what is unknown. However, his results also show considerable slippage in this ability. For instance, items for which students indicated that they had a very strong feeling of knowing, and, apparently were sure they would recognize, actually were recognized only 75 percent of the time (Hart, 1965, Exp. 2). People in Hart's study were, in other words, seriously overconfident in their prediction of later recognition of unrecallable facts.

A failure to accurately assess what is known or unknown is also revealed when attempts are made to "calibrate" subjects' confidence in the correctness of information retrieved from long-

term memory (Koriat, et al., 1980; Lichtenstein & Fischhoff, 1977). A frequent procedure is to present subjects with general information questions and two alternatives as possible answers. Subjects select one alternative and then assign a probability (.50 through 1.00) that the answer is correct. A large number of questions are included and on the basis of subjects' responses a calibration curve is constructed showing the relationship between rated probability and actual probability correct. That is, the calibration curve shows proportion correct for each of the assigned probability levels. The results of many studies, using a wide variety of materials, reveal a "typical" calibration curve, namely, one showing marked overconfidence in the correctness of answers. People tend to overestimate the probability that their answers are correct. It is not unusual, for instance, for subjects to be correct only 80-85 percent of the time when they have indicated an absolute certainty that an answer is right (Fischhoff, Slovic, & Lichtenstein, 1977).

When the goal is to acquire new information, overconfidence is likely to be a source of learning problems. A learner, for instance, who places undue confidence in the correctness of a wrong answer will terminate inappropriately further retrieval efforts. A learner who judges presently studied information to be known, when it is not, would appear to be inviting academic disaster. A learner who mistakenly predicts that recognition will succeed, when recall does not, has failed both types of

retention tests. There is, in fact, a growing literature to suggest that less successful learners are deficient in those metacognitive skills necessary to evaluate the state of their knowledge (e.g., Bransford, et al., 1982; Owings, Petersen, Bransford, Morris, & Stein, 1980). Shaughnessy (1979), for instance, examined confidence judgments which students gave to answers on a series of classroom multiple-choice tests. Although students displayed an ability overall to discriminate known from unknown items, discrimination performance was correlated with test performance. Students who scored high on the classroom tests apparently were better able to judge what they knew than students who scored low.

In the present study an attempt was made to improve confidence judgment (CJ) performance of college students through training in discriminating known from unknown information. The effects of this training procedure were evaluated by looking at differences between pre- and post-training performance on laboratory assessment tasks and by evaluating the appropriateness of confidence judgments of trained and untrained students in the context of a regularly scheduled classroom exam. Groups of high and low achievers, as defined by performance on two introductory psychology exams, served as subjects. This allowed a comparison of CJ performance between these two groups of students prior to training, and provided an opportunity to examine whether the training procedure, to the extent that it worked, was more

effective for one group of students than another.

Previous studies aimed at improving CJ accuracy have shown mixed results (see Lichtenstein, Fischhoff, & Phillips, 1982). In general, these efforts have concentrated on evaluating the effect of one or two factors on CJ performance and mainly have focused on potential changes in CJ accuracy when they are assessed immediately following a training intervention. For example, Koriat, et al. (1980), using a two-alternative, forced-choice procedure, required subjects to list reasons for and against each of the alternatives prior to selecting an answer and rating the probability that it was correct (Exp. 1). This procedure significantly reduced overconfidence. In a second experiment, before rating confidence subjects were asked to list a reason supporting, to list a reason contradicting, or, to list both a supporting and contradicting reason for a chosen alternative. Only the listing of contradictory reasons improved CJ accuracy. These researchers also showed that subjects generally were biased not to consider contradictory evidence for their answers, a fact that was seen as a possible source of overconfidence. Koriat, et al. concluded: "While further research is clearly needed, we can derive some practical advice from the present results. People who are interested in properly assessing how much they know should work harder in recruiting and weighing evidence. However, that extra effort is likely to be of little avail unless it is directed toward recruiting

contradictory reasons" (p. 117).

Overconfidence is not significantly affected when students are given lengthy instructions so that they cannot possibly misunderstand the nature of the task, or when motivation is raised by giving students the opportunity to bet against the experimenter for real money (Fischhoff, et al., 1977). However, Lichtenstein and Fischhoff (1980) found that CJ accuracy is improved if comprehensive feedback is given subjects regarding their performance. In one experiment, subjects were given intensive training consisting of 11 sessions, each involving 200 two-alternative general knowledge items. Subjects indicated which answer they thought was correct and assigned a probability from .50 to 1.00 that they had chosen correctly. Following each session performance was summarized and the results discussed with the experimenter. An effect of training was observed, but all measurable improvement occurred following the first training session. Generalization tests on related probability assessment tasks were given and only modest effects were obtained. Why training worked and why there was not more generalization is not altogether clear. However, the researchers suggested that one critical factor in training appears to be personal feedback, "whose relevance can not be rejected with the claim 'I'm not like that'--as might confront a report that 'most people are overconfident'" (p. 170).

The present experiment differed from previous studies

investigating CJ performance in several important ways. For example, half the subjects were asked during the CJ task to provide the experimenter reasons for their choices. Arkes, Lai and Hackett (Note 1) have reported that informing students that they will have to explain to a group of their peers why they chose a particular answer serves to reduce overconfidence. In addition, training incorporated several factors that appear to be important in improving the accuracy of probability assessments. For instance, both personal feedback regarding CJ performance, and discussion and written exercises related to selecting an appropriate confidence level given the evidence which is available, were included as part of the training session. Also, long-term rather than immediate consequences of training were of interest. Laboratory post-tests occurred approximately two weeks following training and CJ performance of trained subjects was assessed in a classroom exam given about one month after training. Finally, as has been mentioned, effects of training (as well as the requirement to give reasons) were contrasted for high- and low-achieving students.

Method

Design

The design was a pretest-post-test design with a control group. All students were given CJ and feeling-of-knowing (FK) tasks at each of two experimental sessions held approximately one month apart. Between experimental sessions one half the students

were given training aimed at improving CJ performance and half were not (Control). There were two other variables of interest which were combined factorially with the two levels of training and no training. One variable was whether or not students were required to state, for a subset of items on the first CJ task, why they chose a particular answer. The other major variable of interest was achievement level, high or low, of the students. Students were selected to participate on the basis of their performance in an introductory psychology class. One half the students in each of the four groups formed by the combination of training-no training and reasons-no reasons factors were high achievers, and half were low achievers. Finally, following the post-test all students were required to provide confidence judgments when choosing answers as part of a regularly scheduled classroom examination. This exam followed the training session by about one month.

Subject Selection

Early in the semester all students in a large introductory psychology class were asked whether they would give permission to the principal investigator (PI) to use their classroom test scores in association with a research project for which they might be selected. Participation in psychological research was one way for students to earn course credit. Ninety-eight percent gave written permission to use their test scores for this purpose. Subject selection was made immediately after the second

of four regularly scheduled exams (approximately the 7th week of the semester). Students were identified whose z scores were $\geq +.5$ or $\leq -.5$ (approximately the 70th and 30th percentiles) on each of the two classroom examinations. Names of these students were posted with the request that they see the PI in order to make appointments for serving in the research project. (It was not indicated why particular students were selected, and, therefore, it is unlikely that any students were aware that they had been selected because they had either scored high or low on the classroom tests.) A total of 87 students met the criteria for selection, 47 high and 40 low achievers. Seven high achievers were randomly dropped to equalize the groups. Students were asked to make two hourly appointments approximately one month apart. In addition, all students were advised that they might be contacted and asked to come for one more session between the two sessions for which they had made appointments. A total of 33 high achievers and 39 low achievers made appointments. Two students declined to participate because they already had sufficient research credits; others did not participate because the available research times did not meet their schedule. Finally, six students never were heard from despite several pleas made in the classroom for students on the list to see the PI.

The goal was to test 32 students at each achievement level or a total of 64 students. Several additional students, however, were included in the experiment so as to replace any students who

might eventually withdraw from the course. Interestingly, only one subject was lost for this reason, although data from a few students had to be replaced (as noted later) because they produced extreme scores on one or more of the laboratory tests. Only after all students had made appointments were half the high and low achievers randomly sampled for participation in the training session. These students were contacted by mail and asked to sign up for a brief additional session. The training sessions were conducted after all students had been pretested but prior to the second or post-test session. Students were tested individually during pre- and post-training sessions; whereas, they attended the training sessions in small groups.

CJ and FK Tasks

Both the CJ and FK tasks consisted of 100 general information questions. Within each test type, the format of these tasks was the same at pre- and post-test sessions. Therefore, each student attempted to answer a total of 400 general information questions in the context of either CJ or FK tasks. The CJ task was administered using a microcomputer and television monitor. The computer was programmed to present questions one at a time on the monitor screen. Below each question appeared two possible alternatives as answers, and below these appeared a question mark. Students were instructed that when the question mark appeared that they were to enter a 1 or a 2 on the computer keyboard depending on whether they wished to

select the first or second alternative. When an answer had been typed, the question "How confident are you?" appeared on the screen. Students entered one of six possible numbers, 5-10, to identify how confident they were that they had selected the right answer. (The zero key on the keyboard was used to indicate a confidence of 10.) Instructions given prior to the CJ task explained that the numbers corresponded to probability judgments varing from .50 to 1.00. A probability scale showing these proportions was placed above the computer keyboard.

The FK task was given to students immediately following completion of the CJ task; however, students were given the option of completing this task in the laboratory or taking it away with them to be completed at their convenience within the next 48 hrs. (or, if a weekend followed the day after a session, in 72 hrs.). The instructions and materials were placed in several envelopes that were to be opened in a sequence carefully outlined by instructions appearing on the outer envelope. Although this procedure is less than ideal, for example, there is no way to check that students followed instructions exactly, the CJ task required nearly 45 min. to complete and it was not practically possible to schedule the more than 64 students for two 2-hr. appointments. A drop-off location in the Psychology Department was clearly identified in the instructions and a record was kept of the students' delivery of the test materials to this location. Surprisingly, only one student had to be

reminded by a telephone call to bring in the materials. Several students called the experimenter asking for an extra day because they had forgotten to bring the package to school or had not yet gotten to it. Every indication was that nearly all students conscientiously followed the instructions that were given with the materials.

The FK task had three parts. The first envelope contained 100 general information questions. Students were instructed to answer all the questions using an answer sheet provided. Instructions specifically urged students to write down an answer for every question even if it was a guess. Second, after attempting to answer each question students were asked to judge the likelihood that they would be able to recognize the answer if it appeared among several alternatives. This prediction of later recognition was to be made for each question on the test. It was emphasized in the instructions that if recall was judged to be successful then a high likelihood of recognition would be predicted. However, students were instructed that there undoubtedly would be questions for which they had not been able to recall the answers, but, later, would be able to recognize the answers. A 6-point scale, 1-6, was used to indicate likelihood of recognition. The scale points were labeled (1) "Would be guessing" and (6) "Definitely will recognize it." The numbers 1-6 appeared next to each answer space. Instructions also asked students to record the time when they started this task.

Following the recall and rating parts of the recall task they were instructed to record the time and to open the next envelope. Although not specifically forewarned by the previous instructions, students now found a 100-item multiple-choice test using the previously attempted questions. Four alternatives were provided for each question and students circled one of four numbers on an accompanying answer sheet to indicate their choices. Instructions emphasized that an answer was to be circled for every question even if it had been recalled previously. When finished, students were asked to record their time and return the materials to the drop-off location.

Asking Reasons

During the first CJ test, but not the second, half the students were asked, for a subset of 20 items, to give reasons why they had selected a particular answer and why they had given it the confidence that they did. Two "reasons items" appeared nonsystematically in each block of 10 items. Students were carefully instructed prior to the CJ task in this aspect of the experiment. They were informed that reasons would be requested by the experimenter for those items that were followed by a tone. The experimenter prompted the students after the tone was presented and responses of the students were tape recorded. It should be emphasized that students did not know for which items reasons would be requested until they had completed answering the question and had given a confidence judgment.

Test Lists

Items used in the CJ and FK tasks came from several different sources.[1] Prior to the present experiment more than 400 general information questions were presented to introductory psychology students who attempted recall and rated likelihood of later recognizing the answers. Specifically, after writing an answer, students indicated whether they believed the answer was right or wrong. If the answer was judged to be wrong then a prediction was made of later likelihood of recognition. Questions were tape recorded and were presented in sets of 100 to small groups of students until at least 15 students had attempted to recall each of the items. This procedure permitted items to be rank ordered in terms of recall difficulty and also generated alternatives for the CJ and FK tasks. Finally, the pilot testing revealed items which a substantial number of students judged to be correct with a high degree of confidence, but were not. These items were identified as "deceptive" and were systematically included on the test lists. None of the students tested in this pilot task served in the actual experiment.

From the available item pool four sets of 100 questions were prepared under the following restrictions. Items were grouped according to whether they were very hard, moderately hard, moderately easy, very easy, or deceptive, as determined by recall probabilities and subjects' judgments as to whether an answer was correct. Four different 60-item "core" lists were constructed by

randomly assigning items from the four difficulty levels and deceptive category so that there were 10 easy, 20 very hard, 10 moderately hard, 10 moderately easy and 10 deceptive items in each set.

Because overall proportion correct in a CJ task is known to be systematically related to overconfidence (for example, Lichtenstein & Fischhoff, 1977; Nyberg, Engelbrekt, Zechmeister, & Ruble, Note 2), and because high and low achievers normally would be expected to differ in terms of the number of general information questions which they could answer correctly, sets of lists were constructed that would yield approximately the same proportion correct for each achievement group. Specifically, to the core lists described above, either 6 easy and 34 moderately easy items were added, or 14 very hard and 26 moderately hard items were added. Varying these additional 40 items was intended to make the lists approximately equal in difficulty for the low and high achievers tested in this experiment. The four core lists with the additional "easy" items (for low achievers) or "hard" items (for high achievers) were assigned randomly to be used in either the CJ or FK tasks.

In constructing the 100-item lists the items were assigned to positions so that items of various difficulty and type were systematically varied across the list. For example, in each block of 10 items there was one deceptive item and one very easy item. Also, for the CJ lists the 20 items for which half the

subjects were asked to give reasons for their answers were always the same 20 items across subjects. These items, also, were systematically selected. For example, the 10 deceptive items in the list used for the CJ pretest were 10 of the 20 questions for which reasons were asked. Other reasons items included instances from all the difficulty levels.

Finally, the order of the lists was counterbalanced so that various forms of the list were used equally often in experimental conditions and in both pre- and post-training tasks. For example, two forms of the FK test were used equally often (for both high and low achievers) for the first and second sessions. For the CJ test the same procedure was followed except that the 20 items in the reasons subset were held constant for the first session (and 20 similar items were constant on the post-test). The two forms of the CJ lists were counterbalanced with the exception of these items. Therefore, for the CJ test all items except 20 were used equally often at each session and all students in the reasons condition were asked to provide reasons for the same 20 items. Therefore, each subject experienced a different set of general information questions at each stage of the experiment, and, although the lists used for the high and low achievers were not identical, a majority of the items (60 percent) were the same.

Training Sessions

Each training session was 30-35 min. in length with the

general procedure as follows.[2] First, the nature of the research project was explained with particular emphasis on the fact that CJ accuracy, rather than number correct, was of major interest. Then, each student was presented a graph showing two calibration curves, one summarizing performance of a group of 32 students (16 high and 16 low achievers) on the CJ pretest, and the other showing the student's own performance on the CJ task. The manner in which a calibration curve is constructed was discussed, and, in agreement with the group curve, it was pointed out that most people are overconfident in this type of task. Students were asked to examine their own calibration curve and to compare it both to the group curve and to a diagonal line representing "perfect" calibration.

After the calibration curves were examined and any questions answered the question was raised as to why students (and people in general) are not very good predictors of what they know. The answer given by the experimenter was that most people do not adequately mentally cross examine the evidence for and against a particular answer. It was pointed out that people are biased toward considering evidence why an answer might be right rather than considering why it might be wrong. Further, students were told that only by weighing evidence carefully can we predict accurately what we know and do not know. The analogy of a prosecuting attorney cross examining a witness was used to illustrate how we must probe for reasons why our answer is or is

not correct.

At this point in the session students were given a "reasons" test. It consisted of 10 general information questions with two alternatives as possible answers. Unlike the questions presented on the CJ test, however, one of the answers (the correct one) was identified as correct. Next to each question and correctly circled alternative was one of 6 confidence ratings, ranging from 5-10. Finally, beneath this there appeared three possible reasons for selecting the correct answer. Students were asked to assume that someone had answered the question correctly and had given the confidence rating that appeared next to the answer. Their task was to select for each of the 10 items the reason which was most appropriate given the confidence level indicated. It was emphasized that they were not to select the reason which best justified the answer, but, rather, to select the reason most appropriately associated with the confidence assigned to the answer. A sample item from this reasons test is the following:

2. What is the capital of New York?

1) New York *2) Albany

Confidence level: 7

Reasons:

a) "I used to live in New York when I was younger and I know the capital is Albany."

b) "I really have no idea but I picked Albany because it sounds like it would be the capital of a state."

c) "I've never been to Albany, but I've been to New York City and I don't remember seeing a capitol building or state legislature or anything like that. So it's probably not New York City."

Reasons provided as alternatives were similar to those reasons given by subjects who had been asked reasons as part of the first CJ test. Each of the six possible confidence levels (5-10) was used at least once with the items on the reasons test. For each question one reason was judged *a priori* by the investigators to be most appropriate given the confidence level indicated. For example, for the above items, reason (c) was considered the "correct" response. After students completed this test the correct answers were reviewed and the test booklets collected.

Following the reasons test all students were given another short 10-item test. The questions were identical to ones that the students had seen on the prior CJ task. However, 5 of the 10 questions were ones that were known to be particularly "deceptive," in that these questions were often marked wrong but with a high degree of confidence that they were right. For example, one question was:

In what country is the highest waterfall in the world?

1) Venezuela 2) Canada

Students were asked to answer each of the questions and to indicate their confidence in the right answer as they had on the

CJ test given previously. When all students had done this the deceptive items were identified and a poll taken of those present to find out who had answered incorrectly. Not everyone, of course, was deceived by these questions but invariably someone answered one of the five deceptive questions inappropriately. These particular items were discussed and it was apparent that students recognized why these questions might be deceptive even if they had not themselves been led to answer incorrectly.

Finally, the major points of the training session were summarized, and, once again, the image of a tough prosecuting attorney cross examining a witness was used to remind students that evidence for the validity of an answer must be carefully evaluated. A personal anecdote was also mentioned showing how the PI had placed inappropriate confidence in one answer while constructing the test items. It should be noted that the FK test was not mentioned during the training session. Students were told that in the next laboratory session they would have the opportunity to improve their CJ performance. When questioned after the training session students indicated that they had understood the goals of this session. Both high and low achievers were likely to be present at a training session.

General procedure

All students were telephoned and reminded of their appointments the day before they were scheduled. When students arrived at the laboratory they were instructed in the use of the

microcomputer and were presented task instructions using the television monitor. The experimenter then summarized the instructions and presented five practice items. If reasons were to be requested subjects also were given a practice item for which reasons were asked and were informed that their responses would be recorded.[3] The experimenter also recorded the time that subjects took to finish the computer task. Following completion of the CJ task students were given the package of materials containing the FK task and asked whether they would like to take it with them or to complete it now. Most students opted to take the package with them and return it later.

When students not in the training group returned for the second experimental session they were given instructions indicating that the experimenter was interested in whether they could improve their CJ accuracy ("improve your ability to discriminate right and wrong answers") from that of the first session. Students were told that their performance would be compared to others taking this task. Students who had taken part in the training session received similar instructions except that they were asked to apply what they had learned during training. The major points of the training session were reviewed and students were reminded that to be an accurate predictor of what is known that they must challenge their answers just as a prosecuting attorney must challenge the responses of a witness. Average number of days between training and post-tests was 14.47

for high achievers and 13.12 days for low achievers ($p>.05$).

Class Examination

As part of the final examination in the introductory psychology class all students in the class were asked to provide confidence ratings for their answers. The exam covered the material since the previous or third exam and contained 50 four-alternative, multiple-choice items. The request to give confidence ratings was made by the instructor of the course and no deliberate association was made between the request and participation in the present project. It was suggested that providing confidence judgments would help them to think about the right answer to each question as well as provide information that might be used as part of an item analysis of the test. Each student was offered one extra credit point for participation. A 6-point confidence scale was used with the endpoints labeled (1)"Guessing" and (6)"Absolutely Sure." Number of days between the training session and classroom exam was either 33 or 34 for all subjects.

Results

Although various measures have been used to evaluate CJ performance (see Lichtenstein & Fischhoff, 1977), including, most recently, measures derived from signal detection theory (see Ferrell & McGoey, 1980), over-confidence is most often determined by summarizing performance in the form of a calibration curve. Subjects are said to be well calibrated (and neither over- nor

underconfident) when, over many judgments, for all proportions assigned a given probability, the proportion that are true matches the assigned probability (Lichtenstein, et al., 1982). A calibration curve describes, in other words, over-confidence at each level of reported confidence. With general knowledge questions of moderate or extreme difficulty, overconfidence is the most pervasive finding in recent research in this area (Lichtenstein, et al., 1982); although, it is usually greater for high than low levels of reported confidence, and is often most severe for the middle range of the confidence scale. Individual measures of over-confidence can be obtained by calculating for each subject the difference between overall mean confidence and proportion correct (O-U = Mean Confidence - Proportion Correct). A positive O-U score reveals overconfidence and a negative score shows underconfidence. When task experiences and overall proportion correct are similar, differences in O-U only can be obtained when subjects use the confidence scale differently, for example, by using lower values of the confidence scale more frequently, thus, lowering overall mean confidence. The results of the present experiment are, therefore, described in terms of calibration curves summaring CJ performance, as well as in terms of mean O-U and frequency of use of levels of the confidence scale. In addition, CJ performance for certain item types, for example, those considered deceptive, is examined.[4]

In presenting the results, performance on the CJ task will

be described first, followed by a description of performance on the classroom exam, and, finally, data obtained using the FK task will be reported.

Pretraining CJ Performance

Students answering correctly greater than .90 or less than .60 items on the CJ task were not included in the analyses. Three low achievers, one scoring very high, and two with very low scores, were excluded. Several additional subjects were randomly dropped in order to provide equal numbers of high and low achievers (n = 32) and equal numbers of subjects in the experimental groups created by the combination of achievement, reasons, and training factors (n = 8).

There was no significant difference between high and low achievers in the amount of time taken to complete the CJ test. Overall, students not asked to provide reasons spent an average of 17.27 min. on the task; mean time of students asked to give reasons for their answers was 31.94 min.

Proportion correct was .72 for high and .70 for low achievers (n = 32), and did not differ significantly, $t(62)$ = 1.68, $p > .05$. Furthermore, item analyses based on CJ pretest performance revealed that item difficulty distributions for the lists used by high and low achievers were highly similar. Results of this analysis are shown in Table 1. For instance, proportion of very easy items (4 or less errors) was .56 overall for lists used by high achievers and .51 for lists used by low

achievers.

Insert Table 1 about here

Proportion of very hard items (11 or more errors) was .06 and .08 for these same groups, respectively. A chi-square test showed no significant difference between the overall item difficulty distributions of these lists, x^2 (5) = 2.60, $p > .05$. (For this analysis, cells identifying number of items with error frequencies greater than 10 were combined in order to avoid extremely small expected cell frequencies.) The test experience, therefore, of high and low achievers was very similar both in terms of number of items answered correctly and in terms of the relative difficulty of items included in the lists presented to these groups of students. Nevertheless, our experience has shown that CJ performance is very sensitive to differences in overall recall, and, therefore, major analyses also were carried out when subjects were matched in terms of number correct on the CJ test.

Calibration curves based on the performance of high and low achievers on the CJ pretest are shown in Figure 1.

Insert Figure 1 about here

Low achievers were relatively more overconfident than high achievers given essentially the same proportion correct. Mean

O-U was .05 for high and .09 for low achievers, $t(62) = 2.36$, $p < .05$. As must be expected, therefore, students in these two achievement groups used the confidence scale differently. Relative to high achievers, low achievers were more likely to use the extremes of the confidence scale (.5 and 1.0), and, consequently less likely to use middle values (.6-.9); although, this difference in frequency of use of the scale levels was greater for high than for low scale values. Average frequency of scale level use for high achievers, for scale values .5 through 1.0 was, 24.28, 11.28, 9.60, 9.56, 9.56, and 35.72. Mean use of these same scale values by low achievers was 27.34, 7.97, 8.13, 7.31, 8.72, and 40.35. The interaction between achievement level and level of confidence scale use was significant, $F(5,310) = 4.63$, $p < .05$.

Differences in CJ performance between high and low achievers, however, were evident only when students were not asked, as part of the CJ task, to say why they chose a particular answer and why a specific level of confidence was used. Data supporting this conclusion were obtained from the following analysis. High and low achievers not asked to give reasons on the pretest were matched on number correct; students from these two achievement groups who did give reasons for their answers were also matched. Twelve matched pairs in each of the reasons and no reasons conditions were produced. Figure 2 shows that the calibration of these matched groups differs only when reasons

were not requested.

Insert Figure 2 about here

Overall mean confidence of high and low achievers, within the reasons condition, and who were matched for proportion correct, was the same, .77; mean confidence of high and low achievers matched for proportion correct but who were not asked to give reasons for their answers was less for high than low achievers, .77 and .80, respectively. Consequently, mean O-U differed for these latter two groups (.07 vs. .10), although not significantly, $t(11) = 1.07$, $p > .05$.

The calibration curves presented in Figure 2 suggest that the effect of asking reasons on CJ performance was different for low and high achievers. This is more clearly seen when, within low and high achievement groups, CJ performance is compared between students asked for reasons and those not asked for reasons. Within each achievement group students who gave reasons and those who did not were matched for number correct on the CJ test. Thirteen matched pairs were obtained for low achievers, and, among high achievers, this procedure yielded 10 matched pairs. Corresponding calibration curves are found in Figure 3.

Insert Figure 3 about here

Low achievers asked to give reasons were less overconfident than those not asked to give reasons. The pattern is not as obvious for high achievers. High achievers asked to give reasons were actually more overconfident through the middle ranges of the scale than were students not asked to give reasons. However, this is based on only 10 pairs of subjects and it seems reasonable to conclude that there was no difference in CJ performance among high achievers as a function of asking reasons. This conclusion is supported by analyses of mean confidence and corresponding O-U scores. Only differences among low achievers even approached statistical significance. Mean O-U of low achievers was .10 for students asked to give reasons and .07 for students not asked to give reasons, $\underline{t}(12) = 1.81$, $\underline{p} < .10$. Among high achievers, overall mean confidence of students asked to give reasons and those not asked to give reasons (matched for proportion correct) was nearly identical, .762 and .765, respectively.

Differences in the frequency of use of the levels of the confidence scale as a function of whether or not reasons were requested support the conclusion that asking for reasons had substantially more impact on the CJ performance of low than high achievers. Within both high and low achievement groups, students asked to provide reasons were, relative to students not asked for reasons, less likely to use a 1.0 confidence level and more likely to use a .5 level of confidence. However, this difference

was more apparent for low than high achievers. Consider, for example, the mean frequency of use of the extremes of the confidence scale by students asked to provide reasons for their answers and those not asked to provide reasons but who were matched within achievement group for proportion correct (Figure 3). Among low achievers not giving reasons mean use of the .5 and 1.0 confidence levels was 25.92 and 42.46, respectively; the mean use of these same scale values by low achievers asked for reasons was 31.38 and 35.00, respectively. Among high achievers the average use by students not asked to give reasons was 24.1 and 34.2 for the .5 and 1.0 extremes, respectively; whereas, high achievers giving reasons used the lowest end of the scale on the average 25.9 times and the 1.0 level of confidence an average of 31.4 times.

Training Tasks

As part of the training session low and high achievers were given both a "reasons" test and a "deceptive items" test. Performance on these tests by experimental students is summarized in Table 2. High achievers were better able than low achievers to "match" an appropriate reason with a particular level of confidence.

Insert Table 2 about here

They also answered more items correctly on the deceptive items

test and were less likely to be deceived than were low achievers. This latter conclusion is based on the finding that low achievers were more likely than high achievers to select a wrong answer and to indicate with a high degree of confidence that it was right (8-, 9- or 10W items in Table 2). However, as was intended by this exercise, both high and low achievers were deceived by certain questions.

Training and CJ Performance

The effect of training on CJ performance was investigated in several ways. First, calibration curves of students in training and no training conditions were examined following performance on the CJ post-test. These curves are shown in Figure 4.

Insert Figure 4 about here

It is apparent from these curves that students who had training aimed at improving CJ performance were less overconfident on the post-test than students who did not have training. Mean O-U was .03 for trained students and .08 for not trained students and differed significantly, $t(62) = 2.98$, $p < .01$. Students in the training condition were less likely to use the high end of the confidence scale than were those students not in the training condition. Mean frequency of use for the six confidence levels (.5-1.0) was, for trained students, 25.0, 12.6, 9.7, 9.2, 9.4, and 34.6. For not-trained students mean use of these same levels

was 22.0, 10.4, 7.7, 8.3, 8.1, and 43.4. The interaction between training-no training conditions and confidence level was significant, $F(5, 310) = 3.12$, $p < .05$. Although students in the training group performed slightly better on the post-test than students not having training (.74 vs. .72), this difference was not statistically reliable, $t = 1.09$, $p > .05$. However, several additional comparisons of post-test CJ performance were made when students were matched in terms of number correct on either the pretest or post-test. All comparisons yielded the same pattern of results as seen in Figure 3. Moreover, mean O-U of trained and not trained students differed significantly in all comparisons.

Differences between pre- and post-test CJ performance were also examined as a function of training. The pre- and post-test calibration curves for high and low achievers in training and no training conditions are found in the two graphs within Figure 5.

Insert Figure 5 about here

Overconfidence of students in the training group (top graph) decreased from pre- to post-tests, whereas students in the control group (bottom graph) actually were more overconfident on the post-test than they were on the pretest. Mean O-U for pre- and post-test performance of trained students was .07 and .03, respectively ($p < .05$); mean O-U of not trained students on pre-

and post-tests was .07 and .08, respectively. Frequency of use of the confidence scale levels also differed between pre- and post-tests as a function of training. For instance, students in the training condition used a 1.0 confidence level less on the post-test than they did on the pre-test (34.16 vs. 38.38), whereas students not in the training condition actually used this extreme confidence level more on the post-test than they had on the pretest (43.45 vs. 37.88).

The effect of training was greater for low than high achievers as revealed in the calibration curves plotted in Figure 6. Mean O-U on the CJ post-test differed only slightly, and not significantly, for high achievers as a function of training.

Insert Figure 6 about here

Mean O-U was .01 for high achievers in the training group and .04 for high achievers in the control group. A large and significant difference in mean O-U for trained and not-trained low achievers was found. Mean O-U was .05 for trained students from this achievement group and .16 for not-trained students from the same group. Because proportion correct was substantially (although not significantly) different for low achievers in training and no training conditions (.74 vs. .70), an effect of training also was examined for this group when students were matched for number correct on pretest performance. Twelve pairs of subjects were

able to be matched with post-test proportion correct .72 for students in the training group and .71 for those not in the training group. With proportion correct closely equated in this way differences were still found between trained and not trained low achievers. Specifically, mean O-U was .06 for trained students and .11 for not trained students, $t(11) = 2.16$, $p < .06$. The general picture, therefore, is of training leading to significant reductions in overconfidence on post-test CJ performance, but this effect being greater for low than high achievers.

An effect of training was examined in yet another way. Twenty items on the post-test were the same for all students. Among these items were 10 items that pilot testing or results from other published studies had suggested were deceptive in that many subjects tended to select wrong answers for these questions with high degrees of confidence that they were right. It can be pointed out that not all the items originally identified were equally deceptive on the CJ post-test. Among the so-called deceptive items were ones for which in pilot testing many people had recalled a wrong answer and had indicated that it was right. When these questions were used in a multiple-choice format, with both the correct and incorrect answer present, the questions were apparently no longer deceptive. Nevertheless, of the 10 original questions each drew at least one wrong response with a confidence of 1.0 that it was right, and one question resulted in 9 wrong

answers with a confidence of 1.0 (based on 32 students in either training or no training conditions). Mean correct for these 10 items was 5.84 for students not trained and 6.28 for trained students. It is difficult to know whether this difference was due to training since overall proportion correct was slightly greater for trained and not trained students at both pre- and post-tests. However, it is possible that one effect of training was to make students more aware of the reasons why they were selecting an answer. An increased awareness of the evidence for an answer may have led to improved performance on certain items, particularly those that might normally be deceptively difficult. For this possibility to be verified, however, it would likely need to be investigated using other measures, for example, data obtained by content analysis of reasons that subjects provide when saying why they answered particular questions.

When confidence judgments were examined for the 10 so-called deceptive items it is apparent that students in the training condition were less likely to choose a wrong answer and assign it a high confidence than were students in the control condition. Number of wrong answers given a confidence of 1.0 was 22 for the 32 trained students and was 42 for the same number of not trained students. Interestingly, there was little difference between high achievers as a function of training. Within this achievement level, there were 8 items drawing wrong answers and highest confidence for students who were not trained and 6 of

these items for trained students. Among low achievers there were 34 wrong answers with 1.0 confidence in the not trained group and 16 of these answers in the trained group. Thus, the differences parallel the effects of training obtained when overall calibration is examined.

Exam Performance

One student, a high achiever in the no training group, failed to give confidence ratings when answering the 50 multiple-choice questions presented on the fourth or final introductory psychology exam. Performance on this test by the remaining experimental students is summarized in Table 3. Among high achievers, mean O-U of trained and not trained students differed significantly, $t(29) = 2.64$, $p < .05$.[5] It can be noted that high

---

Insert Table 3 about here

---

achievers in the training group were relatively overconfident (-.08) on the final exam, and thus, in one sense, can be judged more poorly "calibrated" than students in the no training group who were neither over- nor underconfident (0.0). A difference in CJ performance also was apparent when test scores of high achievers in these two groups were matched ($n = 13$). With exam performance equated in this manner mean confidence was 4.66 for students in the training group and 5.09 for high achievers in the no training group, $t(12) = 2.78$, $p < .05$.

Another approach to measuring CJ accuracy is to contrast the difference between mean confidence of items answered correctly and mean confidence of items answered incorrectly. The assumption is that an ability to discriminate known from unknown information will be reflected in the difference between these two means. Although only the mean difference may be reported, it is more appropriate to consider this difference relative to the manner in which the subject used the confidence scale. For example, a small difference between means may not necessarily reflect poor discrimination of right and wrong items if a subject used a limited range of scale values when making a discrimination. A measure which takes this into account is the confidence accuracy quotient (CAQ) (see King, et al., 1980, and Zimmerman, Broder, Shaughnessy, & Underwood, 1977), and has been found to correlate positively with performance on classroom multiple-choice exams (Shaughnessy, 1979). The CAQ is a ratio, the numerator of which is the difference between the mean confidence assigned to right items and the mean confidence assigned to wrong items. The denominator is the square root of the pooled variance of the subject's confidence judgments for right and wrong answers. The CAQ is analogous to d' in a signal detection analysis and takes on a value of zero when a subject cannot discriminate right and wrong answers. The CAQ is particulary applicable in an absolute judgment task but is affected by guessing in a forced-choice procedure (see

Shaughnessy, 1979). For example, given a two-alternative choice situation in which a subject is asked to assign confidence judgments to selected answers, a certain proportion of answers given a very low confidence rating (approximately .50) will be correct by chance. Confidence values assigned to these items will tend to lower the mean confidence of right answers, lessening the difference between mean confidence for right and wrong answers. The CAQ score, consequently, will be lowered. Although this problem is particularly severe when only two alternatives are used, as was true for the present laboratory CJ tasks, the severity of the problem decreases with increasing numbers of alternatives. Given that in the present experiment there were four alternatives for each question on the classroom exam, and, thus, scores would be expected to be less affected by chance than they would in the two-alternative situation, performance of trained and not trained students was compared using the CAQ measure. As revealed in Table 3, mean CAQ was slightly greater for trained than not trained high achievers; however, this difference was not statistically significant, and, all but disappeared when students were matched on exam performance (1.12 vs. 1.15).

As would be expected given the differences found in mean confidence, high achievers differed in their use of the 6-point confidence scale as a function of training. Students in the training group were more likely to use low scale values and less

likely to use high scale values relative to students in the no training condition. When students were matched in terms of final exam performance (n = 13), the mean scale use for low (1-2), intermediate (3-4), and high (5-6) scale values was 5.08, 13.77, and 30.69 for trained students, and 3.00, 7.92, and 39.00 for not trained students. The interaction in a matched groups analysis between the levels of training and levels of confidence was significant, $F(2,24) = 7.12$, $p < .05$. High achievers, therefore, showed an effect of training given one month earlier. This effect, however, was reflected mainly in the manner in which trained students used the confidence scale. As assessed by the CAQ measure, there was little evidence that training improved high achievers' ability to discriminate between right and wrong answers.

Low achievers in the training condition scored higher on the final exam than did similar students in the no training condition. This difference was marginally significant, $t(30) = 1.87$, $p < .10$. Nevertheless, a difference in performance was seen between these groups of students on the first two exams taken in the class (see Table 3). Therefore, the superior performance by students in the training condition is most likely the result of differences due to sampling rather than to an effect associated with participation in the training component of this experiment. Although training appeared to reduce overconfidence of low achievers, this effect was only marginally

significant (t = 1.71, p < .10) and its interpretation problematic due to the sizable difference in number correct on the exam. When low achievers were matched on their exam performance (n=9), mean confidence was 3.92 for trained students and 4.11 for not trained students, but did not differ statistically (p > .05).

Although mean CAQ of low achievers was greater for trained than not trained students (see Table 3), the difference was not reliable statistically. Matching low achievers in terms of number correct on the exam also yielded nonsignificant differences, although they were in the direction predicted by a training effect. With exam performance equated within this achievment group, mean CAQ of trained students was 1.00 and that of not trained students, .71 (p > .10). Finally, low achievers also did not differ significantly in their use of the 6-point confidence scale as a function of training, but differences in mean use across the confidence scale were in the same direction as those observed for high achievers for the same comparison. Specifically, when matched on final exam performance, mean use of low, intermediate and high confidence levels by low achievers in the training group was 12.11, 14.44 and 22.22, respectively. Mean use by students not having training was 9.00, 14.78, and 24.89, respectively. Among low achievers, therefore, to the extent that training generalized to classroom performance, the effects were relatively small and nonsignificant. It can be

emphasized, however, that these results and conclusions are based on data obtained with relatively few numbers of subjects (e.g., only nine matched pairs were obtained among low achievers who differed on the training variable). Finally, in agreement with observations made by Shaughnessy (1979), average CAQ scores were greater for high than low achievers (see Table 5).

Performance on FK test

Following both pre- and post-test CJ tasks administered via the laboratory computer, all students were given a paper and pencil FK task patterned generally after that of Hart's (1965). It was of interest whether high and low achievers would differ on this type of memory monitoring task, and whether training, to the extent that it was effective in modifying CJ performance, would generalize to this test. Mean total times to do the post-training FK task based on reports by students in the training and no training groups were 48.55 min. and 51.54 min., respectively ($p > .05$).

Performance by high and low achievers on pre- and post-training FK tests is summarized in Table 4.

Insert Table 4 about here

Because failure to attempt to answer questions in the recall stage of the FK task may reflect withholding of correct answers for which a subject is uncertain, and because these same items

may receive high FK ratings, the result of withholding may be to bias proportion correct recognition of incorrect items given high FK ratings. Therefore, number of items for which recall was not attempted was examined for all subjects, and three subjects (all low achievers) were not included in the FK analysis because they failed to attempt answers for more than 15 of the 100 questions on one or both of the FK tests. Only two subjects could be replaced using the small pool of surplus subjects made available for this type of situation, and, therefore, FK results are based on 31 low achievers. Unexpectedly, low achievers both recalled and recognized significantly more answers on the pre- and post-training FK tests than did high achievers. The attempt to equate retention of general information questions for low and high achievement groups, which had proved so effective for the computer CJ task, apparently was not successful for the FK task.

Within both achievement groups, proportion correct recognition of items recalled incorrectly was compared for low (1-3) and high (4-6) FK ratings. Unexpectedly, on the pretest, there was little difference in recognition memory performance as a function of FK rating for either low or high achievers. Low achievers actually recognized slightly fewer answers following a high FK rating than following a low FK rating. On the post-training FK task, recognition memory was significantly better for high FK ratings than for low ratings, but only for high achievers, $\underline{t}(31) = 2.14$, $\underline{p} < .05$. Moreover, training did not

appear to be a factor in FK performance as assessed by recognition of nonrecalled items. The difference in recognition memory performance between low and high post-test FK ratings was not substantially greater for students in the training group than those not in the training group, either overall or within low and high achievement groups. Students in the training condition correctly recognized .37 answers following a low FK rating (1-3) and wrong recall, whereas they recognized .48 answers after a high FK rating (4-6). Not trained students recognized .35 answers given a low FK rating and .45 answers following a high FK rating.

Performance on the FK post-test was also examined by looking at overall proportion correct recognition as a function of FK ratings for both correctly and incorrectly recalled answers. In other words, probability of correct recognition, ignoring whether an answer was recalled correctly or not, was calculated for each level of reported confidence. The resulting "calibration" curves revealed no apparent differences as a function of training, either overall or within achievement levels. However, when only frequency of use of the confidence scale was considered, a shift in use of the levels of the scale was apparent which was similar to that seen when training was assessed in the computer-run CJ task. Specifically, when predicting recognition trained subjects used the highest confidence level (6) less often than did not trained subjects (.30 vs. .38), and, conversely, used the lowest

level (1) more frequently than did not-trained students (.30 vs. .25). This difference between trained and not trained students was apparent for both high and low achievers to approximately the same degree. Therefore, a slight effect of training CJ accuracy was apparent in the FK task. However, what effect was present was limited to a modest shift in use of the level of the confidence scale as a function of training which did not vary with achievement level. It is likely, however, that the present procedure did not permit a sufficiently sensitive test by which to assess generalization of training, a point that will be developed in the discussion that follows.

Discussion

When confidence judgments are made for answers given to general knowledge questions, the most salient finding of a host of studies is that people are overconfident (see Lichtenstein, et al., 1982). Students, as well as nonstudents, most people, in fact, tend to overestimate the likelihood that their answers are correct. The implications of this metamemorial bias are obvious in situations where individuals are called upon to acquire new information. For instance, it seems reasonable to assume that initial, as well as subsequent, attempts to acquire information are based on individuals' assessment of the state of their knowledge. Learning strategies to be efficient would seem to depend on accurate judgments of this kind. Nevertheless, to the extent that learners are overconfident regarding what they know,

learning is not likely to be efficient, and, is likely to lead to performance that falls considerably short of that which is expected on the basis of the judgment, "I know this."

The goal of the present study was to examine the effect of a modest training exercise on improving the appropriateness of confidence judgments of answers given to general information questions, as well as to investigate the effect of training on confidence judgments made in the context of a regularly scheduled classroom examination. The results were encouraging in that students who attended the training session were significantly less overconfident when answering questions on a post-training laboratory task, and an effect of training was found to generalize to the classroom. Generalization of training occurred even though the class exam followed training by one month. However, the effect of training was not the same for low and high achieving students, and, the results obtained from the classroom examination are problematic given that some students, namely high achievers, were now found to be substantially underconfident in judging what they knew. These major results, as well as several related findings, require comment and elaboration.

An initial finding was that low achieving students were significantly more overconfident than high achieving students. With item difficulty held relatively constant, mean confidence of low achievers was greater than that of high achievers. This result was unexpected for two reasons. First, previous studies

have suggested that students who are more intelligent, or who have greater expertise in a subject area from which questions are drawn, are not better calibrated, or necessarily less overconfident, than are students of lesser intelligence or who lack expertise in an area (Lichtenstein & Fischhoff, 1977). Second, in pilot work carried out prior to the present study, students of different achievement levels, similar to those of the actual study, were tested and calibration curves did not differ. Why, then, do the present results indicate that students who are doing very poorly in a academic setting are likely to be more overconfident about what they know than students who are doing very well?

Several reasons can be suggested for why the present findings might differ from those previously reported, including those obtained as part of the pilot work. First, with regard to the findings of other investigators, specifically those of Lichtenstein and Fischhoff (1977), it can be pointed out that these investigators examined CJ accuracy of individuals for a relatively limited range of intelligence, namely that existing between "usual" volunteer undergraduate students and psychology graduate students. Furthermore, comparisons between these groups of students were made when item difficulty was equated by sampling items from a larger set of items for which overall performance differed between groups of subjects. In other words, when <u>items</u> were matched for difficulty graduate students were not

better calibrated than the usual undergraduate students. In the present study, the difference between the general aptitude of low and high achievers, essentially those students who were consistently doing very poorly in an introductory college course and those who were doing very well, is likely to be significantly greater than that between typical undergraduate volunteers and psychology graduate students. Moreover, the present procedure, which involved matching overall test performance, provides a more appropriate test of the effect of different aptitude level on CJ performance than is the case when items from a larger item set are matched. When performance is equated for subsets of items the effect of overall test context on CJ performance is a confounding factor. The present results suggest that when test context is the same, low achievers are less well calibrated, that is, more overconfident than are high achievers. As was shown, low achievers are more likely than high achievers to use the most extreme levels of the confidence scale, with this difference being greatest for the high end (1.0) of the scale.

That the present results do not agree with findings in the pilot study are more difficult to deal with. There were numerous small differences in procedure between the pilot study and the present experiment. For instance, the study was conducted in a setting somewhat less formal than that used here. Items were presented on cards, rather than using a microcomputer, and instructions were less formalized than the present ones. In

addition to obvious differences in the "atmosphere" under which testing took place, the overall level of performance in terms of number correct was greater in the pilot study than here. It is possible that differences in over-underconfidence between students of different aptitude levels may be seen at some levels of proportion correct and not others. This would occur, for instance, if a particular level of overall performance had a psychologically greater impact for one group of students than another. It has been suggested, for example, that individuals may have an "ideal" test, one whose difficulty level leads to neither under- nor overconfidence (see Lichtenstein, et al., 1982). Analogously, different "types" of individuals, for instance low and high achievers, may respond differently depending on the overall difficulty of the test even when overall proportion correct between groups of these individuals is the same. Those researchers who are interested in individual differences in CJ performance may want to consider alternative ways to assess CJ accuracy than that associated with calibration of probabilities based on tasks of this kind.

It is clear from the present results that training was more effective for low than high achievers, at least as assessed by the laboratory-based CJ test. It is also obvious that the present design does not permit one to evaluate the contributions of the various components of the training exercise. Whether personal feedback in the form of a calibration curve, learning

what is appropriate evidence to support different levels of confidence, recognizing the deceptive nature of certain items, realizing the need to weight carefully evidence for the validity of an answer, or the clearly implied message to "be careful" when assigning confidence ratings, was most effective in reducing overconfidence of low achievers is not clear. Given the lack of success in improving the appropriateness of confidence judgments that has previously been reported, and which generally involved experiments focusing on only one procedure, it may be that a combination of the present components was what was most effective.

Training would be expected to have more of an impact on low than high achievers if high achievers were already doing those things that were taught as part of the training exercise. Several findings suggest that this is the case. First, high achievers were significantly less overconfident to begin with, as seen in the calibration curves of Figure 1. Second, data obtained as part of the training exercise show that low achievers were significantly less able than high achievers to match an appropriate reason with a specific level of confidence. In fact, high achievers made very few errors on this reasons test, averaging more than 9 out of 10 correct (see Table 2). Moreover, high achievers were less likely than low achievers to give extremely high confidence ratings to wrong answers. This was apparent in the training session (Table 2) as well in the

responses seen to so-called deceptive items on the CJ post-test. For example, high achievers did not differ in the number of times they were "deceived" by these item types as a function of training; whereas, low achievers without training were much more likely to be deceived than were low achievers with training. Finally, it is clear that an effect on CJ performance of requiring students to provide reasons for selecting an answer and assigning a particular level of confidence, was greater for low than high achievers (see Figure 3). It may be suggested that high achievers are more likely than low achievers to engage spontaneously in the kinds of mental cross examination necessary for appropriate CJ performance. High achievers, in other words, more often than low achievers, ask themselves the kinds of questions which are elicited from low achievers only through prompting. As a practical measure, therefore, the present results suggest that overconfidence of low achievers can be reduced by directing them to produce reasons why they have selected a particular answer as being correct.

Two other findings of the present study require comment. First, there was no significant effect of training on FK performance. Trained subjects were no more likely than not trained subjects to predict accurately recognition of nonrecalled facts. Second, CJ performance associated with the classroom exam revealed an effect of training, but this appeared to be limited to changes in frequency of use of the levels of the confidence

scale, which affected over-underconfidence, and did not apparently lead to significantly greater discrimination between right and wrong answers. However, both the FK tasks and the classroom exam may not have provided sufficently sensitive tests of generalization of training.

In hindsight, the procedure associated with the FK task could be improved upon. In the standard FK procedure subjects are asked to make FK judgments only for those items for which recall is unsuccessful or for which recall has been indicated as wrong. However, the present task required subjects to make predictions of later recognition without knowing whether a response that was produced was right or wrong. No doubt in some cases subjects would assume incorrectly to have recalled the right answer and be led to predict recognition with a high degree of certainty, only to find among the recognition alternatives what they then realize is the right answer. Also, the fact that subjects in the present situation were allowed to take the FK task without being monitored may have led to careless responding or even "looking ahead" behavior since the recognition test with the correct answers was included in the same package of materials as the recall test.

The only findings of note with regard to the FK task were that high achievers were more accurate in their FK judgments than low achievers on the second FK test, and, an analysis of the frequency with which various levels of the FK scale were used

indicated that subjects in the training group were less likely to use the extreme high end of the scale. Thus, some evidence of generalization of training was observed although the task as presented did not appear appropriate to adequately validate this aspect of the training effect. Moreover, Nelson and Narens (1980) have suggested that the FK procedure as typically presented confounds the subject's metamemorial knowledge of nonrecalled items and the subject's "know/don't know threshold." They recommend, instead, a procedure relying on relative FK decisions involving paired-comparisons of nonrecalled items that leads to a rank ordering of nonrecalled item in terms of their predicted likelihood of recognition.

The findings with respect to the classroom exam are also difficult to interpret due to the extremely high performance by the high achievers on this task and because relatively few numbers of subjects in each achievement level were able to be observed. As overall number correct on a CJ task increases, overconfidence is often reduced to the point that underconfidence is seen when proportion correct is very high (Lichtenstein & Fischhoff, 1977). Therefore, it is to be expected that, overall, high achievers would be less overconfident than low achievers on the final exam. In fact, high achievers not in the training group were neither under- nor overconfident. The effect of training for high achieving students was to reduce overall mean confidence on the exam, and, consequently, to produce

underconfidence. As assessed by the CAQ measure, however, trained high achievers were no more accurate when discriminating right and wrong answers than not trained high achievers. Nevertheless, such a result may be difficult to obtain given the relatively high number of correct answers. There were clearly very few items for which these students would have the opportunity to be wrong and be led to assign a low level of confidence that the answer was right. It is also to be expected that guessing will reduce somewhat the overall difference between mean confidence of right and wrong answers. Therefore, it is perhaps not surprising not to see a training effect as assessed by the CAQ measure. These results should not, however, take away from the fact that, for high achievers, training had a significant effect on CJ performance as assessed in a classroom task one month later. Had the classroom task been more difficult it is possible that an effect of training could be more appropriately evaluated; but, of course, students were selected to participate because they had been doing very well on the classroom tests.

Although effects of training appeared to be present for low achievers on the classroom exam, these effects were small and nonsignificant. Nevertheless, it is important to note that some effects were seen. This seems especially significant given the relatively modest investment in training and given that no deliberate association was made between participation in the

training session one month earlier and the classroom test. It appears worthwhile to consider introducing aspects of the training program, at least to low achieving students, in a way that would permit a clearer examination of training effects on classroom performance. Should these students become fully aware of the metacognitive bias that is present when judging what is known, it may have the effect of prompting more careful evaluation of evidence for answers retrieved from long-term memory, and the realization that less is often actually known than is generally assumed. Training, in other words, might have the important effect of motivating low achieving students to work harder to determine unambiguously what it is that they know.

Finally, although the present experiment was successful in improving the appropriateness of confidence judgments given to answers retrieved from long-term memory, it must be admitted that this effect basically was limited to reductions in overconfidence that were obviously the result of changes in the manner in which the confidence scale was used. The clearest result of training was that subjects were less likely to say that they were absolutely sure that an answer was right. Although this outcome was obviously one that was being sought, the results do not necessarily speak to the issue of whether training led to an increased sensitivity to what is a right or wrong answer. Use of the CAQ measure in the context of the class exam was one approach

to this question, although the results were not conclusive. Other approaches to this question have involved other measures, including those associated with signal detection theory (e.g., Ferrell & McGoey, 1980). Nevertheless, to properly assess CJ accuracy using these measures it is often necessary to obtain confidence judgments for hundreds of responses. Such a procedure is very inefficient and those who wish to undertake the investigation of CJ accuracy might more appropriately consider different approaches to this problem. For example, more systematic use of the reasons test designed for the present training session, or a content analysis of reasons given by trained subjects prior to and after training, might reveal more clearly the processes underlying sensitivity to right and wrong answers. Finally, it is important that we attempt to elucidate the link between metamemorial judgments and acquisition of knowledge. Assumptions regarding the nature of these links are much easier to make than they are to verify. While it can be assumed, for example, that being able to judge appropriately what is known will lead students to perform better on tests of memory than those who continue to believe that they know more than they do, valid evidence for this assumption is sparse.

Reference Notes

1. Arkes, H. R., Lai, C., & Hackett, C. A. Two methods for reducing overconfidence. Paper presented at the 54th Annual Meeting of the Midwestern Psychological Association, Minneapolis, May, 1982.

2. Nyberg, S. E., Englebrekt, B. J., Zechmeister, E. B., & Ruble, N. Can metamemory be calibrated? Paper presented at the 51st Annual Meeting of the Midwestern Psychological Association, Chicago, May, 1979.

References

Bisanz, G. L., Vesonder, G. T., & Voss, J. F. Knowledge of one's own responding and the relation of such knowledge to learning. Journal of Experimental Child Psychology, 1978, 25, 116-128.

Bransford, J. D., Stein, B. S., Vye, N. J., Franks, J. J., Auble, P. M., Mezynski, K. T., & Perfetto, G. A. Differences in approaches to learning: An overview. Journal of Experimental Psychology: General, 1982, 111, 390-398.

Ferrell, W. R., & McGoey, P. J. A model of calibration for subjective probabilities. Organizational Behavior and Human Performance, 1980, 26, 32-53.

Flavell, J. H. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 1979, 34, 906-911.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 552-564.

Hart, J. T. Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 1965, 56, 208-216.

Hart, J. T. Methodological note on feeling-of-knowing experiments. Journal of Educational Psychology, 1966, 57, 347-349.

Hart, J. T. Second-try recall, recognition, and the memory-

monitoring process. Journal of Educational Psychology, 1967, 58, 193-197.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. Judgments of knowing: The influence of retrieval practice. American Journal of Psychology, 1980, 93, 329-343.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal o. Experimental Psychology: Human Learning and Memory, 1980, 6, 107-118.

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. Training for calibration. Organizational Behavior and Human Performance, 1980, 26, 149-171.

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.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Nelson, T. O., & Narens, L. A new technique for investigating the feeling of knowing. Acta Psychologica, 1980, 46, 69-80.

Owings, R. A., Petersen, G. A., Bransford, J. D., Morris, C. D., & Stein, B. S. Spontaneous monitoring and regulation of learning: A comparison of successful and less successful fifth graders. Journal of Educational Psychology, 1980, 72, 250-256.

Shaughnessy, J. J. Confidence judgment accuracy as a predictor of test performance. Journal of Research in Personality, 1979, 13, 505-514.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. A recognition test of vocabulary using signal-detection measures and some correlates of word and nonword recognition. Intelligence, 1977, 1, 5-31.

Table 1

Item Difficulty Distributions of Test Lists

Used by High and Low Achievers*

		Number of Errors							
Group	Form	0-2	3-4	5-6	7-8	9-10	11-12	13-14	15-17
High	A	32	21	23	11	7	3	1	2
	B	41	19	14	10	10	4	1	1
	Total	73	40	37	21	17	7	2	3
Low	A	30	18	17	10	15	5	4	1
	B	36	18	17	14	10	5	0	0
	Total	66	36	34	24	25	10	4	1

*Table reports the number of test items at each difficulty level. Data are based on 16 high achievers using each test form, and 17 low achievers using form A and 15 low achievers using form B.

Table 2

Summary of Performance in Training Session

I. Reasons Test	$\bar{X}$ Correct*
Low (n=16)	7.94
High (n-16)	9.19

*$t(30)=2.12$, $p<.05$

II. "Deceptive" Questions Test

	$\bar{X}$ Correct**	Number of 10W Items	Number of Subjects with 8-, 9-, or 10W
Low (n-16)	5.81	13	13
High (n-16)	7.12	5	9

**$t(30)=2.19$, $p<.05$

Table 3

Exam Performance of High and Low Achievers in Training (T) and No Training (NT) Conditions

Measure	HT(n=16)	HNT(n=15)	LT(n=16)	LNT(n=16)
$\bar{X}$ Correct				
Exams 1 & 2	42.81	42.73	29.16	26.47
Exam 4	44.19	43.33	34.75	29.69
$\bar{X}$ Confidence	4.71	5.10	4.09	4.03
Over-Under Conf.	-.08	.00	.02	.11
CAQ	1.32	1.10	.90	.70

Table 4

Performance on FK Tests

by Low and High Achievers

	Achievement Level	
Pretest	Low (n=31)	High (n=32)
Prop. Recall	.48	.41
Prop. Recognition	.67	.62
Prop. Recog./FK(1-3)	.41	.39
Prop. Recog./FK(4-6)	.38	.42
Post-test		
Prop. Recall	.50	.43
Prop. Recognition	.66	.65
Prop. Recog./FK(1-3)	.34	.38
Prop. Recog./FK(4-6)	.40	.52

Footnotes

1 A principal source of general information questions was an extensive list kindly provided by Decision Research, Eugene, Oregon.

2 A transcript of the training session as well as copies of the materials used during training can be obtained by writing to the first author.

3 A content analysis was performed on reasons that subjects gave for choosing a particular answer. Qualitative differences were apparent between high and low achievers, but should be expected in a free response situation due to correlated differences in verbal abilities of these groups of subjects.

4 Analyses were also performed using several other measures associated with calibration of probabilities, for example, measures of calibration and resolution (see Lichtenstein et al., 1977, for a definition). However, these measures are likely to show considerable chance fluctuation unless based on literally hundreds of responses (see Lichtenstein & Fischhoff, 1980). The CJ tests in the present experiment contained only 100 items. Moreover, these measures are often moderately to highly correlated with the O-U measure (see Nyberg, et al., Note 2). For example, calibration scores based on pretest performance in

the present experiment correlated .71 with O-U scores (n = 64). In short, these other measures did not contribute significantly to the interpretation of the present results. For example, no statistically significant changes in resolution were found in any analyses except one. Mean resolution decreased significantly between pre- and post-test for low achievers not in the training group.

5 In order to obtain a measure of over-underconfidence the 6-point confidence scale was treated as an equal interval (.15) probability scale with .25 assigned to a confidence of 1.0 (guessing), .40 assigned to a confidence of 2, and so forth.

Figure Captions

Figure 1. Calibration curves of low and high achievers on the CJ pretest.

Figure 2. Calibration curves of low and high achievers who were asked for reasons, and of low and high achievers who were not asked for reasons. Low and high achievers in both the reasons and no reasons groups were matched on number correct.

Figure 3. Calibration curves of low achievers who were asked reasons and who were not asked reasons, and of high achievers who were asked reasons and who were not asked reasons. Low achievers in the reasons and no reasons groups, as well as high achievers in these groups, were matched on number correct.

Figure 4. Calibration curves of students in training and no training groups based on CJ post-test performance.

Figure 5. Pre- and post-test differences in calibration curves of low and high achievers as a function of training.

Figure 6. Calibration curves of low and high achievers in training and no training conditions based on CJ post-test performance.

FIGURE 1

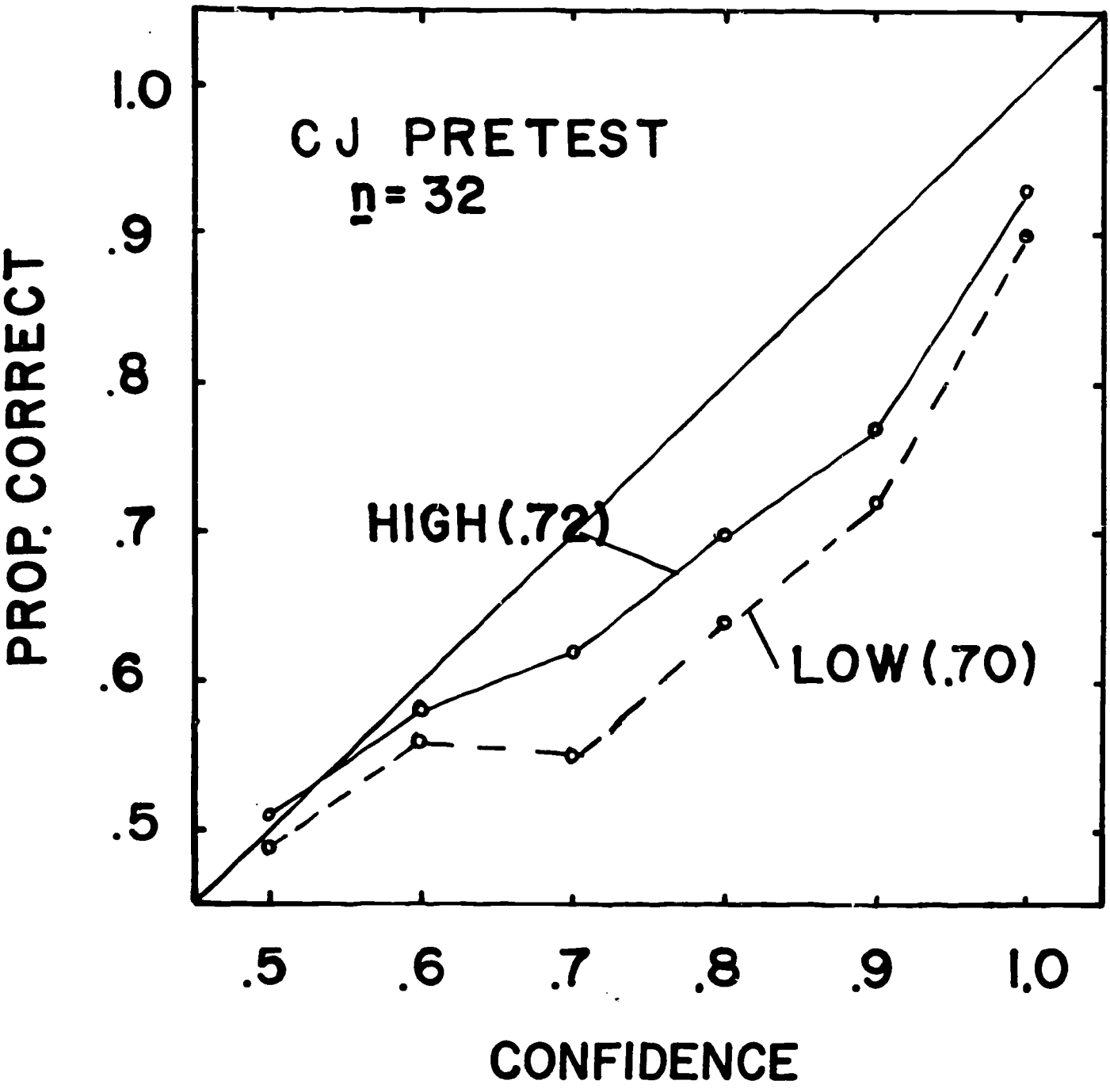

CJ PRETEST
n = 32
HIGH (.72)
LOW (.70)
PROP. CORRECT
CONFIDENCE
1.0
.9
.8
.7
.6
.5
.5
.6
.7
.8
.9
1.0

FIGURE 2

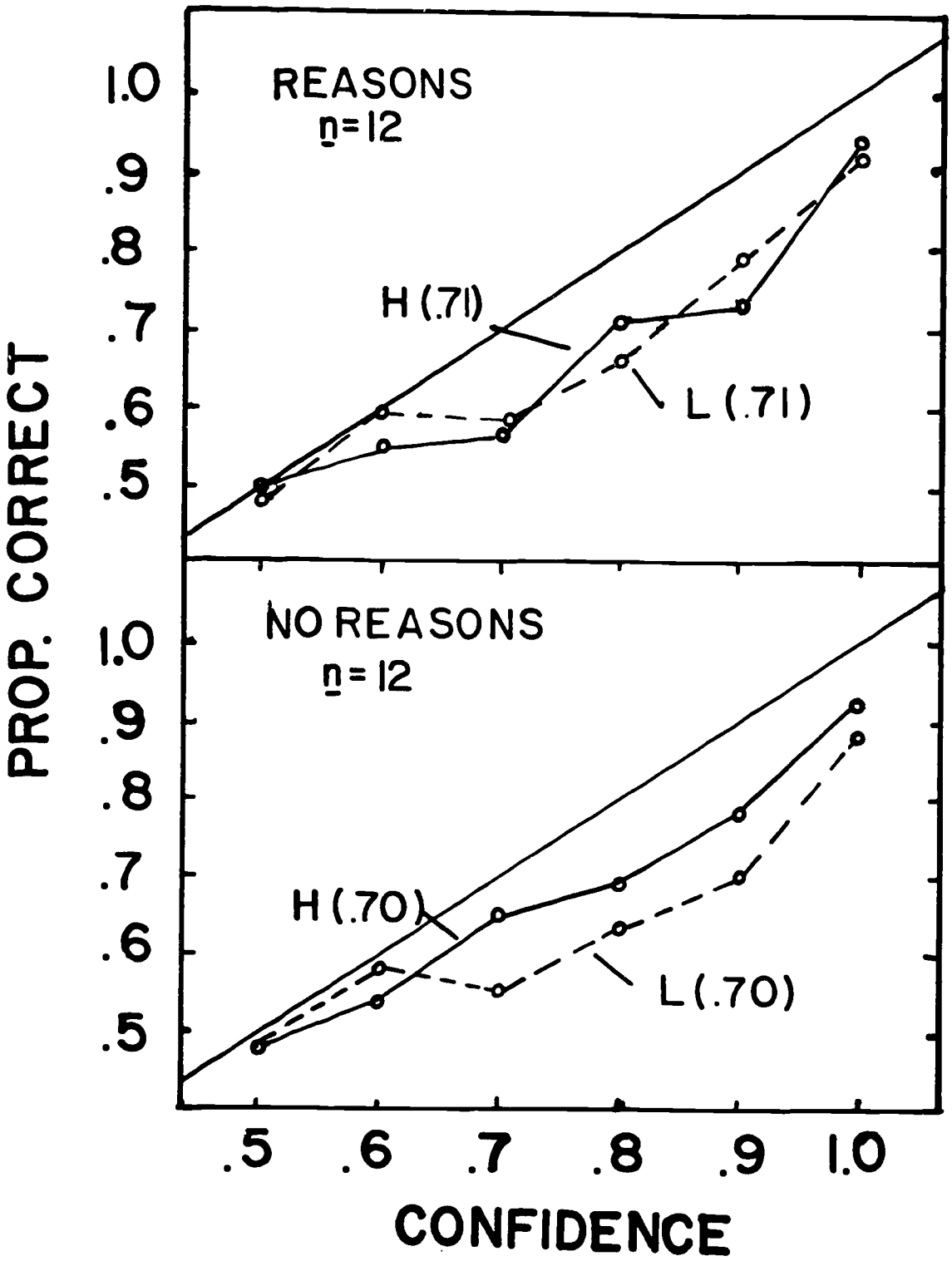
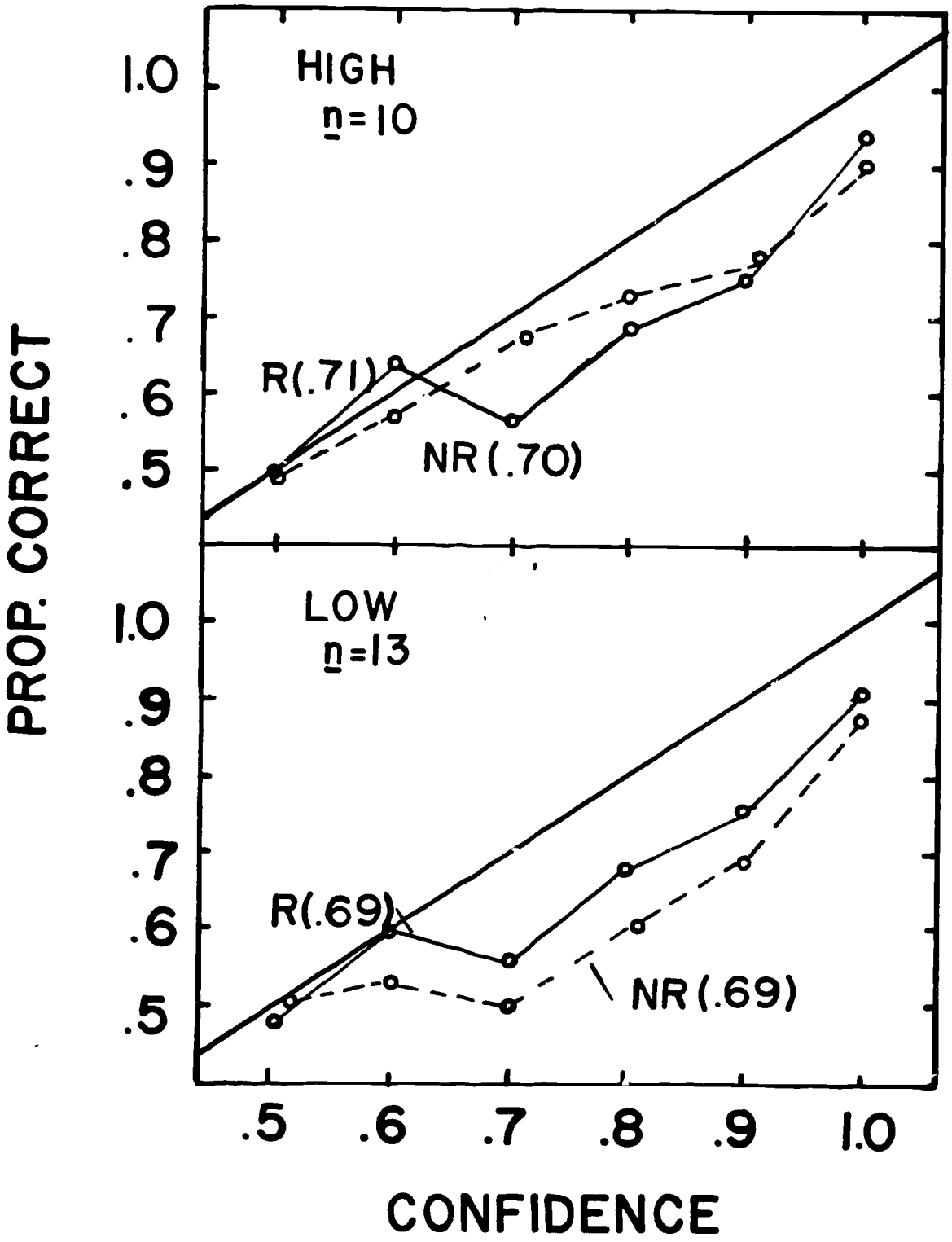

REASONS
n=12
H (.71)
L (.71)
NO REASONS
n=12
H (.70)
L (.70)
PROP. CORRECT
1.0
.9
.8
.7
.6
.5
.5
.6
.7
.8
.9
1.0
CONFIDENCE

FIGURE 3


HIGH
n=10
LOW
n=13
R(.71)
NR(.70)
R(.69)
NR(.69)
PROP. CORRECT
CONFIDENCE
1.0
.9
.8
.7
.6
.5

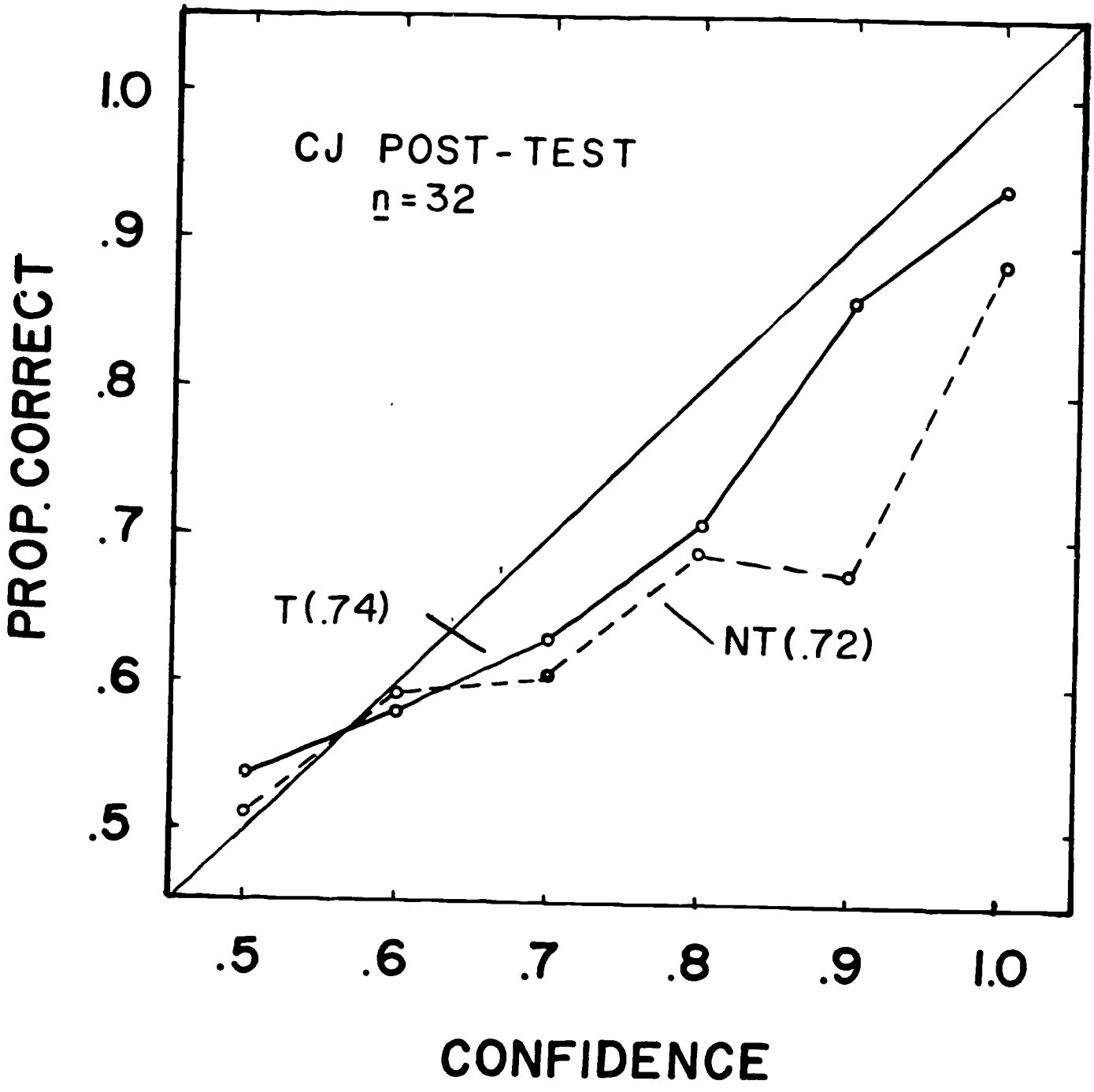

CJ POST-TEST
n = 32
PROP. CORRECT
1.0
.9
.8
.7
.6
.5
T(.74)
NT(.72)
.5
.6
.7
.8
.9
1.0
CONFIDENCE

FIGURE 5

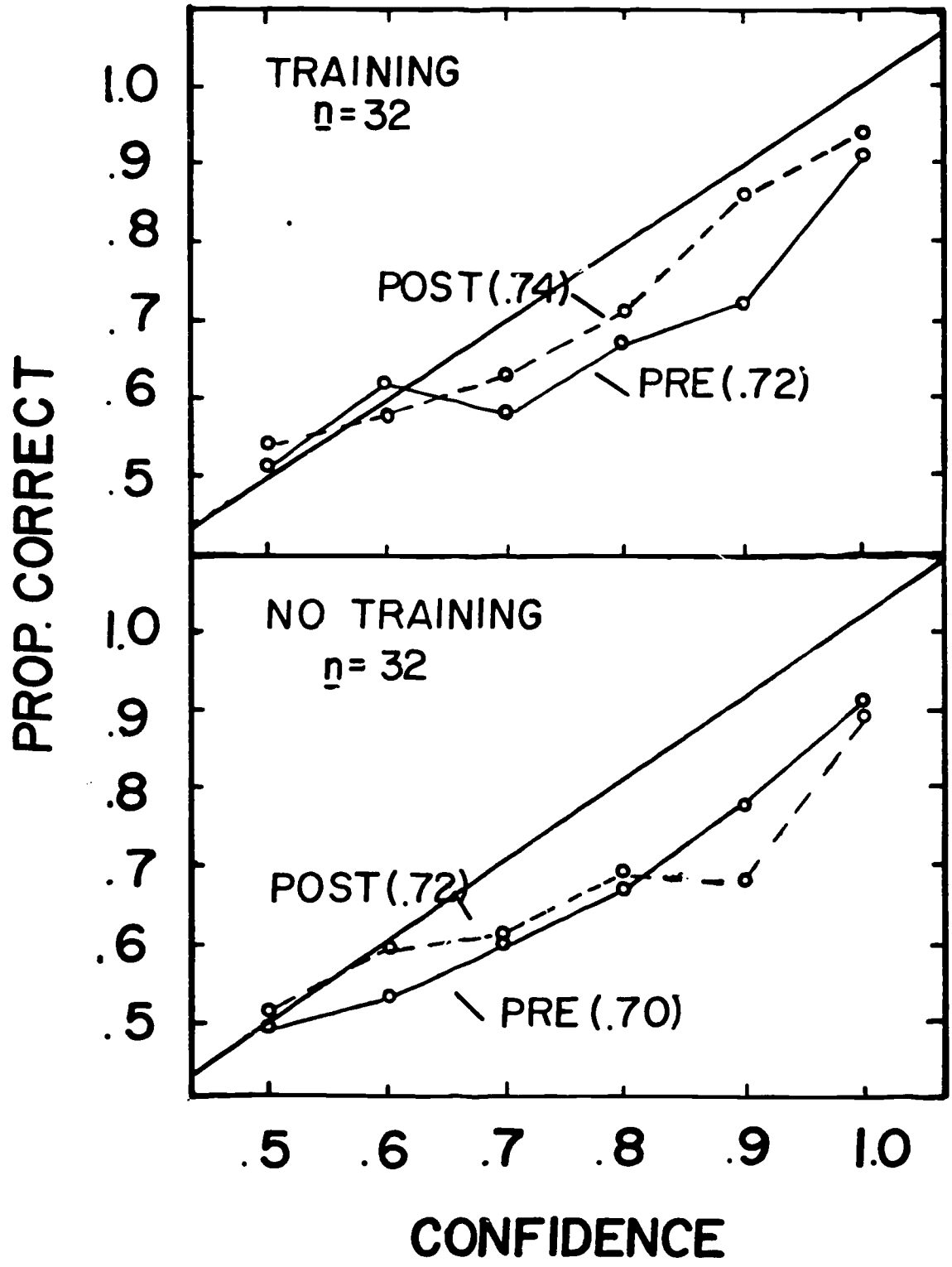
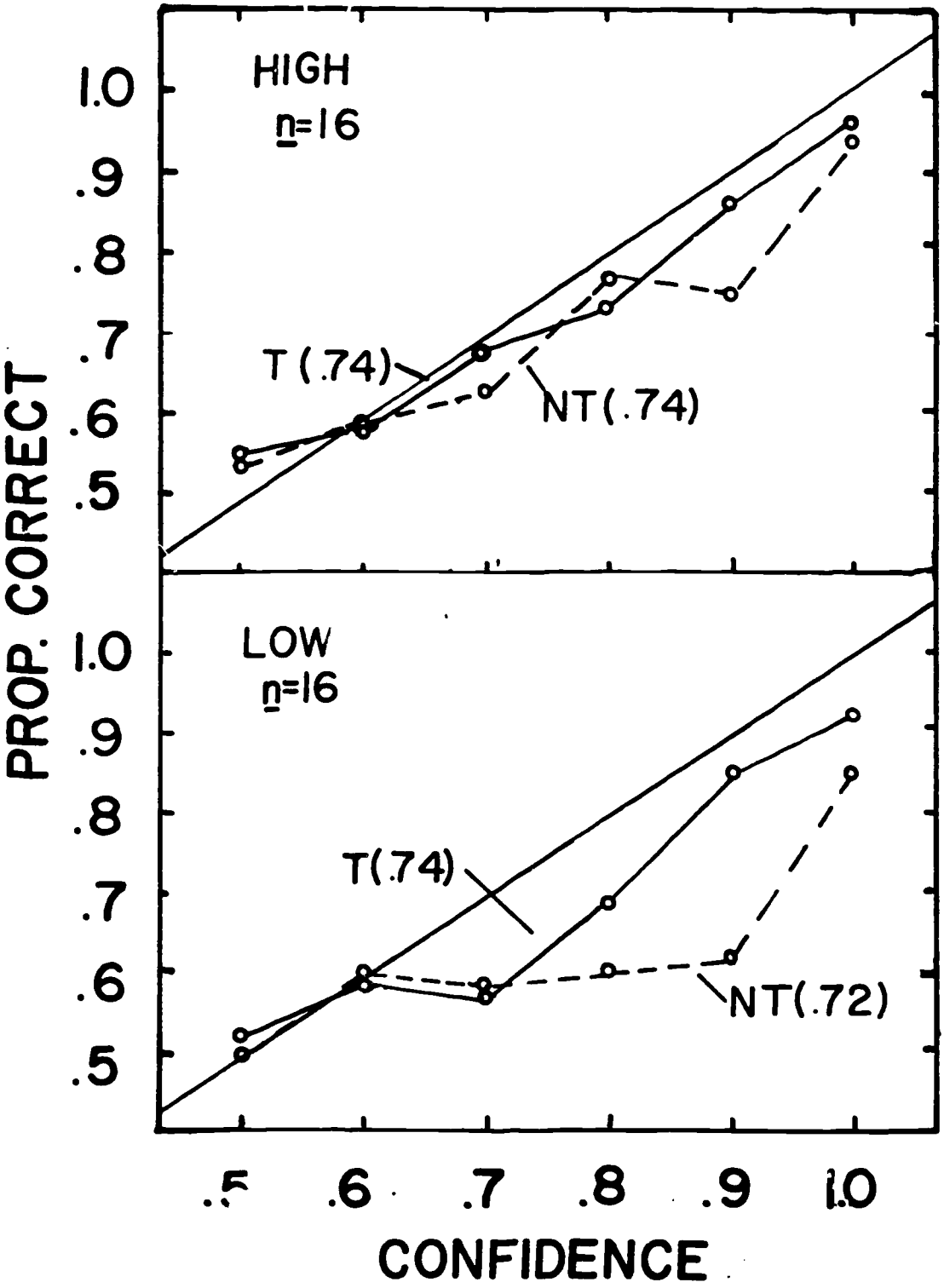

TRAINING
n=32
POST(.74)
PRE(.72)
NO TRAINING
n=32
POST(.72)
PRE(.70)
PROP. CORRECT
1.0
.9
.8
.7
.6
.5
.5
.6
.7
.8
.9
1.0
CONFIDENCE

FIGURE 6


HIGH
n=16
T (.74)
NT(.74)
LOW
n=16
T(.74)
NT(.72)
PROP. CORRECT
1.0
.9
.8
.7
.6
.5
.5
.6
.7
.8
.9
1.0
CONFIDENCE