1986

Metacognition and Aging: The Effect of Monetary Incentive on Confidence Ratings

Kathryn A. Markell
Loyola University Chicago

3 ✓

METACOGNITION AND AGING: THE EFFECT OF MONETARY INCENTIVE ON CONFIDENCE RATINGS

by

Kathryn A. Markell

A Dissertation Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

August

1986

ACKNOWLEDGMENTS

I would like to thank the members of my committee for their patience and support: Dr. Eugene Zechmeister, my committee chairperson, Fr. Daniel O'Connell S.J., and Dr. Jill Nagy Reich. I am grateful for the encouragement given to me by Dr. Jeanne Foley throughout my graduate career. I would also like to thank James Carr for his encouragement and support.

VITA

The author, Kathryn Ann Markell, is the daughter of Mary Jennings Markell and Wilbur Markell. She was born on May 17, 1956 in St. Paul, Minnesota.

Her elementary education was obtained at McKinley Elementary School in Owatonna, Minnesota. Her secondary education was completed in 1974 at Owatonna High School, Owatonna, Minnesota.

In September, 1974, Ms. Markell entered the College of St. Benedict, receiving the degree of Bachelor of Art in English and Psychology in May 1978.

In September, 1978, Ms. Markell was granted a graduate assistantship in psychology at Loyola University of Chicago, enabling her to complete the Masters of Arts degree in 1981. Ms. Markell was granted a Schmitt fellowship in 1981 to pursue her dissertation work. She has been teaching since 1982, first at The College of St. Thomas, and then at St. John's University, both in Minnesota.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CONTENTS FOR APPENDICES

INTRODUCTION

Cognitive psychologists have recently been concerned not only with how people comprehend and memorize information, but also with the ability people have to know how much information they have stored in their memory, and what information they may be able to retrieve. The term metacognition refers to a person's knowledge about anything related to cognitive processes (Brown, 1978). Metacognitive skills, for instance, enable people to discriminate between information they "know they know" and information that they do not know or are less sure of.

The metacognitive skills of children and college students have been studied under many conditions, but until recently, few studies have concentrated on the metacognitive abilities of older adults. Much of the emphasis of research on metacognitive ability has therefore been concerned with how student's knowledge about what they do and do not know affects their efforts in learning material and their performance on tests, and how student metacognitive abilities may be improved. These may be the easiest and most direct applications of "knowing about knowing" skills. Yet as people grow older, they are exposed to more and more information. Therefore, it becomes increasingly important for them to be able to discriminate between the information they know and the information they do not know, so that they can make accurate decisions in all aspects of their lives.

The focus of the present study is to examine one metacognitive skill, that of confidence judgments, in older adults. This study will examine whether there are any differences between young and older adults in this metacognitive skill, and whether there are differences between young and older adults in their reactions to feedback about their performance of this skill. The study will also examine the effects of "monetary incentive" on the confidence judgments of young and older adults.

LITERATURE REVIEW

Metacognitive research, in its broadest sense, includes any research concerned with "knowing about knowing". However, a large amount of the available research on metacognition centers around subject performance on three specific metacognitive tasks: judgment-of-knowing tasks, feeling-of-knowing tasks, and confidence ratings of responses to general information questions. Most research on these tasks has used young adults and children as subjects. However, recent interest in the psychology of aging has led several researchers to include older adult subjects in these metacognitive studies.

The following review is organized from a developmental perspective of subject performance, from childhood to old age, in each of these three metacognitive tasks. Although all three tasks are concerned with aspects of metacognitive ability, the research studies employing these tasks differ in the methodology that they use. Therefore, a brief review of the methodology of each task will be presented first. This will be followed by a review of the research studies that use these tasks to study metacognitive abilities for three age groups: children, young adults, and older adults.

Review of Research Methodology: Judgment-of-Knowing Predictions

Judgment-of-knowing prediction tasks test metacognitive abilities by asking subjects to make judgments about the likelihood that they will be able to remember information they are presently studying (e.g.Gardiner & Klee, 1976). For example, subjects may be given a list of words to study, and asked during study to predict how likely they think it is that they will recall each word. Later they are given a recall test on the list, and their actual performance compared to their predicted performance.

Feeling-of-Knowing Predictions

If everyone had a perfect memory, then people would be able to retrieve any information that they had stored in their memory, and the only information that they would not be able to retrieve would be information that they had never encoded. In reality, people often fail to recall information that they actually do have stored in their memory. They may "feel that they know" the information, that it is on the "tip-of-their-tongue", and yet be temporarily unable to recall it. This Tip-of-the-Tongue (TOT) phenomenon, first mentioned by William James (1893), is a "feeling-of-knowing" occurrence, and has been explored by several researchers. Naturally occurring TOT states were first studied by Woodworth (1934) for English words, and Wencl (1934) for German words. They found that when complete recall of a word is not present, people can often recall part of the word, such as a letter or syllable from the word, or something abstract about the word, such as how many syllables it has. Brown and McNeill (1966) also noted that subjects in a TOT state had knowledge of certain letters in the word, and also knew where the primary stress in the word was found. Yarmey (1973) explored verbal and non-verbal imagery codes involved in the TOT phenomenon by presenting subjects with pictures of famous people and asking them to try to recall their names. His TOT state subjects also used knowledge about the letters and syllables in the famous person's name for retrieval, and in addition to this information they relied on information about the target person's profession, or the last time that they had seen the target.

The TOT phenomenon is related to the "feeling-of-knowing" paradigm introduced by Hart (1965, 1967). Hart's hypothesis was that people may be more likely to recognize information that they feel that they know but can't recall, than information for which they have no "feeling-of-knowing". The method that Hart used to study "feeling-of-knowing" skills was to first ask subjects to recall general information items, and then, for those items not recalled, to judge whether an answer would be recognized if it was presented among several alternatives. After these feeling-of-knowing predictions, subjects were given a multiple-choice test in order to evaluate their actual recognition of these items.

Confidence Judgements

A third task used to assess metacognitive abilities is to ask subjects to make a judgment about how confident they are that an answer they have given to a question is correct. These confidence judgments are usually made by asking subjects to estimate the probability that each of their answers is correct. A "calibration" measure of how accurately subjects have made their confidence judgments (how well "calibrated" they are) is achieved by having subjects answer a series of questions, and give a confidence rating (in the form of a probability) for each. This rating reflects how confident they are that their answer is correct. A comparison is then made between the number of items receiving any given rating, and the actual proportion correct for items at that rating. For example, if subjects gave a series of their answers a .8Ø confidence rating, stating that they are 8Ø% sure that these answers are correct, then to be well-calibrated, they should actually get an average of 8Ø% of these answers correct.

An examination of research employing these tasks, with subjects from different age groups, follows. Many of these studies indicate "developmental trends" in these metacognitive tasks, with older children and young adults displaying better metacognitive skills than younger children.

Developmental Studies In Metacognition

Flavell (1971) has termed knowledge that people have about their own memory "metamemory". This term has often been used interchangeably with "metacognition", especially in developmentally based studies. Wellman (1977) has pointed out that a distinction can be made betweeen two types of metamemory: (1) Timeless facts that people could know about memory (i.e. short lists are easier to learn than long lists; young children are usually worse at memorizing lists than adults, etc.); (2) Ongoing assessments people could make about information in their own memory (i.e. this information is in my memory; this information is definitely not there, etc.). The developmental studies that follow are concerned with this second type of metacognition, exploring assessments that subjects can make about information in their own memories.

Childhood

Several studies using metacognitive tasks have compared the performance of young, school age children to that of older children or college students. Some of these studies have shown that even very young children can accurately discriminate between items they have missed and items they have identified correctly on previous testing.

For instance, Masur, McIntyre and Flavell (1973) tested the ability of first graders to judge which items they had recalled correctly and which items they had missed on a recall test. The children showed high identification accuracy: 98% of their identifications were correct for recalled items and 96% were correct for nonrecalled items.

Berch and Evans (1973) presented kindergarten and third grade subjects with a list of items. After the items were presented once, the children were given a second list of items, and asked to judge whether any of the items had been viewed on the first list. The children were also asked to make judgments about how sure they were that their identifications were correct. Their results showed that the probability of recognizing an item, for both age groups of children, varied directly as a function of their certainty judgments. This indicates that children of both ages were capable of monitoring the certainty of their recognition.

Other studies indicate that although children do show metamemory skills very early, there are some developmental trends in metacognitive accuracy. Flavell, Friedrichs and Hoyt (1970) instructed children to study a set of stimuli until they could recall all the items. It was emphasized that the children were not to signal for a recall test until they were sure that they could achieve perfect recall. Second and fourth graders were relatively accurate on this task of item recallability, but nursery school and kindergarten students frequently called for the recall test before they were able to correctly recall all the items.

Kreutzer, Leonard and Flavell (1975) asked children in grades K,1,3, and 5 to make ease of learning judgments for lists of paired words that were "opposites" versus lists of paired words of people and things they may do. The opposites list was shown to be easier to learn by all the children. However, older children were more likely than younger children to be able to identify that they would have an easier time learning the lists of opposites than the list of people-things they do. Even when younger children chose the opposites list as easier to learn, they were often not able to explain why it was easier. The young children did, however, show some knowledge of the relationship between being familiar with a list of items and the ease with which the items could be learned.

Young children appear to have some trouble not only in discriminating how easy lists of items will be to learn, but also, for prose passages, in discriminating which items will be important to learn in order to remember the main ideas of the prose passage, and which items would be less important to spend time studying (Brown & Smiley ,1977). Similiar differences in these types of discrimination abilities have been found in comparing academically successful elementary students, who can make these discriminations accurately, with less successful students, who have trouble making these discriminations (Smiley, Oakly, Worthen, Campione, & Brown, 1977; Owings, Petersen, Bransford, Morris, & Stein, 1980).

Wellman (1977) designed a study using Hart's(1965) feeling-of-knowing paradigm. He asked kindergarten, first and third grade children to recall the name of items depicted in pictures. When the children failed to recall any name, they were asked to give a feeling-of-knowing judgment about their ability to recognize the name if it was presented to them. They were also asked to judge whether or not they had ever seen the depicted item before. They were given a recognition test for all those items they had been unable to recall. There was a significant increase with age in the subjects' ability to predict which items they would and would not be able to recognize. Kindergartners were only somewhat better than chance in their predictions, whereas third graders were fairly accurate in their recognition predictions. All subjects were able to predict whether or not they had seen an item before, but the kindergartners seemed to ignore their "seen" judgments when making their feeling-of-knowing predictions.

Although even very young children seem to have some metacognitive abilities, it is clear that these abilities improve as children grow older. Wertsch (1979) has proposed a theory of how metacognitive abilities originate and develop in preschoolers based on social interaction. His theory states that the crucial element in metacognitive development is adult-child interaction in a problem-solving setting (Wertsch, 1979; Wertsch, McNamee, McLane & Budwig, 198Ø). Kontos (1983) tested this hypothesis by observing the performance of pre-school children during a problem-solving task. The preschoolers, ages 3-5, were asked to solve three peg-puzzles that required putting pegs in holes. Each child was given a 5 minute interval to attempt to solve the puzzle, and after each puzzle session, children were asked to explain how they had tried to solve the puzzle (to ascertain any "metacognitive strategy" they had used). During the second puzzle solving session, an adult, either the preschooler's mother or father, was allowed to be present during the session to help her or his son or daughter solve the puzzle by giving verbal strategy clues.

A third puzzle solving session, with the preschoolers working alone again, followed the second session. Changes in the preschoolers' metacognitive abilities were assessed by comparing their performance and strategy use on puzzle 1 to that on puzzle 3. Kontos found that children who had not received help from their parents during the puzzle 2 session improved just as much in performance and knowledge of strategy use as children who received many verbal directives from their parents. It was concluded that practice may be just as important in the development of metacognitive ability in children as adult-child interactions. The study also indicates that children may be able to increase their metacognitive abilities through training and practice.

A review of metacognitive studies employing children as subjects indicates that, although even young children (i.e., kindergarten age) show some metacognitive abilities, these abilities seem to improve with age. This is especially true for more complicated tasks, such as judging the amount of study time needed to learn a list of information, or making feeling-of-knowing judgments about information that cannot be recalled.

Young Adulthood

Most metacognitive studies have employed young college-age adults (usually 18-22 years of age) as subjects. Since the major emphasis in a college student's life is on how to study effectively, many metacognitive studies with young adults have used judgment-of-knowing tasks.

For example, Groninger (1979) presented subjects with a list of 60 words and asked them to rate, as they heard each word, their confidence that they would later be able to recall that word. They were then tested for recall, and, after the recall test, given a surprise recognition memory test. In the recognition task the subjects heard the target words interspersed with distractor words. Their task was to rate how sure they were that each word was a target word. Subjects' actual recall and recognition scores related to their prediction ratings, although they considerably overestimated their performance abilities. A second part of the study found that accurate recall predictions were also significantly related to word frequency and imagery variables.

In addition to predictions of what will be recalled, judgment-of-knowing tasks include studies examining the ability of people to discriminate what they have and have not been able to recall on previous study-test trials. These studies examine the relationship between knowledge of previous test-trial performance and a subject's performance on later study trials.

One of the studies examining this relationship is by Gardiner and Klee (1976). They gave college students a series of free-recall lists, each of which they studied and attempted to recall on a recall test. Following the recall test on the last list, the subjects were given the words from all the lists, and were asked to identify which words they had recalled on the initial tests. Even though this final test was unexpected, subjects were generally accurate in being able to distinguish previously recalled and non-recalled items.

A study by King, Zechmeister, and Shaughnessy (198Ø) also underlined the importance of previous test trials for accurate judgment of knowing ratings. They gave subjects several exposures to pairs-of-items from two lists, and then asked them to predict for each pair the likelihood that the response term would be recalled when the stimulus term was presented on subsequent trials. Half of the subjects received only study trials prior to the prediction task. The other half of the subjects received alternating study and test trials. All the subjects were also required to learn a third paired-associate list and make judgments of knowing without receiving any test trials. The results showed that prediction accuracy was consistently higher for those subjects who had been given test trials prior to the prediction task. Also, those subjects who had been given test trials on the first two lists showed a decrease in prediction accuracy on the third list for which they did not receive test trials.

One theory explaining the increase in judgment of knowing accuracy with the use of test trials has been suggested by Bisanz, Vesonder, and Voss (1978). Their hypothesis states that learners make decisions regarding memory processing based on their discimination between known and not yet known items. Accurate discriminaton between information a learner already knows, and information that he or she needs to spend more time studying, would enable the learner to shift processing from well-learned items to processing less well-learned items, thus letting the learner distribute study time efficiently.

Other theories have been suggested to explain discrimination of known and unknown information in the absence of test trials. Arbuckle and Cuddy (1969) have proposed that subjects make these judgments of knowing on the basis of perceived item difficulty, since they found ease of learning (EL) ratings to correlate highly with JK responses made during study.

Zechmeister, Christensen, and Rajkowski (198Ø) examined the relationship between EL ratings and JK ratings by presenting two groups of students with a list of facts. One group of subjects was asked to rate each sentence in terms of how easy or hard it would be for someone in general to remember, while the other group rated items in terms of how easy or hard each item would be for themselves to remember. On every trial a fact was presented followed by study, a judgment-of-knowing rating, and then, after a filler task, recall was tested. Although JK ratings were found to be better predictors of item difficulty than EL ratings, EL ratings did reliably predict item difficulty. Zechmeister et al. also found evidence to suggest that there was a difference in the JK performance between good and poor learners in that good learners were, in one condition, more accurate in "knowing what they know".

Judgment-of-knowing predictions, then, have been shown to be quite accurate after only study trials are used. However, the method used to present items to a subject can affect JK predictions. When memory for a lengthy list of verbal items is tested, items that have been repeated within the list in a distributed manner (DP) are more likely to be remembered than are items repeated in a massed fashion (MP) (Hintzman, 1974). Subjects spend less time studying massed presentations of an item than studying distributed presentations when study is self-paced (Shaughnessy, Zimmerman & Underwood, 1972). Zechmeister and Shaughnesy (1980) found that in a judgment of knowing task, MP items were consistently judged to be more recallable than the DP items, even though recall was actually higher for DP items. Learners were, however, accurately able to predict that twice-presented items would be easier to recall than once-presented items.

A summary of judgment-of-knowing predictions for young adults indicates that their predictions about what information they will and will not be able to recognize and recall after study are related to their actual performance. However, young adult subjects do show some overconfidence in their judgments, predicting that they will recognize or recall more items than they are able to, and erroneously judging that MP items will be more recallable than DP items.

The feeling-of-knowing (FK) task has also been used to study the metacognitive abilities of young adults. As stated earlier, Hart was the first researcher to use the feeling-of-knowing paradigm. In Hart's first experiments (1965), subjects were asked to attempt recall of general information items, and, then, for those items not recalled, to judge whether an answer would be recognized if it was presented among several alternatives. Following these recall and judgment phases, a multiple-choice recognition test was given. The basic test of the accuracy of the FK responses (feeling-of-knowing items that recognition is predicted for) and FK responses (those the subject predicts they probably will not recall) is made by looking only at the test items that subjects predicted they had missed, and in actuality had missed, on the test of recall. If the feeling-of-knowing judgments are accurate indicators of memory storage, the proportion correct recognition for FK items should be significantly greater than the proportion correct for FK items. Hart's results showed that FK predictions are accurate indicators of memory storage.

Memory performance on FK items was correct 67% of the time, whereas performance on $\underline{\text{FK}}$ items proved correct only about 40% of the time, although this also shows that subjects were overconfident in their feeling-of-knowing judgments.

Hart emphasizes that overcautious withholding of correct answers can falsely inflate a subject's memory-monitoring accuracy by producing correct recognition responses that should have been eliminated from the scoring as correct recall responses. Therefore, it is important to encourage subjects to guess. Hart has shown memory monitoring accuracy to occur in college students for general information questions (1965) as well as for paired associate materials (1967) and results are the same whether simple FK or $\underline{\text{FK}}$ dichotomous ratings or 6 pt. rating scales for feeling-of-knowing judgments are used.

Blake (1973) points out that even though subjects in Hart's experiments (1965,1967) recognized more items given FK ratings than $\underline{\text{FK}}$ ratings, subjects showed overconfidence in some of their judgments by failing to recognize about 44% of the items they felt they knew. They also recognized 42% of the items they felt they did not know. He suggested that one of the problems in Hart's procedure is that there were substantial time lapses between attempted recall of a given item, FK judgments, and recognition of items. This could possibly reduce the predictive power of the FK judgments. Blake used trigram stimuli presented so that all three phases, test, recall, and recognition, could be obtained on each item before presentation of the next item. His results indicated that a reduction in the time between item recall, FK judgments and item recognition can increase FK rating accuracy.

Freedman and Landauer (1966) investigated both the TOT and FK phenomena, replicating the results of previous findings. Their subjects were also able to accurately predict which items missed on the recall test would be accurately identified on the recognition test. They also found, similar to the tip-of-the-tongue studies, that providing subjects with the initial letter of the correct answer on the recognition test significantly increased recognition of the answer.

The FK and related metacognitive tasks, then, indicate that young adults are fairly accurate at knowing whether or not they will be able to recognize information that they cannot, in some given time period, recall, but that there is a tendency for people to be overconfident in their judgments.

Metacognitive studies of college-aged students have also included research on the confidence judgments (CJ) that young adults ascribe to the accuracy of their answers to general information questions. Murdock (1966) assessed subject confidence judgments by presenting subjects with lists, each composed of five paired associate words. After the presentation of each list, a probe was given for one of the pairs, and subjects were asked to recall the word associated with the probe and give a rating of how confident they were that their answer was correct. They used a 6-point scale ranging from a point indicating that they were positive that their response was correct to a point indicating that they were positive that their response was incorrect. Murdock found that subjects could assess their performance quite accurately. When they gave the highest rating they were nearly always correct, and when they gave the lowest judgment they were nearly always wrong.

A group of researchers at a Decision Research Center in Oregon has focused many of their research studies on examining how subjects are calibrated. The basic design of their research is to give subjects general information questions, and have them respond to each question by choosing the most likely answer from two alternatives provided. They are then asked to indicate their degree of certainty that the answer they have selected is correct. These studies have found that college students tend to be overconfident in their confidence ratings to general information questions (Lichtenstein & Fischhoff, 1977; Lichtenstein & Fischhoff, 1980) as well as in their responses to more practical information questions, such as which diseases or accidents are more likely to be fatal

(Fischhoff, Slovic, & Lichtenstein, 1977).

Some confidence judgment studies have reported that people who know more are better calibrated. Nickerson and McGoldrick (1963,1965) and Pitz (1974) have reported that people who know more about the materials they are being tested on are more accurate in their confidence ratings, and Maki and Berry (1984) found that high achieving students were better able to accurately predict their future test performance than were lower achieving students. Shaughnessy (1979) has reported a positive relationship between confidence judgment accuracy and test performance.

Lichtenstein et al. (1977) in an in-depth study, examined the relationship between knowledge and accurate calibration. Using general information, two-alternative choice questions, they found that, if percent of items answered correctly is held constant between subjects, there is no evidence that expertise in a particular area leads to better calibration. When subjects were not matched for percent of items answered correctly, subjects who knew more clearly outperformed those who knew nothing. The latter situation tended to lead to high levels of overconfidence, poor calibration and little accurate discrimination between use of numbers on the probability scale. With increasing knowledge came decreasing overconfidence until, for those whose percentage correct exceeded 8Ø%, there was moderate underconfidence. This study indicates the importance of matching subjects for percent of items correct, before assessing their confidence ratings.

It has already been cited that the major finding from the Decision Research Center is that college-aged subjects are overconfident in evaluating the accuracy of their knowledge. Fischhoff et al. (1977) found that subjects were so confident in the confidence judgments they had made that many were willing to stake money on the accuracy of their judgments. The overconfidence of all of these subjects willing to gamble was so great that all would have actually lost money if the gamble had been real.

Koriat, Lichtenstein, and Fischhoff (198Ø) hypothesized that overconfidence is due to relying more on reasons consistent with a chosen answer than on considerations contradicting it. To test this hypothesis, they first had each of their subjects choose the correct alternative for a series of general information questions, and then had them judge the probability that their choice was correct. They used two conditions for this task: one where subjects were not required to give reasons for their choices, and another where subjects were required to specify all possible reasons that they could give for favoring and opposing each choice. They found that the calibration scores for the subjects under the reasons condition were superior to those under the control condition. In a second experiment they found that a group listing only contradictory reasons also improved significantly in calibration. These results strongly suggest that confidence in an answer depends on the reasons a subject can provide to support or contradict the answer.

Several recent studies have addressed the possibility of training realistic confidence. Lichtenstein and Fischhoff (198Ø) have reported that people can improve their confidence accuracy if comprehensive feedback on their performance is provided. They gave subjects feedback on the accuracy of their confidence ratings over multiple training sessions. They found that feedback did lead to improved calibration, but that almost all of the improvement in the quality of subject ratings took place after the first feedback session.

Zechmeister, Rusch, and Markell (1986) also found that confidence judgment accuracy improved for subjects who were provided with feedback, although they found that training had more of an effect on improving the calibration scores of subjects defined as low achievers than on those designated as high achievers. Arkes, Lai, and Hackett (1982) have shown that simply informing subjects that they will have to explain why they have chosen each of their answers to a group of fellow subjects reduces subject overconfidence, even in the absence of training sessions or the requirements to list contradictory reasons.

In general, confidence judgment studies indicate that young adults tend to be overconfident in evaluating the accuracy of their knowledge of general information. However, subjects are better "calibrated" if they are asked to give support for and, more importantly, against their answers before they assign a confidence rating to their response. Training sessions, giving subjects feedback about their confidence ratings, have also been shown to improve the confidence judgment accuracy of young adults.

Older Adults

The metacognitive abilities of older adults have been ignored until recently. Research has shown that metacognitive abilities indicate a developmental trend, with older children and young adults showing more accurate metacognitive skills than younger children (Kreutzer, Leonard, & Flavell, 1975). Research on the metacognitive abilities of older adults explores the idea that the development of metacognitive abilities may extend throughout adulthood, and the possibility that older adults may show metacognitive skills that are different from those of younger adults.

A study by Lachman, Lachman, and Thronesbury (1979) assessed this possibility by examining the metamemory abilities of young, middle-aged, and older adults. They employed feeling-of-knowing tasks dealing both with questions of accuracy (are items predicted to be known actually known) and efficiency (do people spend more time searching for an answer they believe they know than one they don't know). The task was similar to Hart's (1965) study using general information questions. No age group showed better metamemorial accuracy or efficiency than any other. All of the age groups answered more items correctly that they thought they knew and fewer items they thought they did not know, although all groups showed some overconfidence in their judgments. All subjects spent more time responding to items they thought they knew and less time responding to items they thought they did not know. It did appear, however, that relative to the other groups, the oldest group may have suppressed some correct answers and included them in the most confident feeling-of-knowing category. This possibility may make the feeling-of-knowing ratings for the oldest group misleading, since it was previously mentioned by Hart (1965) that cautiousness of responding can lead to inflated estimates of feeling-of-knowing accuracy. The oldest group may not have had feeling-of-knowing ratings that were as accurate as the other age groups if they had not suppressed some of their correct answers.

Perlmutter (1978) has assessed the memory monitoring skills of older and younger subjects' at two education levels: high school and doctoral. She tested subjects' memory prediction judgments, feeling-of-knowing judgments and confidence judgments for words and facts. The word tasks involved having subjects study 24 words under incidental and intentional conditions. They were then asked to predict how many words they felt they would be able to recall, and were then given a recall test. After the test, they were asked to rate, for the words they had recalled, how confident they were, on a 4 point scale, that each word they had recalled was on the originally presented list. They were also asked to predict how many of the words they had not been able to recall they felt they would be able to recognize. For the fact portion of the study, 24 general information fact questions were presented and subjects were instructed to answer the questions, make confidence ratings for as many of the questions as they could recall, and for those they could not recall, predict how many they would be able to recognize. No age differences in accuracy of confidence ratings, or recognition predictions were found, although more education at any age was associated with more accurate memory monitoring skills. Perlmutter suggests that lack of age differences in metacognitive abilities may indicate that these abilities do not contribute to age differences in adult memory.

Murphy, Sanders, Gabriesheski, and Schmitt's (1980) two-part study examined judgment of knowing skills of older adults. The first part of the study examined two JK tasks. All subjects were first asked to estimate their memory span for a series of line drawings, and then each subject was given a span test to assess their estimation accuracy. After this task was completed, subjects were instructed to study each of three different lengths of line drawing lists (with the length of each based on variations of their previous memory span accuracy). They were told to spend as much time as they felt was necessary to accurately recall each of the lists. The results for the estimation task showed that young and older adults were equally accurate in their memory span estimation, although older subjects memory span performance was considerably less than that of the younger adults. The older adults performed more poorly than the younger adults in the recall readiness task in that they chose to study less time in response to increasing task difficulty than did the younger subjects. In a second part of the study it was shown that differences in recall readiness accuracy between the age groups could be eliminated if older subjects were forced to spend at least a set minimal amount of time studying the lists.

Differences between the metacognitive abilities of young and older adults may not always reflect "age" differences. Like Perlmutter (1978), Zivian and Darjes (1983) have suggested that since school provides students with opportunities to practice a variety of mnemonic strategies, and to evaluate their abilities to memorize, it may have a positive effect on judgment-of-knowing tasks. They compared four groups of female subjects on memory performance for a list of 30 words, and on metacognitive strategies used to learn the list. The four groups consisted of young college students, middle-aged college students, and middle-aged and older (over 65) women who had not attended college in the last 5 years. Subjects in school performed better on the memory recall test, and reported using more mnemonic strategies to learn the word list, than did subjects not in school. However, there were no significant differences between the young and middle-aged subjects in school, or the middle-aged and older subjects out of school. Zivian and Darjes concluded that being in school may be a better predictor of metacognitive and memory performance than age differences.

These developmental studies examining metacognitive skills indicate that, although there are developmental trends showing older children to be more accurate in these skills than younger children, accuracy in metacognitive judgments for some tasks can be seen in children as early as kindergarten. Young adults are fairly accurate in assessing their metacognitive skills, although they are often "overconfident" in their assessments. And of the few studies examining metacognitive ability in old age, only one study (Murphy et al. 198Ø) showed older subjects to be less accurate at making these judgments, with the other two studies showing no age decrements between young and older adults. No calibration curves on the metacognitive abilities of older adults have yet been obtained.

Age Differences in Cautiousness and Risk-Taking

Studies on age differences in cautiousness and risk-taking may help lead to predictions about age differences in confidence judgment tasks, since "well-calibrated" people could be seen as being "more cautious" in their confidence judgments than a person who is overconfident in using confidence judgment ratings.

Several studies examining age differences in risk-taking responses have used a "choice-dilemma questionnaire" originally developed by Kogan and Wallach (1961). The questionnaire is made up of a series of everyday life situations. The central person in each situation is forced with a choice between two courses of action, one which is more risky than the other, but also more rewarding if the outcome is successful. The subject must indicate the probability of success that he or she feels would be sufficient to warrant the risky choice. They select from six probability of success alternatives presented after each situation is given: 1, 3, 5, 7 or 9 chances out of 1Ø that the risky alternative will be a success, and an alternative not to choose the risky course no matter what the probability of success. Many studies employing this task to test age differences in risk-taking have found older adults to respond cautiously, choosing higher probabilities of success than younger adults before they feel it would be desirable for the person described in the dilemma to take the more risky course of action (Botwinick, 1966,1969; Kagan & Wallach, 1961; Vroom & Pahl, 1971). Botwinick (1969) found that the main reason for this age difference in responding was the fact that elderly subjects were more likely to choose the "no-choice of risky alternative no matter what the probability of success" option than were younger adults. When this option was unavailable, Botwinick (1969) found that elderly and younger subjects were similar in their risk-taking responses.

More recent studies have examined risk-taking in the elderly in terms of task performance under different conditions of reinforcement. Reinforcement has either been studied in terms of instructional set, where subjects receive instructions reinforcing or discouraging risk-taking responses (Okun & Di Vesta, 1976), or monetary incentive, where subjects are reinforced with money for risk-taking behavior (Birkhill & Schaie, 1975; Okun & Cherin, 1977; Robins, 1969; and Winefeld & Mullins, 1980).

All of these studies on the relationship of reinforcement to risk-taking support Botwinick's (1969) finding that when a risk-taking option is not available, or in the case of reinforcement, made somehow undesirable, older subjects and younger subjects will show similar risk-taking responses. The elderly will, however, tend to choose to take fewer risks, or not to respond at all, when given the option. This often results in increased omission errors in the elderly, and more cautious responding.

These cautiousness studies indicate that older subjects are more likely to choose the most extreme and cautious response, that of taking no risks, than are younger subjects, when that response is available. However, when older subjects are forced to use a scale without a "no risk" option, they are similar to younger subjects in their risk-taking responses. This implies that the type of scale used to make risk-taking ratings may affect older subject responses more than it affects the responses of younger subjects.

RATIONALE FOR THE PRESENT STUDY

The present experiment will examine the effect of three variables on confidence ratings. The first variable is age. No confidence rating studies have examined calibration differences between young and older adults. Most of the studies looking at age differences in metacognitive tasks have not found older subjects to be significantly different from young subjects in their metacognitive skills. However, Lachman et al. (1979) indicated that the older adults in their study may have been more cautious than the younger subjects in the recall phase of the experiment, being less likely than the younger subjects to say that they recalled an answer that they were not positive about. This could lead to an "inflated" feeling-of-knowing performance for older adults, since they may have been more likely than younger adults to make a "feeling-of-knowing" rating on items that they could actually recall. Older subjects in the Lachman et al. (1979) study did show feeling-of-knowing ratings that were similar to those of the younger subjects; however, if their performance was artificially "inflated" due to cautious responding, they may actually be less accurate than young adults in their feeling-of-knowing assessments. Murphy et al. (198Ø) suggested that older subjects may have more problems than younger subjects in making judgment-of-knowing assessments. Thus, older subjects may have more problems than younger subjects in making confidence ratings, perhaps showing "poorer calibration".

However, the Lachman et al. (1979) study indicates that older subjects may show more "cautiousness" in responding than younger subjects, and research on cautiousness indicate that older subjects will respond more cautiously (Botwinick, 1966), and will show more omission errors than young subjects, in some "risk-taking" tasks, when given the chance. This might lead to a prediction that older subjects will be less overconfident than younger subjects, perhaps showing "better calibration".

In the present study, subjects did not have the choice of "not responding". All subjects were required to respond to a series of general information questions, and to rate their confidence that their response is correct. However, all subjects were able to choose a "cautious" rating that indicated that they had no idea whether or not their response was correct.

The second variable to be examined is monetary incentive. Past studies have all employed number scales for subjects to use to make their confidence judgment ratings. The present study will provide half of the subjects with a "money incentive" scale, where they will "bet money" to make their ratings, and half of the subjects will be provided with the usual "number" scale (no money incentive) to make their ratings. The prediction is that subjects may be more accurate in their confidence ratings if they think they can win money by being accurate, than if they have no money incentive to be accurate.

The third variable studied is the effect of feedback on confidence ratings. Past studies have indicated that training subjects by providing them with information about their confidence judgment performances can help them to improve their confidence judgment accuracy on future tests (Lichtenstein et al., 198Ø). The present study asked subjects to go through the same set of general information questions twice, assigning confidence ratings to their responses both times. In between trials, subjects were given brief feedback that they may have been overconfident on their ratings during the first trial, and encouraged to try to make accurate ratings on the second trial. The hypothesis is that subjects will be less overconfident in their ratings on Trial 2 than they were on Trial 1, even though the feedback was be brief, and not directed to any specific Trial 1 responses.

METHOD

Subjects

A total of 34 college students, ($\bar{X}$ age = 19.38 yrs.) from Loyola University in Chicago and the College of St. Thomas in St. Paul volunteered to participate in the study. These students received credit towards their grade in Introduction to Psychology classes in exchange for their participation. A total of 28 older adults, ($\bar{X}$ age= 71.61 yrs.),from the Roger's Park area in Chicago were also recruited, with the majority located through senior citizen centers. All older subjects were offered $1 to participate in the study.

One half of the subjects at each age level were randomly assigned to the "Questions-only Condition"(No Money Reinforcement). All subjects who had been assigned to this condition were asked to give true/false responses to each of 100 general information statements, and to rate each of these responses in terms of their confidence that it was correct. The other half of the subjects were randomly assigned to the "Monetary Incentive condition"(Money Reinforcement). Subjects assigned to this condition were asked to give true/false responses to each of 100 general information statements; however they were asked to predict how certain they were that each response was correct by stating the number of pennies they would be willing to bet that their response was correct, from Ø to 5 cents. It was explained to them that they could win money if they were accurate in rating the confidence of their responses in this way (see Appendix C).

Materials

The subjects were given a packet of 1ØØ 4x6 in. index cards, each card containing one general information statement (see Appendix A). They were also given an answer sheet for their responses and ratings (see Appendix B). Most of the general information questions used were taken from a study by Nelson and Narens (198Ø) giving norms for 3ØØ general-information questions. Pilot studies were done to design the present series of questions so that, on the average, subjects from both age groups would be able to answer approximately half of the questions accurately (corrected for chance guessing).

Procedure

At the beginning of each session, subjects in both groups were given Form 1 of the Quick Vocabulary test. The Quick Test is a brief individual intelligence test based on perceptual-verbal performance. Form 1 consists of four line drawings, and subjects are asked to point to which of the line drawings best represents each of the 5Ø word-items. Three to ten minutes are required to administer the

QT to any person who can see the drawings and hear or read the word items. Scores on the test were calculated immediately after it was delivered, and subjects were required to score 43 or above (approximately equivalent to 100 on a standardized I.Q. test) in order to receive the 100 general information question cards. The young subjects' $\bar{X}$ QT score was 44.74; for older subjects the $\bar{X}$ QT score= 46.75. The subjects were told that they were being given a vocabulary test. If they asked for further information, they were told that it was to help the experimenter decide on which packet of questions to use. All subjects were told that they had done well on the vocabulary test, and then the rest of the procedure was explained to them.

All subjects were told that this was a two-part study, each part taking approximately 30 min.. The subjects were then given instructions relaying information about the condition to which they had been assigned. (See Appendix C for specific instructions). All subjects were told to respond true or false to each of the 100 statements on the answer sheet provided, and to rate their confidence that each of their responses was correct on the six-point scale provided for each response. Subjects were told to take as much time as they needed to make their responses and ratings.

The experimenter stayed in the room with all the subjects in order to answer any questions they might have. The experimenter read the questions to any of the older subjects who requested this assistance. Approximately 1/4 of the subjects requested this assistance. After the subjects had responded to and rated the responses to all 1ØØ questions, the subjects were given a 5 to 1Ø minute break. During this time the experimenter corrected each subjects' answer sheet, and computed the number of points or pennies a subject had given as a confidence rating to each incorrect decision they had made.

During the second part of the study, the subjects were asked to do the same task again, using the same cards in the same order. They were then (a) given feedback about how many points or pennies they had placed on their incorrect decision while going through the task the first time, (b) urged to approach the questions as though they were answering them for the first time, and (c) told to try to be as accurate as they could be in their confidence ratings (See Appendix C for detailed instructions for both conditions).

RESULTS

Many measures have been used in metacognitive studies to evaluate confidence judgment performance. Assessments of over-confidence are most often evaluated by using calibration curves. The measures reported here include calibration curves, along with numerical assessments of over/underconfidence, calibration and resolution introduced by Lichtenstein and Fischhoff (1977), and confidence accuracy quotients, introduced by Zimmerman, Broder, Shaughnessy and Underwood (1977).

Calibration curves provide a graphic evaluation of how well calibrated subjects are. A subject, or group of subjects, are perfectly calibrated if, for all responses assigned the same probability correct, the proportion correct is equal to the probability assigned. Therefore, responses to which a perfectly calibrated subject assigns a probability of being correct 8Ø% of the time will be correct 8Ø% of the time. A graph showing the hit rate (percentage correct) for each probability rating given is called a calibration curve.

A perfectly calibrated subject would have a "calibration curve" that lay completely on the diagonal; meaning that responses they assigned .5Ø probability of being correct would be correct 5Ø% of the time, the responses they assigned a .60 probability of being correct would be correct 60% of the time, and so on. An underconfident subject would have a calibration curve that lay above the diagonal. That would mean, for example, that responses assigned a .50 probability correct may be correct 60% of the time, and those assigned a .60 probability correct may be correct 65% of the time, and so on up the probability scale. The most common finding in confidence judgment research is that subjects tend to be overconfident (Lichtenstein, et al., 1982). Overconfident subjects show calibration curves that lay below the diagonal. For example, when overconfident subjects gave responses a .50 probability rating, they may only be correct 45% of the time, and for a .60 rating be correct only 50% of the time, etc.

An equation measuring the adequacy of calibration was proposed by Murphy (1973): Calibration = $1/N\sum_{t=1}^{T} n_t(r_t - c_t)^2$.*

*N=total number of responses
n_t=number of times r_t was used
r_t=probability rating
c_t=probability correct for all items assigned r_t
T=total number of response categories used

A perfectly calibrated subject would score Ø on this measure. The worst possible score, 1.Ø, could be obtained only by those who always give the highest probability rating (absolutely sure) when wrong, and always give the lowest rating (total guess) when right.

Murphy (1973) also proposed an equation to measure resolution. Resolution measures the ability of the responder to discriminate different degrees of subjective uncertainty by sorting the responses into categories whose respective ratings of percentage correct are maximally different from the overall percentage correct. A flat (horizontal) calibration curve shows no resolution; a steeper curve shows good resolution. The higher the resolution score, the better the subjects resolution ability.

The over/underconfidence measure is a simpler and more commonly used measure of confidence judgment accuracy than the calibration and resolution measures. The equation given by Lichtenstein et al. (1977) is:

Over/underconfidence = $1/N\sum_{t=1}^{T}n_t(r_t - c_t)$.

A rearrangement of the terms in this equation shows that over/underconfidence is equal to the differences between the mean of the probability responses and the overall proportion correct. Overconfidence is shown by a positive score, underconfidence by a negative score.

The final measure used to analyze confidence judgment accuracy is the confidence accuracy quotient (CAQ). The CAQ is a ratio, the numerator of which is the difference between the mean confidence assigned to right items and mean confidence assigned to wrong items, and a denominator that is the square root of the pooled variance of the subject's confidence judgments for right and wrong answers. The CAQ is similar to _d'_ in a signal detection analysis, and equals zero when a subject cannot discriminate right and wrong answers. The CAQ is affected by guessing in a forced-choice procedure, like the two-alternative choice situation (true or false) used in the present study (Shaughnessy, 1979). In this type of task, a certain proportion of responses given a very low confidence rating (for example .5Ø) will be correct by chance. Confidence values assigned to these responses will tend to lower the mean confidence of right answers, lessening the difference between mean confidence for right and wrong answers. Therefore, although the CAQ is still an accurate measure of confidence accuracy in a two-alternative choice situation, the CAQ scores in designs like the present study are likely to be lower than for studies using an increased number of alternatives (for example, four-item multiple choice questions) from which to choose responses and give confidence ratings.

To be included in the following analyses, subjects had to meet several criteria. First, any subjects who answered less than a chance level (50% of the items) correct were not included in the final analyses, and only subjects who participated in both Trials 1 and 2 of the experiment were included. The difference in the number of subjects used in each comparison group is the result of careful subject matching for proportion of correct responses. Lichtenstein et al., (1982) report that calibration differences between groups of subjects can be affected by differences in proportion of items subjects have responded to correctly. The proportion of responses correct between groups must be controlled. The maximum number of young subjects in the young subject comparison groups that resulted in the best proportion correct "match" was 17 per group. The maximum number of subjects in the young versus older subjects comparison groups was 14 per group, and the number of older subjects being compared to other older subjects that resulted in the best match was 12 per group. (All the young subjects are taken from the same pool of 34 subjects that met all the criteria for the experiment; the older subjects are taken from the same pool of 28 subjects that met the same criteria.)

No significant differences were found in proportion correct between matched subjects in any of the comparison groups, since the subjects had been matched so that proportion correct would be as similar as possible between comparison groups. The comparison groups to be discussed include: comparisons between young and older subjects (Tables 1-4) in both the money and no money conditions for Trials 1 and 2; comparisons between young subjects in the money and no money condition, and within each condition, between Trials 1 and 2 (Tables 5-8); and comparisons between older subjects in the money and no money condition, and within each condition, between Trials 1 and 2 (Tables 9-12). (Appendix D includes calibration curves for all the subjects before matching took place).

Item analysis based on response performance revealed that item difficulty distributions (ranging from number of items correct by 100% of the subjects to number of items correct by 0% of the subjects) for lists used by young and older subjects in both money and no money conditions (n=14 in each condition) on Trial 1 (Table 13) and Trial 2 (Table 14) were not significantly different. (Chi-square analysis was used, and cells with errors of 9 or above were grouped together to enable large enough frequencies for chi-square analysis). The item difficulty distribution pattern was similar between the young and older subjects. For example, in Table 13, 33% of the items on Trial 1 were answered incorrectly by 2 or fewer older subjects in the money group, and similarly 32% of the items were answered incorrectly by 2 or fewer of the younger subjects in the money group. Within the same groups, 1Ø% of the items were answered incorrectly by 11-14 of the older subjects, and 12% of the items were answered incorrectly by 11-14% of the younger subjects.

As previously discussed, the use of calibration curves is the most common way to display confidence rating results. Calibration curves are shown in Figures 1-12. In order to construct the figures, the Ø-5 point rating scale that subjects used for their confidence judgments was first converted to a .5Ø to 1.ØØ probability scale. This converted scale was also used to carry out the analyses of the dependent variables (over/underconfidence, calibration, resolution, etc.) summarized in Tables 1-12. Significant differences in these variables will be noted in Tables 1-12, and cited in the text.

Comparisons between young and older subjects

The calibration curve comparing young (*n*=14) and older subjects (*n*=14) in the money condition of Trial 1 is shown in Figure 1. Both groups showed overconfidence at each confidence rating level, except for the .5Ø level. The calibration curve for older subjects was closer to the "perfect calibration" diagonal line at the lower rating levels (.5Ø-.7Ø) than the curve for younger subjects, but the younger subjects were better calibrated than the older subjects at the .9Ø and 1.ØØ rating levels. Table 1 shows that there were no significant differences between the two groups in any of the confidence rating measures analyzed.

Figure 2 displays the calibration curves for younger and older subjects on Trial 1 of the no money reinforcement condition. Although both groups again show overconfidence in scale use, the older subjects are closer to the diagonal calibration line than the younger subjects at the .60 and .70 levels. The only significant difference found in the analyses between these groups shown in Table 2 was in the resolution measure, where older subjects showed poorer resolution, with a mean resolution score of .024, than the younger subjects with .034, t (26)= 2.13, $p < .05$. The calibration curve for older subjects in Figure 2 is "flatter" than the curve for younger subjects, reflecting the resolution score difference between groups.

The calibration curves for young and older subjects in the second trial of the money reinforcement condition is seen in Figure 3. The curve for younger subjects is closer to the perfect calibration line for the .60 and .70 ratings, and also for the .90 and 1.00 ratings. Significant differences in over/underconfidence and calibration were found between the young and older groups (Table 3). Mean over/under confidence was .102 for young subjects and .174 for older subjects, t(26)= -2.60, $p < .05$. The mean calibration score was .029 for young subjects and .059 for older subjects t(26)= -3.04, $p < .05$. Younger subjects had significantly higher resolution scores and lower confidence scores than older subjects; young subjects average resolution score was .Ø37, and older subjects score was .Ø18, *t*(26)=3.51, *p* $<$.Ø5. The average confidence score for young subjects was .751, and for older subjects .823, *t*(26)= -3.Ø9, *p* $<$.Ø5. On the second trial of the money condition, then, younger subjects were better calibrated, especially at the higher confidence rating levels. Older subjects were better calibrated at the lower levels, but had flatter calibration curves, again reflecting their poorer resolution abilities.

Figure 4 contains the calibration curves for young and older subjects on Trial 2 of the no money reinforcement condition. Both groups of subjects were again overconfident in their confidence ratings, although the older subjects had a flatter line, being closer to the perfect calibration line than the younger subjects at the .6Ø and .8Ø levels. As shown in Table 4, young and older subjects were significantly different in only the CAQ and resolution measures. Young subjects were better than the older subjects at discriminating correct from incorrect answers. The mean CAQ rating was .8Ø1 for younger subjects and .488 for older subjects, *t*(26)=2.73, *p* $<$.Ø5; young subjects showed better resolution, mean score .Ø46, than older subjects, with mean score .Ø23, *t*(26)= 5.Ø8, *p* $<$.Ø5.

Therefore, in comparisons between young and older subjects, in both money and no money conditions, on Trials

1 and 2, significant differences in over/under confidence and calibration were found only on the second trial of the money condition (Table 3), where older subjects were significantly more overconfident than younger subjects. This performance difference on Trial 2 suggests that young subjects seemed to be affected by simple feedback about their overconfidence on Trial 1 more than the older subjects were. Younger subjects showed better resolution than older subjects in each comparison except for that of the first trial of the money condition. This resulted in the "flatter" calibration curves seen for the older subjects, since they appeared less overconfident than the younger subjects at the lower end of the rating scale, and more overconfident at the higher end of the scale.

Young Subject Comparisons

Calibration curves for young subject comparisons between Trials 1 and 2 of the money (*n*=17) and no money (n=17) condition are shown in Figures 5-8. Figure 5 shows that the calibration curves for young subjects between the first trials of the money and no money condition were very similar. No significant differences were found between any of the calibration measures listed in Table 5, suggesting that young subject's made similar confidence rating judgments, whether they made confidence ratings in terms of a money scale or a number scale.

Calibration curves for the young subjects of the second trial for those in the money and no money conditions are displayed in Figure 6. Again, the calibration curves for both groups are similar, and no significant differences were found between the calibration measures listed in Table 6.

Figure 7 presents the calibration curves for young subjects in the money condition on Trials 1 and 2. Calibration improved at the .70 confidence rating level on Trial 2, with some improvement also shown on Trial 2 at the .90 and 1.00 level as well. Significant differences between these trials were found in the mean confidence, over/under confidence and calibration measures with subjects showing less overconfidence in their confidence rating judgments on Trial 2. On Trial 1, subjects showed a higher mean confidence level of .780 as compared with .753 on Trial 2, $\underline{t}(16)=3.59$, $\underline{p} < .05$. On Trial 1, subjects showed a mean over/underconfidence score of .117 and on Trial 2 of .085, $\underline{t}(16)= 3.37$, $\underline{p} < .05$, and on the first trial, subjects in the money condition had an average calibration score of .035, and on the second trial a mean calibration score of .025, $\underline{t}(16)=3.42$, $\underline{p} < .05$. Therefore, it appears that young subjects in the money condition were affected by feedback about their Trial 1 performance, making more cautious ratings during Trial 2 than they had during Trial 1.

Calibration curves on Trials 1 and 2 of young subjects in the no money condition are shown in Figure 8. The curves are similar, with the exception of the .70 confidence rating. Subject ratings on Trial 2 were closer to the perfect calibration line than the ratings on Trial 1. Significant differences were found between Trials 1 and 2 for mean confidence ratings, .797 on Trial 1 and .772 on Trial 2, $\underline{t}(16)= 4.16$, $\underline{p} < .05$, and for resolution, with a mean resolution of .033 on Trial 1 and .044 on Trial 2, $\underline{t}(16) = -3.68$, $\underline{p} < .05$. However, no significant differences were found between trials in calibration or over/underconfidence measures. Feedback appeared to have some affect on young subjects in the no money condition, although it did not affect their confidence judgments as much as it affected the young subjects in the money condition.

For young subjects then, comparisons between young subjects in the money and no money conditions displayed significant differences in over/underconfidence and calibration measures only between Trials 1 and 2 of the money condition, where subjects were better calibrated on the second trial, after feedback had been given.

Older Subject Comparisons

Calibration curves are presented in Figure 9 for older subjects on Trial 1 between the money and no money condition. Both groups showed overconfidence, with the curves overlapping so that the no money subjects were better calibrated at the .60,.70 and .90 rating levels, and the money subjects better calibrated at the other rating levels. No significant differences were found between the money and no money conditions for the confidence rating measures shown in Table 9, indicating that on Trial 1, older subjects, like younger subjects, showed few differences between confidence ratings made using a money scale versus those made using a number scale.

Figure 10 shows calibration curves for older subjects in the money and no money conditions of Trial 2. Subjects in the no money condition were closer to the perfect calibration line for the .50 and .70 confidence rating levels than were subjects in the money condition, although the calibration lines are similar for .80-1.00 confidence ratings. Again, no significant differences were found between the money and no money conditions for the confidence rating measures shown in Table 10.

Figure 11 presents comparisons of calibration curves for older subjects in the money condition between Trials 1 and 2. The curves here overlap, with subjects on Trial 1 being closer to the diagonal line for confidence ratings .60 and .70, and subjects on Trial 2 having points on the curve closer to the diagonal line for the .80 and .90 levels. No significant differences between trials were found for the measures listed in Table 11. The confidnence ratings of older subjects, then, were not effected as much as the ratings of younger subjects after feedback about their Trial 1 performance in the money condition.

Figure 12 displays calibration curves for older subjects in the no money condition between Trials 1 and 2. Subjects in Trial 1 have a calibration curve closer to the perfect calibration line at the .7Ø, .9Ø and 1.ØØ levels. Significant differences between Trials 1 and 2 were found for mean confidence and over/under confidence measures. Mean confidence on Trial 1 was .8Ø4, and Trial 2 .825, $\underline{t}(11)=-2.85$, $\underline{p} < .Ø5$, and mean over/under confidence was .144 on Trial 1 and .167. In the no money condition, older subjects displayed <u>more</u> overconfidence in their confidence ratings after they had received feedback about their Trial 1 performance. This is a pattern that is nearly opposite to that shown by younger subjects, who were less overconfident in the no money condition.

For older subjects then, significant differences in confidence and over/under confidence were found only between Trials 1 and 2 of the no money condition. These subjects appeared to be more overconfident in their confidence judgments on Trial 2 than they were on Trial 1.

Confidence Rating Scale Comparisons

Young and older subjects did use the confidence scale differently. Table 15 displays the total distribution scale for young and older subjects in both money and no money conditions (n=14 each), on Trial 1 , and Table 16 on Trial 2. Chi-square analysis of both tables yielded significant results. On Trial 1, $\underline{\chi}^2(15) = 278.45$, $\underline{p} < .05$, and for Trial 2, $\underline{\chi}^2(15) = 444.06$, $\underline{p} < .05$. On Trial 1, the largest differences in scale use are seen between the young and older subjects in the money condition. Younger subjects were more likely to choose the lower ratings (0,1 and 2) with 45.71% of their ratings given to these ratings as compared to 34.57% of ratings at these lower levels for the older subjects in the money condition. These older subjects were also more likely to choose the highest rating of 5 (41.86% of their ratings) as compared to 28.29% of the younger subjects ratings given to the highest scale level.

On Trial 2, differences in scale use are seen between younger and older subjects in both the money and no money conditions. For example, older subjects in the money condition used the lower scale levels (0-2) 33.64% of the time, and in the no money condition, 28.86%, compared to younger subjects who used the lower scale levels more frequently, 50.86% for those in the money condition, and 41% for those in the no money condition.

The frequency of answer and rating changes between Trials 1 and 2 for all conditions, and the patterns of those changes are shown in Tables 17 and 18. Table 17 shows the average number of answers and ratings that were changed from incorrect to correct, or from correct to incorrect. Descriptive analysis of the pattern of the number of answers and ratings was similar between the four groups, with a mean of 14.6 answers changed between Trials 1 and 2, approximately half of these changes (7.33) from correct to incorrect, and the other half (7.29) from incorrect to correct. Also, approximately half the answers were changed without an accompanying rating change. Of the average total ratings changed (41.34), 82% were changed without an accompanying answer change. For all groups then, subjects were unlikely to change their answers between Trials 1 and 2, but they did change a substantial number of their ratings.

The direction of the rating changes is shown in Table 18. Overall, 58.66% of the ratings given to responses on Trial 1 were not changed by subjects on Trial 2. For young subjects in the money condition 56.5% of their ratings remained the same between Trials 1 and 2; for young subjects in the no money condition, 60.2% were not changed. For older subjects in the money condition, 59.4% of the ratings were not changed from Trial 1 to 2, and 58.7% were not changed for older subjects in the no money condition.

Of the rating changes made, young subjects in both conditions were more likely to change from higher ratings to lower ratings, showing less confidence in their answers on Trial 2 than on Trial 1. For young subjects in the money condition, 68% of their rating changes were from higher ratings to lower ratings, and similarly, 69% of the changes for young subjects in the no money condition were from higher ratings on Trial 1 to lower ratings on Trial 2. This suggests that rating changes were not made randomly, and may have been a reaction to the feedback that subjects received between Trials 1 and 2.

For older subjects in the money condition, almost as many of their rating changes were from high to low (45.20%) as from low to high (54.8%) between Trials 1 and 2. Older subjects in the no money condition showed a rating change pattern that was different from the younger subjects pattern, with 61.2% of their rating changes going from lower ratings on Trial 1 to higher ratings on Trial 2. Older subjects in the money condition seemed to display a random pattern of rating changes, whereas the pattern for subjects in the no money condition indicated a tendency to give higher ratings to answers on Trial 2 than on Trial 1.

In summary, young and older subjects showed significant calibration differences only on Trial 2 of the money condition, where younger subjects were better calibrated than older subjects. Also, younger subjects showed an overall better resolution ability than older subjects.

Comparisons for younger subjects in the money and no money conditions indicated that feedback did have an effect on these subjects. Subjects in the money condition showed better calibration on Trial 2, after feedback on their Trial 1 performance, than they had shown on Trial 1. Young subjects in the no money condition had lower mean confidence on Trial 2 than on Trial 1, also indicating lower confidence judgments after feedback.

Comparisons for older subjects found significant differences only between Trial 1 and 2 of the no money condition, where subjects appeared more overconfident on Trial 2, after feedback on their performance, than on Trial 1.

DISCUSSION

The basic goal of the present study was to examine the effects of three variables on subject confidence judgment accuracy. These three variables were: feedback given to subjects about their performance; monetary incentive; and most importantly, the age of the subjects.

The effects of feedback on confidence judgments will be discussed first. An overview of the results indicated that feedback seemed to have more of an effect on the confidence judgments of younger subjects than on older subjects. The next effect to be discussed will be the effect of monetary incentive on confidence judgments. The results suggested that, although there seemed to be no dramatic changes between the confidence judgments of subjects using "money bets" to make their confidence ratings, and those using a simple number scale, the young "money incentive" subjects, at least, seemed to show less overconfidence between Trials 1 and 2 than the subjects in the "no money incentive" condition. Finally, overall confidence judgment differences between young and older subjects will be discussed. In general, older subjects seemed to use the confidence rating scale differently than young subjects. They showed poorer "resolution", or ability to sort their ratings into different scale levels.

This resulted in "flatter" calibration curves for the older subjects than for the younger subjects, since they were less overconfident than younger subjects at the lower end of the confidence rating scale (.5Ø and .6Ø ratings) and more overconfident at the higher end of the scale (1.ØØ ratings).

Feedback

Several previous studies have indicated that subjects can be "trained" to produce more realistic confidence ratings by giving them "feedback" about their performance. Lichtenstein et al. (198Ø) gave subjects comprehensive feedback, over multiple training sessions, about their confidence judgments. Subjects did show improved ratings, with most of the improvement occuring after the first training session. Zechmeister et al. (1986) have also shown that one training session can help subjects, especially low achieving subjects, improve their calibration scores, and Arkes et al. (1982) found that even in the absence of training, simply informing subjects that they will have to explain their reasons for their answer choices, to other subjects, helps to improve subject calibration.

In the present study, the feedback was brief and intentionally "vague". Subjects were simply provided with a number indicating, in the no money condition, the number of "points" that corresponded to their ratings of incorrect answers. For the money condition, the feedback given was the number of pennies that the subjects would have lost during Trial 1 on "bet ratings" given to incorrect answers. All subjects were then told that this feedback meant that they had been overconfident in some of their ratings, giving high confidence judgments to some of their incorrect answers. They were told to be as accurate as they could be on the second trial. "Money condition" subjects were told that they could win back some of the money they had lost if they were more accurate in their ratings on the second trial.

Young subjects did seem to be affected by this feedback, showing less overconfidence in the confidence judgments on Trial 2, after feedback, than they had shown in their Trial 1 judgments. This is reflected in the lower confidence, over/under confidence, and calibration scores shown on Trial 2 for young subjects in the money condition, and lower confidence ratings on Trial 2 for young subjects in the no money condition. Young subjects in both groups changed approximately 40% of their ratings between Trials 1 and 2, with almost 70% of these changes indicating less overconfidence on Trial 2, i.e., with lower ratings on Trial 2 than had been chosen on Trial 1. It seems, then, that for college aged subjects, even brief feedback about "overconfidence" can affect subject confidence judgment ratings.

The effect of feedback on older subjects is not clear. There were no statistically significant differences between the Trial 1 and Trial 2 performances for older subjects in the money condition. These older subjects, like the younger subjects, did change about 40% of their ratings between Trial 1 and Trial 2. However, these changes seemed almost "random", with approximately half of the changes going from lower ratings on Trial 1 to higher ratings on Trial 2, and the other half of the changes going from high ratings on Trial 1 to lower ratings on Trial 2.

Older subjects did seem to show more "fatigue" on Trial 2 than younger subjects, and may have been exhibiting less concentration on their Trial 2 ratings than they exhibited on their Trial 1 ratings. But even if this were true, it does not adequately explain the rating changes seen between Trial 1 and 2 for the older subjects in the no money condition.

In the no money condition, older subjects appeared to be more overconfident after feedback than they were before they received feedback. This result was opposite of what had been expected. Again, like older and younger subjects in the other conditions, these older subjects changed approximately 40% of their ratings. Of these rating changes, over 60% were from lower ratings on Trial 1 to higher ratings on Trial 2. Although this pattern of rating changes is not as dramatic as the rating changes shown for young subject (from high ratings on Trial 1 to lower ratings on Trial 2), it was reflected in significantly higher confidence and over/under confidence measures on Trial 2 than on Trial 1. One explanation for this pattern may be the greater "familiarity" that subjects had with the material on Trial 2 than they had had with it on Trial 1. These older subjects may have reasoned "I've heard that answer before, therefore I'm confident it is correct". For example, Hasher, Goldstein and Toppino (1977) found that repeated general information statements were more likely to be judged as "true" than similar, non-repeated statements.

Both young and older subjects were likely to make the same answer responses on Trial 2 that they had made on Trial 1. For example, these older subjects in the no monay condition changed only 13% of their answers between Trial 1 and 2, and so they were "familiar" with these answers on Trial 2. In the absence of any "monetary" incentive to temper their responses, the older subjects in the no money condition may have been more likely to choose their confidence ratings based on "familiarity" with the material.

Interaction Between Feedback and Monetary Incentive

Fischhoff et al. (1977) found that subjects were "overconfidently" willing to stake money on confidence judgments that they had already made. "Betting" money on their judgments did not cause subjects to respond cautiously, since most subjects in the study would have lost money if these bets had been real. The present study examined whether having subjects make money bets as confidence ratings would make subjects more cautious in their responses.

Monetary incentive did not have a significant effect on subject confidence judgments on Trial 1, since the confidence judgment measures between the money and no money conditions for the Trial 1 ratings were not significantly different for the younger or older subject groups. Although there were no significant differences between the money and no money conditions on Trial 1 for either age group, young subjects in the money condition did use the scale differently than young subjects in the no money condition. In the money condition, young subjects were less likely to use the highest rating level of 1.ØØ (used only 28.3% of the time) than were subjects in the no money condition, who used the highest rating level 37.5% of the time on Trial 1. This indicates the possibility that money could have influenced the young subjects in the money condition by causing them to be less likely to want to "bet" the highest amount of money on the accuracy of their responses than young subjects in the no money condition. It is possible that the low amount of money at stake (penny bets) made it less likely that there would be "significant" differences between the two groups than if higher money stakes had been used.

Younger subjects did appear to be less overconfident in both the money and no money conditions after they had received feedback. This effect was most apparent in the money condition, where subjects showed lower over/underconfidence and calibration scores, as well as lower confidence scores, on Trial 2. Since using a "money" scale did not cause subjects to make different confidence judgments than those made when using a "point scale" on Trial 1, it is likely that the use of this "low wager" money scale cannot completely account for the differences between Trials 1 and 2 for the money and no money conditions.

It is true that subjects in the money groups were given a monetary incentive to work toward on Trial 2 that was not given to the no money groups, since the money groups were told that they could win back some of the money they had lost by givng more accurate ratings on the second trial. But the feedback given to subjects in the money groups between Trials 1 and 2 was also more "concrete" than the feedback given to subjects in the no money group. Subjects in the money group were given feedback about the amount of pennies (a concrete example) that they had lost by giving overconfident ratings to incorrect answers, whereas the no money group was given feedback about the amount of points (a more abstract concept) that they had given in ratings to incorrect answers.

Both groups of young subjects, therefore, were more likely to make confidence rating changes that resulted in less overconfidence on Trial 2. It has also been shown that monetary incentive alone did not have a significant effect on the young subject's ratings on Trial 1. It may be, then, that the greater difference between Trials 1 and 2 of the money condition as compared to the difference in calibration of Trials 1 and 2 of the no money condition can be accounted for more by the "concreteness" of the feedback given, than by the monetary incentive indicated.

Confidence judgment results were also different between Trial 1 and Trial 2 for older subjects in both the money and no money conditions. As stated earlier, older subjects in the money condition appeared to make random rating changes between Trials 1 and 2, with no significant confidence judgment differences between trials. Older subjects in the no money condition showed more overconfidence through their ratings on Trial 2. The brief feedback given to subjects about their ratings on Trial 1 did not seem to have much of an effect on the Trial 2 confidence ratings of the older subjects.

It may be that there was a tendency for both groups of older subjects to become more overconfident on Trial 2, because of the greater "familiarity" of the material. As already suggested, it is also possible that there is a "fatigue" factor involved, with older subjects being more tired of the task on Trial 2. This may have caused them to pay less attention to their Trial 2 ratings.

This possible tendency to be overconfident on Trial 2 may have been "tempered" somewhat by the more "concrete" feedback given to the older subjects in the money group. This hypothesis could help to explain the "seemingly random" rating changes shown by the older subjects in the money group. They may have been likely to give higher ratings to answers that now, on Trial 2 seemed more familiar, but also to keep the feedback about their overconfidence in mind, which could have resulted in less overconfidence in those answers that still seemed unfamiliar.

Age Differences

Other studies examining age differences in confidence ratings have indicated no significant differences between young and older adults (Perlmutter,1978;Lachman et al,1979). It is interesting to note, therefore, the different confidence rating patterns seen between young and older subjects in this experiment.

Most of the differences between age groups, as already mentioned, seemed to occur during Trial 2 of the study.

For example, in the "money condition" comparisons, young subjects were significatly less overconfident than older subjects, on Trial 2. The rating scale was used differently by young and older subjects of both groups for both trials, but the dramatic differences were seen on Trial 2. For example, 43.4% of the older subject's ratings on Trial 2 of the money condition were made at the highest 1.ØØ level, and 43.7% of the older no money subjects ratings. This contrasts with only 27% of the young money subjects ratings and 36.7% of the money subject's ratings given to the highest 1.ØØ rating level.

Although the highest number of ratings for both age groups occurred at the 1.ØØ level, older subjects were more likely to choose this rating on both trials than were the younger subjects. Botwinick (1969) indicated that the older subjects in his experiment were more likely to choose the most extreme response (in the case of his experiment, the most cautious response) than were younger subjects. When this extreme response was not available to them, they showed the same pattern of responses as younger subjects. It may be that older subjects, who may have had less recent experience with test taking than younger subjects, have a more difficult time in simply using the rating scale than younger subjects. The older subjects did tend to have "flatter" calibration curves, reflecting problems in sorting their ratings into different levels of uncertainty that could best reflect the actual confidence they had in their responses. As Perlmutter et al.(1978) and Zivian et al. (1983) pointed out, being "in or out of school" may be more of a factor in explaining differences between groups in metacognitive skills than are age differences.

Implications of the Present Study

Cavanaugh and Perlmutter (1982), in a critical examination of metacognitive research, concluded that the value in this research has been its demonstration of metacognitive ability differences between different groups of subjects, and the weaknesses in this research have centered around the inability of these studies to show a direct relationship between metamemory ability and memory performance. The present study is valuable as a demonstration of differences in the confidence rating patterns of young and older adults. However, future research needs to examine how metacognitive knowledge is acquired, and how it is related to memory performance.

There is also a need to demonstrate how the metacognitive tasks that researchers have used relate to "real" memory monitoring skills of people in everyday situations. In confidence judgement tasks, for example, researchers need to demonstrate that changes made by subjects in the use of the confidence scale reflect actual changes in their metacognitive skills.

Tulving and Madigan (1970) commented that research concerned with "knowledge of our own knowlegde" may be one of the most important areas to explore in advancing our insights about memory processes. In order to provide this insight, metacognitive research will need to go beyond testing subject performance on single metacognitive tasks, and instead establish a standard procedure of using multiple assessments of memory knowledge to analyze metacognitive abilities.

In conclusion, the results of the present study indicate that even brief feedback about overconfidence may have some effect on lowering the overconfidence of young subjects. The results suggest that using a "money" rating scale, where subjects could win or lose money depending on the accuracy of their confidence ratings, may lead young subjects to be less overconfident in their ratings, especially if more money was at stake than in the present study. Finally, older subjects seem to have more "resolution" problems in using the rating scales provided than do younger subjects, and older subjects are more likely than younger subjects to choose the most extreme 1.ØØ rating.

Money incentives and feedback seem to have little effect on the confidence ratings of older subjects. Older subjects may have become more tired and/or bored with this task as time went on, than younger subjects. Older subjects may also have been effected by the "familiarity" of the task on Trial 2.

Since young and older subjects did show similar ratings on Trial 1, there is no reason to believe that older subjects may actually be different in confidence rating skills than are younger subjects. Subjects of any age who are not familiar with rating scales, and not used to taking tests, may show the same confidence judgment "patterns" shown by the older subjects in this experiment.

Table 1
Performance of young and older subjects on the first trial of the money reinforcement condition *N*=14.

Measure	YM1	OM1	t value
$\bar{X}$ Correct	64.5	64.5	Ø.Ø
$\bar{X}$ Confidence	.773	.816	-1.8Ø
$\bar{X}$ Over-Under Conf	.128	.171	-1.58
$\bar{X}$ CAQ	.6Ø5	.465	1.21
$\bar{X}$ Calibration	.Ø39	.Ø58	-1.97
$\bar{X}$ Resolution	.Ø34	.Ø24	1.53

Table 2
Performance of young and older subjects on the first trial of the no money reinforcement condition *N*=14.

Measure	YNM1	ONM1	t value
$\bar{X}$ Correct	66.93	66.79	Ø.Ø5
$\bar{X}$ Confidence	.812	.8Ø8	Ø.15
$\bar{X}$ Over-Under Conf	.143	.14Ø	Ø.Ø7
$\bar{X}$ CAQ	.681	.485	1.46
$\bar{X}$ Calibration	.Ø45	.Ø5	-Ø.4Ø
$\bar{X}$ Resolution	.Ø34	.Ø24	2.13*

*$p < .Ø5$

Table 3
Performance of young and older subjects on trial 2 of the money reinforcement condition N=14

Measure	YM2	OM2	t value
$\bar{X}$ Correct	64.93	64.86	0.03
$\bar{X}$ Confidence	.751	.823	-3.09*
$\bar{X}$ Over-Under Conf	.102	.174	-2.60*
$\bar{X}$ CAQ	.694	.428	2.28*
$\bar{X}$ Calibration	.029	.059	-3.04*
$\bar{X}$ Resolution	.037	.018	3.51*

Table 4
Performance of young and older subjects on the second trial of the no money reinforcement condition N=14

Measure	YNM2	ONM2	t value
$\bar{X}$ Correct	65.71	65.79	-0.03
$\bar{X}$ Confidence	.785	.833	-1.93
$\bar{X}$ Over-Under Conf	.128	.175	-1.35
$\bar{X}$ CAQ	.801	.488	2.73*
$\bar{X}$ Calibration	.045	.060	-1.01
$\bar{X}$ Resolution	.046	.023	5.08*

*$p < .05$

Table 5
Performance of young subjects on trial 1 of the money and no money conditions.*N*=17

Measure	YM1	YNM1	t value
$\bar{X}$ Correct	66.29	66.76	-Ø.21
$\bar{X}$ Confidence	.78Ø	.797	-Ø.93
$\bar{X}$ Over-Under Conf	.117	.13Ø	-Ø.54
$\bar{X}$ CAQ	.726	.698	Ø.25
$\bar{X}$ Calibration	.Ø35	.Ø39	-Ø.58
$\bar{X}$ Resolution	.Ø37	.Ø33	Ø.71

Table 6
Performance of young subjects on trial 2 of the money and no money conditions. *N*=17

Measure	YM2	YNM2	t value
$\bar{X}$ Correct	66.82	66.18	Ø.28
$\bar{X}$ Confidence	.753	.772	-1.Ø5
$\bar{X}$ Over-Under Conf	.Ø85	.11Ø	-Ø.99
$\bar{X}$ CAQ	.78Ø	.777	Ø.Ø3
$\bar{X}$ Calibration	.Ø25	.Ø4Ø	-1.71
$\bar{X}$ Resolution	.Ø38	.Ø44	-1.27

Table 7
Performance of young subjects in trials 1 and 2 of the money reinforcement condition. N=17

Measure	YM1	YM2	t value
$\bar{X}$ Correct	66.29	66.82	-Ø.67
$\bar{X}$ Confidence	.78Ø	.753	3.59*
$\bar{X}$ Over-Under Conf.	.117	.Ø85	3.37*
$\bar{X}$ CAQ	.726	.78Ø	-Ø.9Ø
$\bar{X}$ Calibration	.Ø35	.Ø25	3.42*
$\bar{X}$ Resolution	.Ø37	.Ø38	-Ø.Ø5

Table 8
Performance of young subjects for trials 1 and 2 of the no money reinforcement conditions. N=17

Measure	YNM1	YNM2	t value
$\bar{X}$ Correct	66.76	66.18	Ø.71
$\bar{X}$ Confidence	.797	.772	4.16*
$\bar{X}$ Over-Under Conf.	.13Ø	.11Ø	2.Ø8
$\bar{X}$ CAQ	.698	.777	-1.51
$\bar{X}$ Calibration	.Ø39	.Ø4Ø	-Ø.33
$\bar{X}$ Resolution	.Ø33	.Ø44	-3.68*

*$p < .Ø5$

Table 9
Performance of older subjects on the first trial of the money and no money reinforcement conditions. N=12

Measure	OM1	ONM1	t value
$\bar{X}$ Correct	65.58	66	-Ø.14
$\bar{X}$ Confidence	.8Ø6	.8Ø4	Ø.Ø4
$\bar{X}$ Over-Under Conf.	.15Ø	.144	Ø.16
$\bar{X}$ CAQ	.5Ø7	.453	Ø.34
$\bar{X}$ Calibration	.Ø5Ø	.Ø54	-Ø.29
$\bar{X}$ Resolution	.Ø26	.Ø24	Ø.39

Table 1Ø
Performance of older subjects on trial 2 of the money and no money reinforcement conditions. N=12

Measure	OM2	ONM2	t value
$\bar{X}$ Correct	65.75	65.83	-Ø.Ø2
$\bar{X}$ Confidence	.814	.825	-Ø.37
$\bar{X}$ Over-Under Conf.	.156	.167	-Ø.28
$\bar{X}$ CAQ	.475	..483	-Ø.Ø6
$\bar{X}$ Calibration	.Ø52	.Ø57	-Ø.31
$\bar{X}$ Resolution	.Ø2	.Ø22	-Ø.34

Table 11
Performance of older subjects on the first and second trials of the money reinforcement condition. N=12

Measure	OM1	OM2	t value
$\bar{X}$ Correct	65.58	65.75	-Ø.12
$\bar{X}$ Confidence	.8Ø6	.814	-.84
$\bar{X}$ Over-Under Conf.	.15Ø	.156	-.4Ø
$\bar{X}$ CAQ	.5Ø7	.475	Ø.37
$\bar{X}$ Calibration	.Ø5Ø	.Ø52	-Ø.32
$\bar{X}$ Resolution	.Ø26	.Ø2	1.28

Table 12
Performance of older subjects on trial 1 and 2 of the no money reinforcement condition. N=12

Measure	ONM1	ONM2	t value
$\bar{X}$ Correct	66.ØØ	65.83	Ø.23
$\bar{X}$ Confidence	.8Ø4	.825	-2.85*
$\bar{X}$ Over-Under Conf.	.144	.167	-2.36*
$\bar{X}$ CAQ	.453	.483	-Ø.51
$\bar{X}$ Calibration	.Ø54	.Ø57	-.68
$\bar{X}$ Resolution	.Ø24	.Ø22	Ø.57

*p<.Ø5

Table 13

Item Difficulty Distributions of Test Lists Used by Young and Older Subjects on Trial 1 of the Money and No Money Conditions*

	Number of Errors						
	Ø-2	3-4	5-6	7-8	9-1Ø	11-12	13-14
Older Money Trial 1	33	15	18	13	11	7	3
No Money Trial 1	34	2Ø	17	13	8	7	1
Young Money Trial 1	32	17	2Ø	13	6	9	3
No Money Trial 1	31	2Ø	19	19	7	1	3

*This table lists the number of test items at each difficulty level. There were 14 subjects in each of the four conditions shown

Table 14

Item Difficulty Distributions of Test Lists Used by Young and Older Subjects on Trial 2 of the Money and No Money Conditions*

	Number of Errors						
	Ø-2	3-4	5-6	7-8	9-1Ø	11-12	13-14
Older Money Trial 2	32	12	24	16	9	5	2
No Money Trial 2	34	13	18	18	1Ø	7	Ø
Young Money Trial 2	27	22	19	15	6	9	2
No Money Trial 2	31	18	18	19	1Ø	1	3

*This table lists the number of test items at each level. There were 14 subjects in each of the four conditions shown.

Table 15

Total Frequency of Use of the 6 Point Rating Scale (Ratings from Ø to 5)*

	Rating Scale					
	Ø	1	2	3	4	5
Older Money Trial 1	227	91	166	244	86	586
Older No Money Trial 1	312	74	99	174	185	556
Young Money Trial 1	259	198	183	178	186	396
Young No Money Trial 1	286	82	1Ø6	161	24Ø	525

*Trial 1 of Young and Older Subjects in the Money and No Money Conditions (_n_=14 per condition)

Table 16

Total Frequency of Use of the 6 Point Rating Scale (Ratings from Ø to 5)*

	Rating Scale					
	Ø	1	2	3	4	5
Older Money Trial 2	226	43	2Ø2	249	72	6Ø8
Older No Money Trial 2	256	59	89	177	2Ø7	612
Young Money Trial 2	327	23Ø	155	156	153	379
Young No Money Trial 2	349	118	1Ø7	16Ø	152	514

*Trial 2 of Young and Older Subjects in the Money and No Money Conditions (_n_=14 per condition)

Table 17

Average number of rating and answer changes between Trials 1 and 2 for each condition.

Groups	$\bar{X}$ Answers Changed	$\bar{X}$ Ratings Changed	Answers Changed to Wrong	Answers Changed to Right	Answers Changed Alone	Ratings Changed Alone	Ratings &Answers Both Changed
Young Money (n=17)	13.12	43.47	6.29	6.83	6.18	36.53	6.94
Young No Money (n=17)	16.00	39.82	8.29	7.71	9.00	32.82	7.00
Older Money (n=12)	16.17	40.58	8.00	8.17	7.67	32.08	8.50
Older No Money (n=12)	13.17	41.25	6.67	6.50	5.75	33.83	7.42
Overall Means	14.60	41.34	7.33	7.29	7.22	33.97	7.38

Table 18

Direction of rating changes between Trials 1 and 2.

Young Money (n=17)

	Trial 2					
Trial 1	.50	.60	.70	.80	.90	1.00
.50	242	45	9	3	2	3
.60	96	82	41	12	2	4
.70	51	81	44	27	8	5
.80	21	39	46	68	27	14
.90	4	14	25	46	65	34
1.00	5	11	10	17	37	460

Young No Money (n=17)

	Trial 2					
Trial 1	.50	.60	.70	.80	.90	1.00
.50	310	22	9	6	4	1
.60	64	50	11	13	3	1
.70	28	49	29	24	8	4
.80	26	36	49	41	32	8
.90	10	24	30	80	99	64
1.00	1	1	8	16	45	494

Older Money (n=12)

	Trial 2					
Trial 1	.50	.60	.70	.80	.90	1.00
.50	144	5	25	19	3	14
.60	20	11	30	17	2	11
.70	22	12	53	37	7	21
.80	15	9	32	83	18	39
.90	3	1	7	17	24	19
1.00	19	4	8	36	15	398

Older No Money (n=12)

	Trial 2					
Trial 1	.50	.60	.70	.80	.90	1.00
.50	172	23	27	26	16	8
.60	16	7	14	12	11	9
.70	19	7	10	27	15	11
.80	16	9	19	41	36	24
.90	2	3	12	28	65	44
1.00	3	5	5	21	27	410

The boxed numbers show the amount of ratings that were not changed between Trials 1 and 2 at each rating level. Numbers to the left of the boxed-in values are the amount of ratings that were changed from a higher rating on Trial 1 to a lower rating on Trial 2. (Note: underlined numbers show levels at which subjects were more than twice as likely to change their ratings to lower ratings, i.e. show less confidence in the accuracy of their answer, than to change their ratings from lower ratings on Trial 1 to higher ratings on Trial 2.) Numbers to the right of the boxed-in numbers show the amount of ratings that were changed from a lower rating value to a higher rating value between Trials 1 and 2.

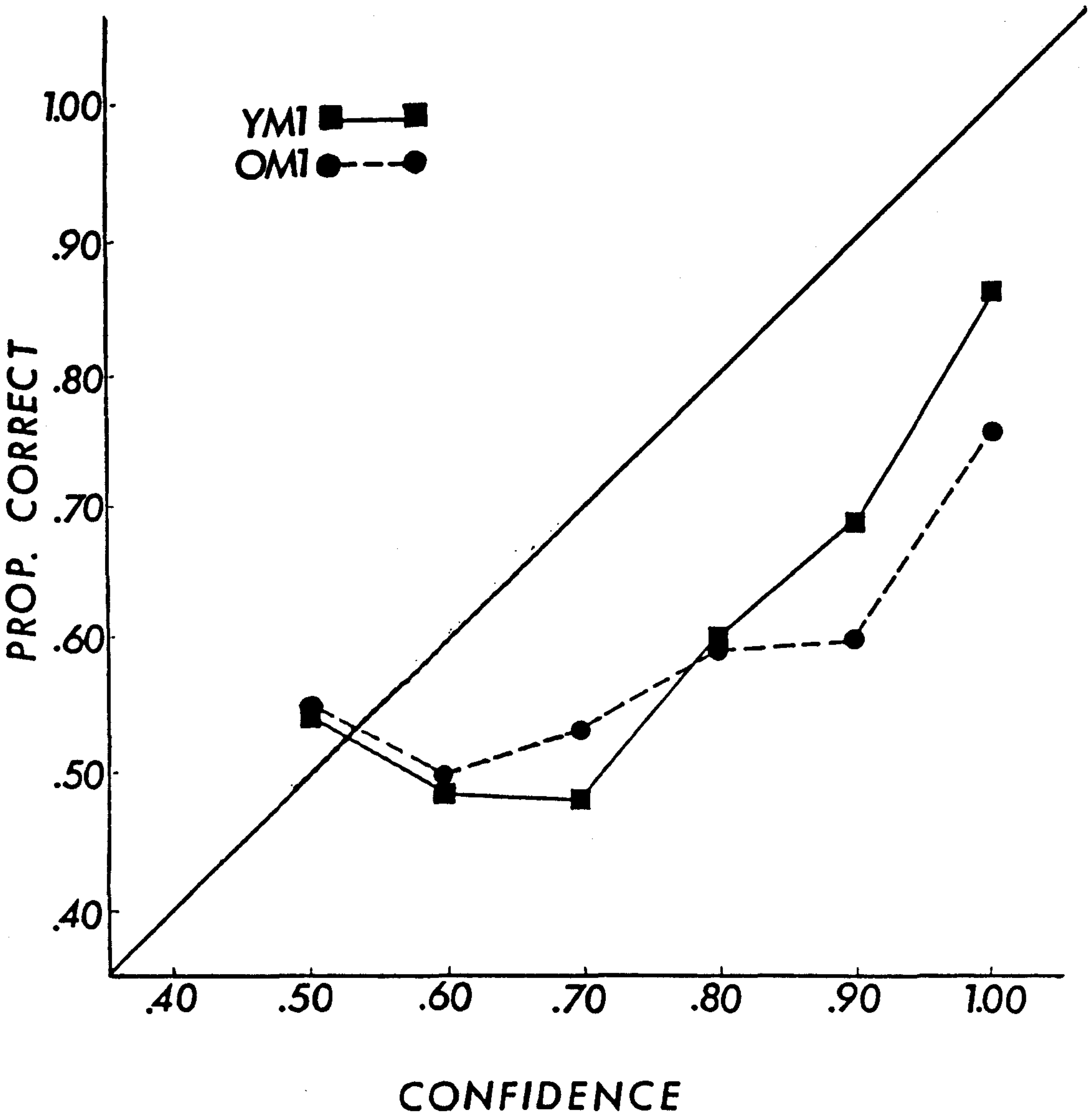

Figure 1. Calibration curves of young and older subjects on Trial 1 of the money reinforcement condition. The young and older subjects were matched for proportion correct on the general information test (N=14).

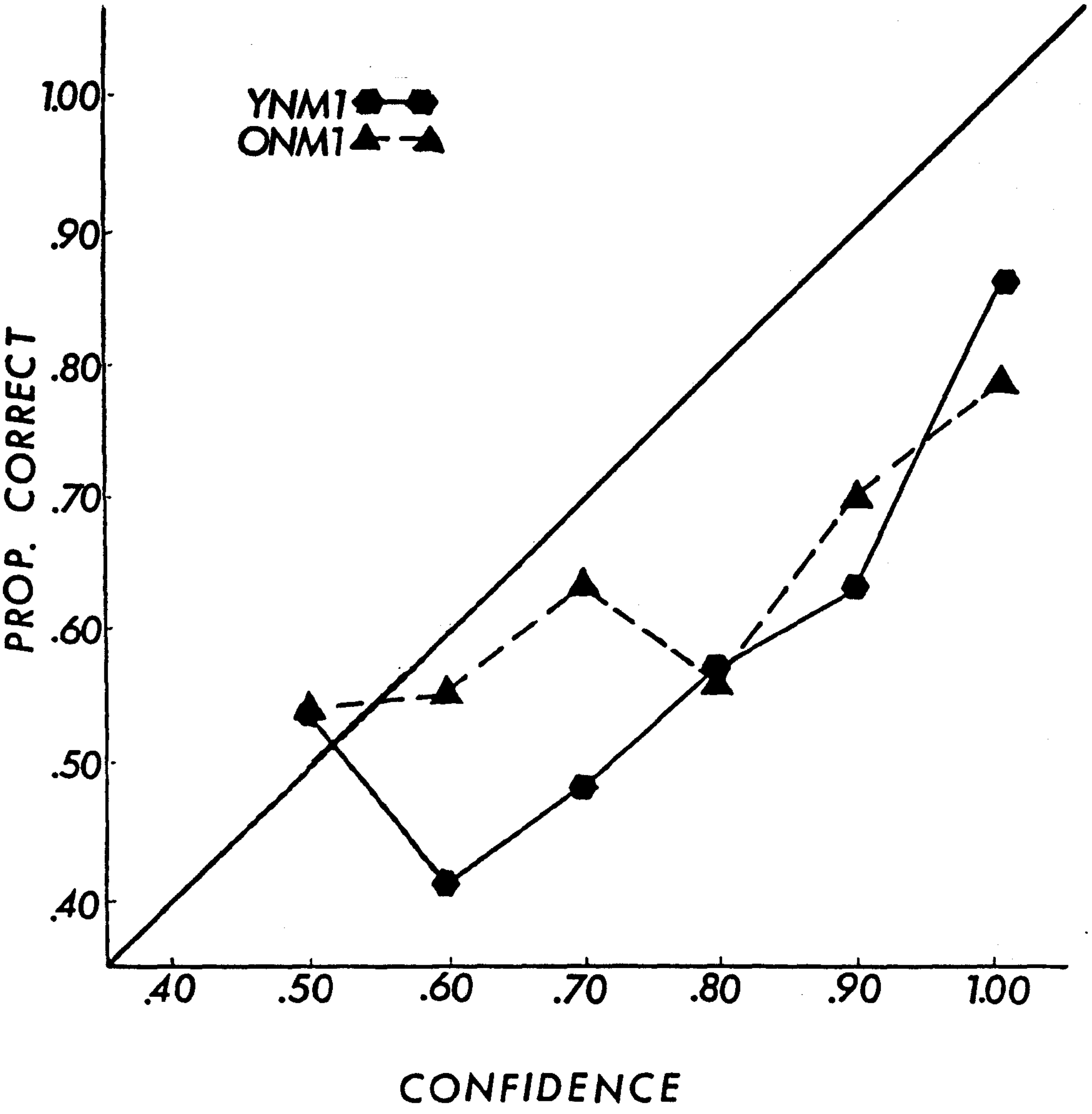

Figure 2. Calibration curves of the young and older subjects on Trial 1 of the no money reinforcement condition. Young and older subjects were matched for number correct (N=14).

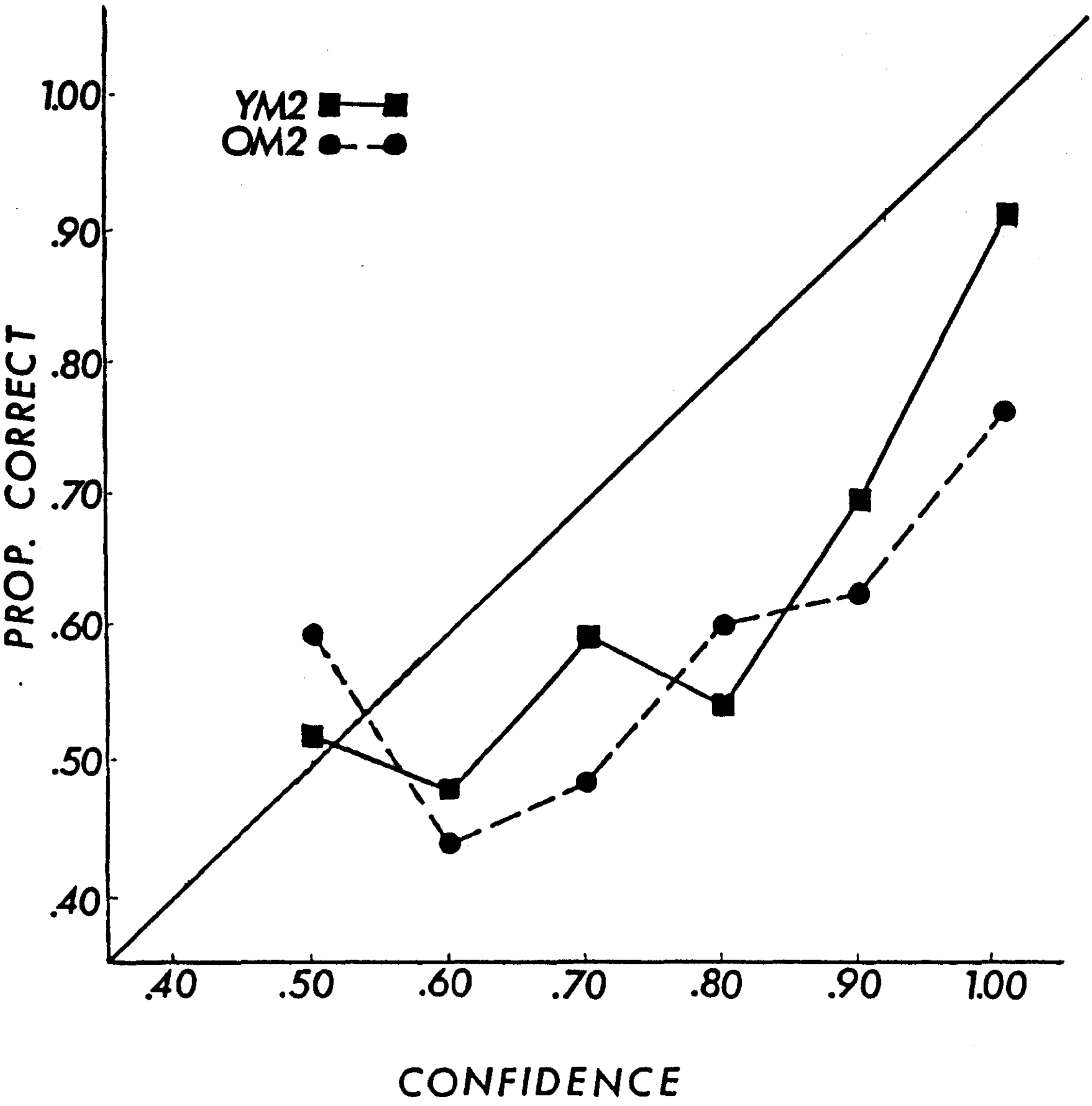

Figure 3. Calibration curves for the young and older subjects on Trial 2 of the money reinforcement condition. Young and older subjects were matched for proportion correct (N=14).

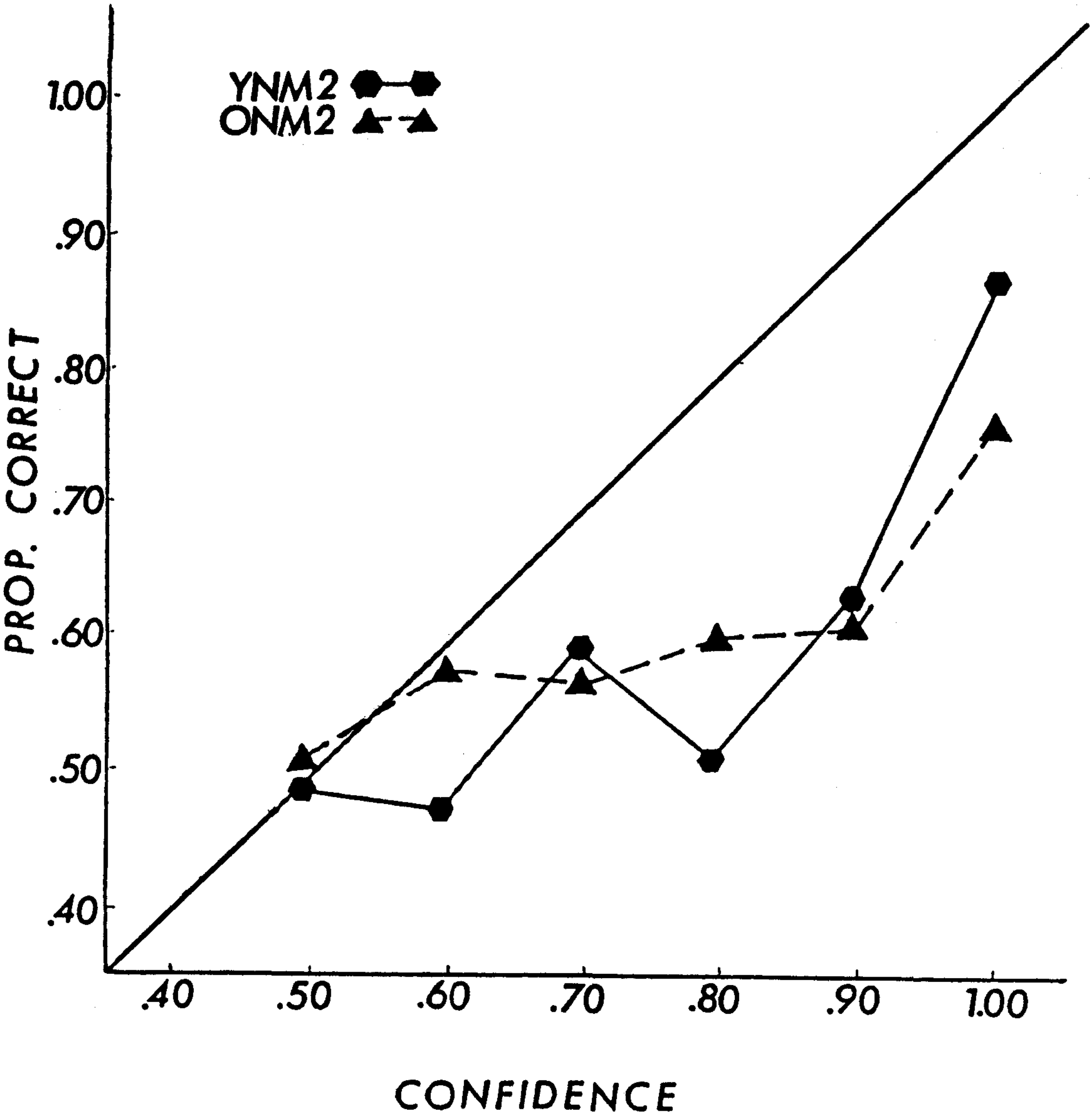

Figure 4. Calibration curves for the young and older subjects on Trial 2 in the no money reinforcement condition. Young and older subjects were matched for proportion correct (N=14).

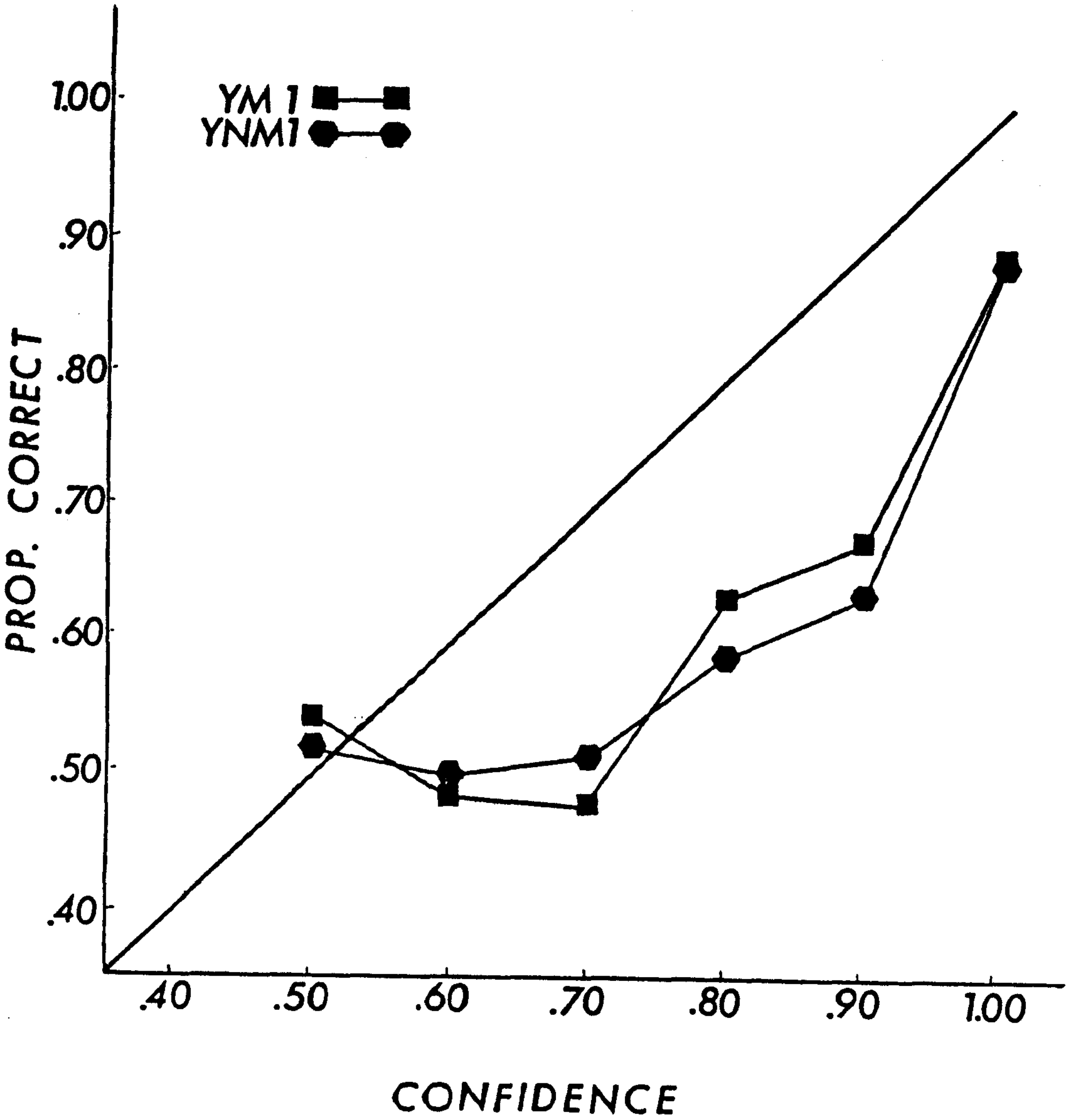

Figure 5. Calibration curves for young subjects on Trial 1 of the money and no money reinforcement conditions. Money and no money subjects were matched for proportion correct (N=17).

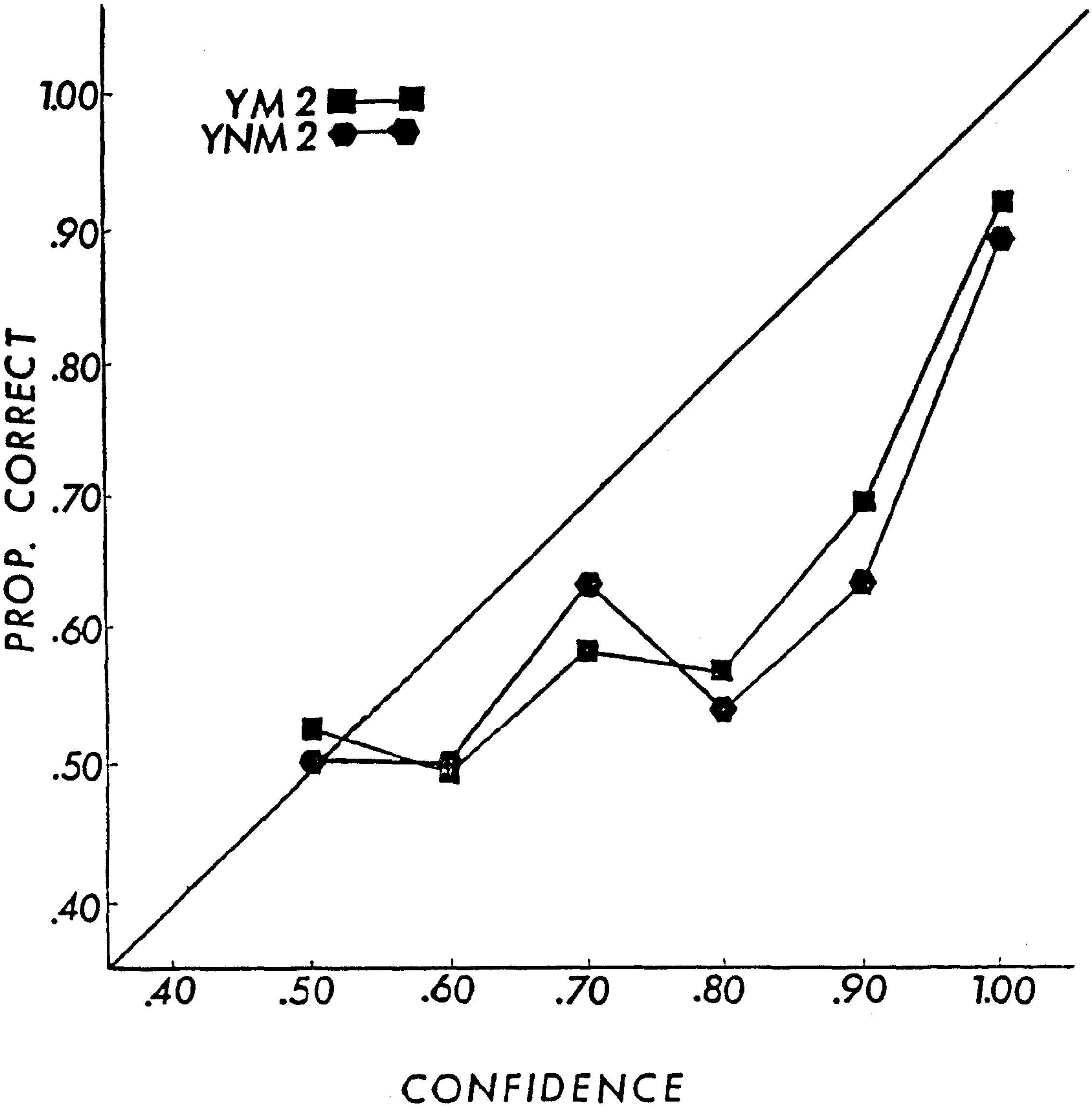

Figure 6. Calibration curves for young subjects on Trial 2 of the money and no money conditons. Money and no money subjects were matched for proportion correct (N=17).

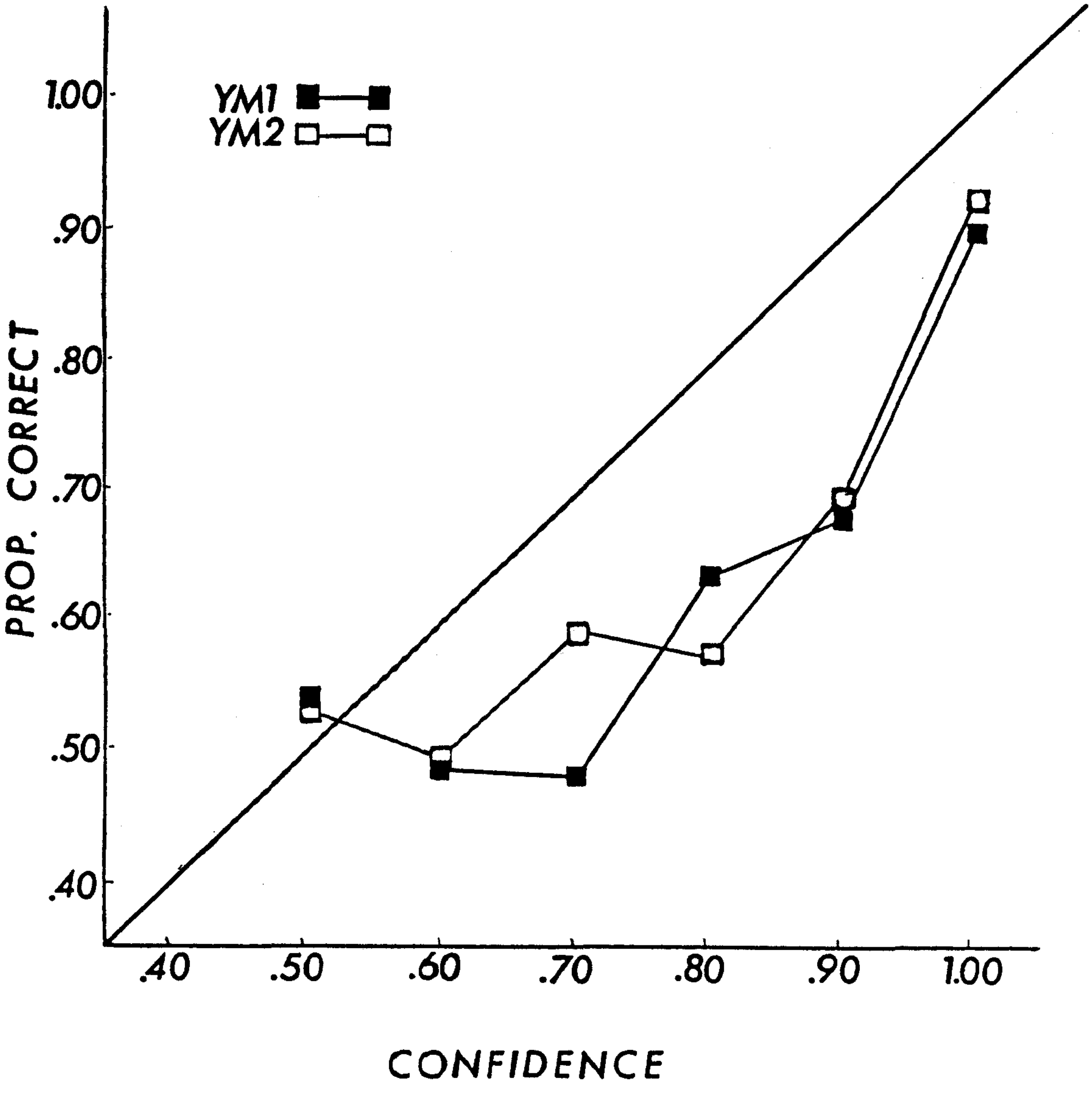

Figure 7. Calibration curves for young subjects on Trial 1 and 2 of the money reinforcement conditions. Subjects were matched for proportion correct so that the subjects on Trial 1 are the same as those in Trial 2 (N=17).

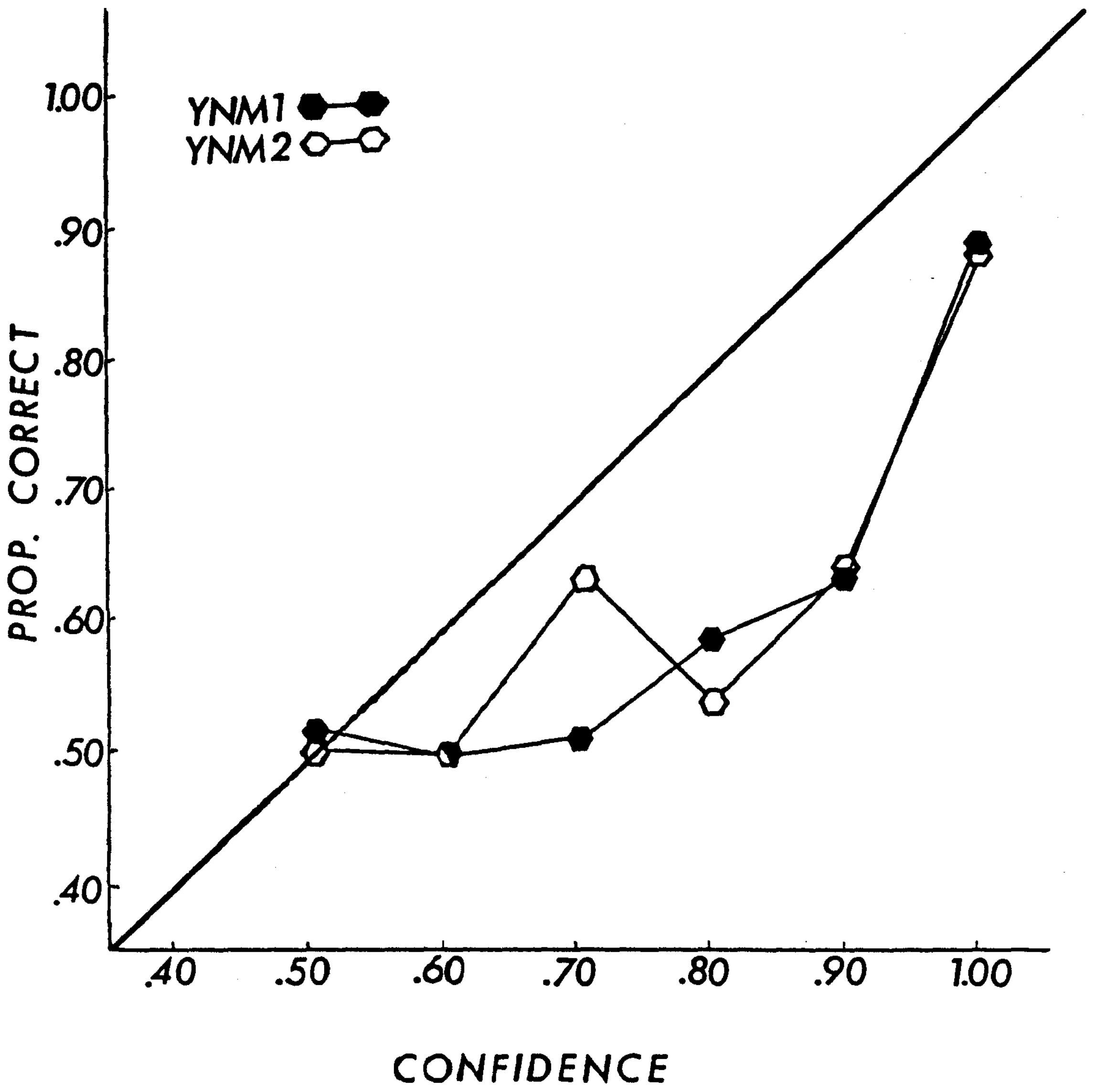

Figure 8. Calibration curves for the young subjects between Trial 1 and Trial 2 of the no money reinforcement condition. Subjects were matched for proportion correct so that subjects on Trial 1 are the same as those on Trial 2 (N=17).

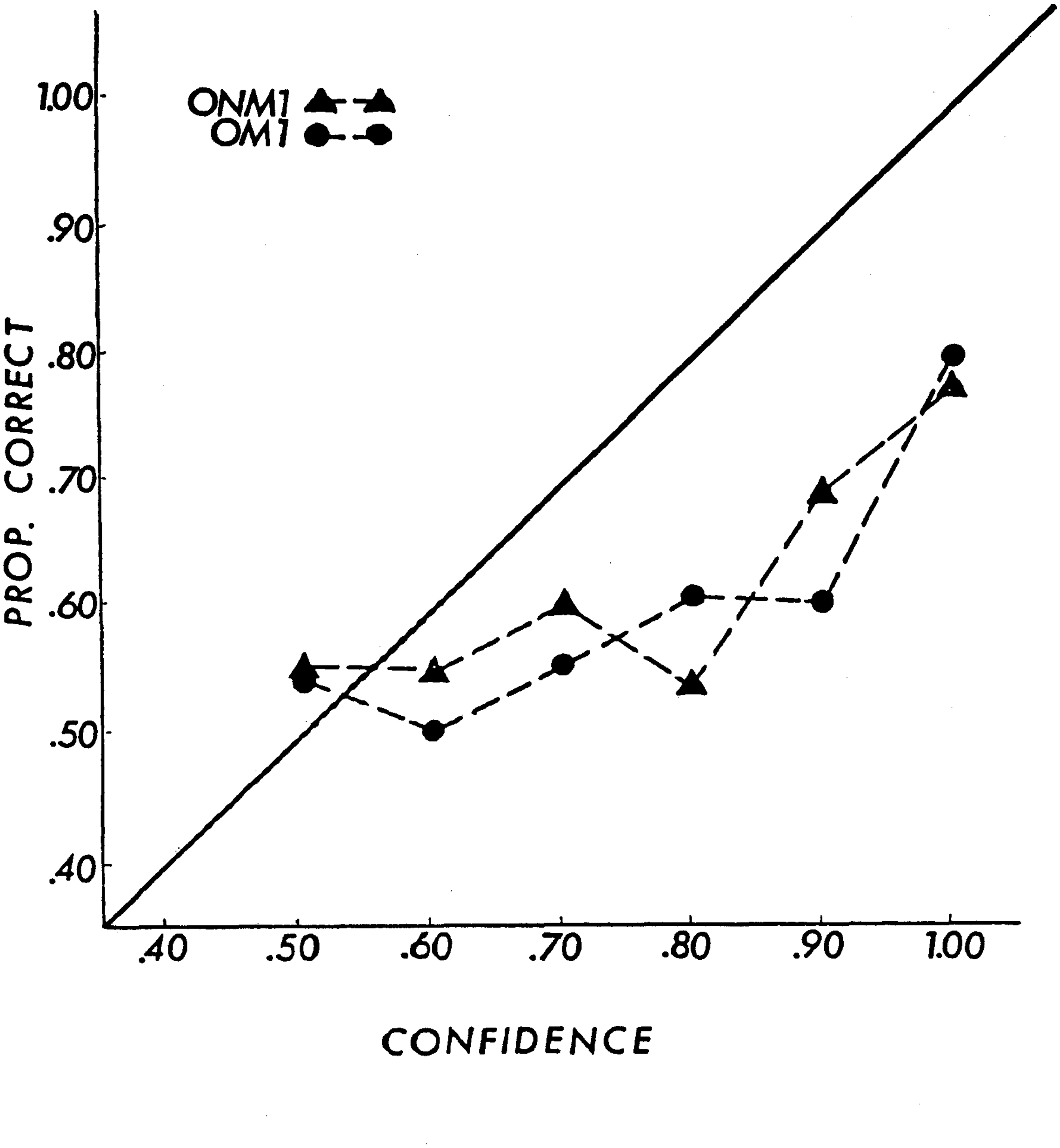

Figure 9. Calibration curves for the older subjects on Trial 1 of the money and no money reinforcement conditions. Subjects in the money and no money conditions were matched for proportion correct (N=12).

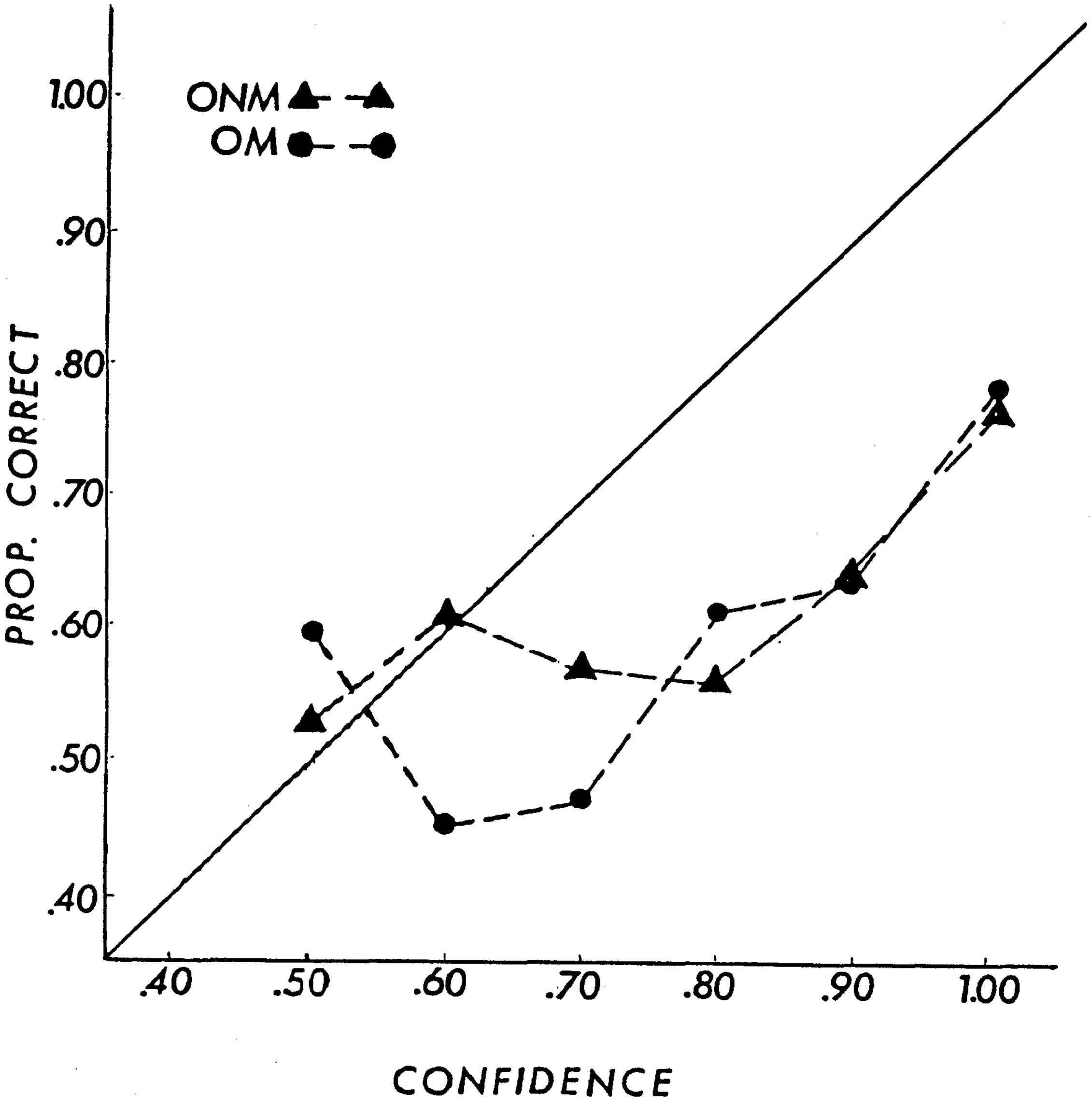

Figure 1Ø. Calibration curves for older subjects on Trial 2 of the money and no money reinforcement conditions. Subjects in the money and no money conditions were matched for proportion correct (N=12).

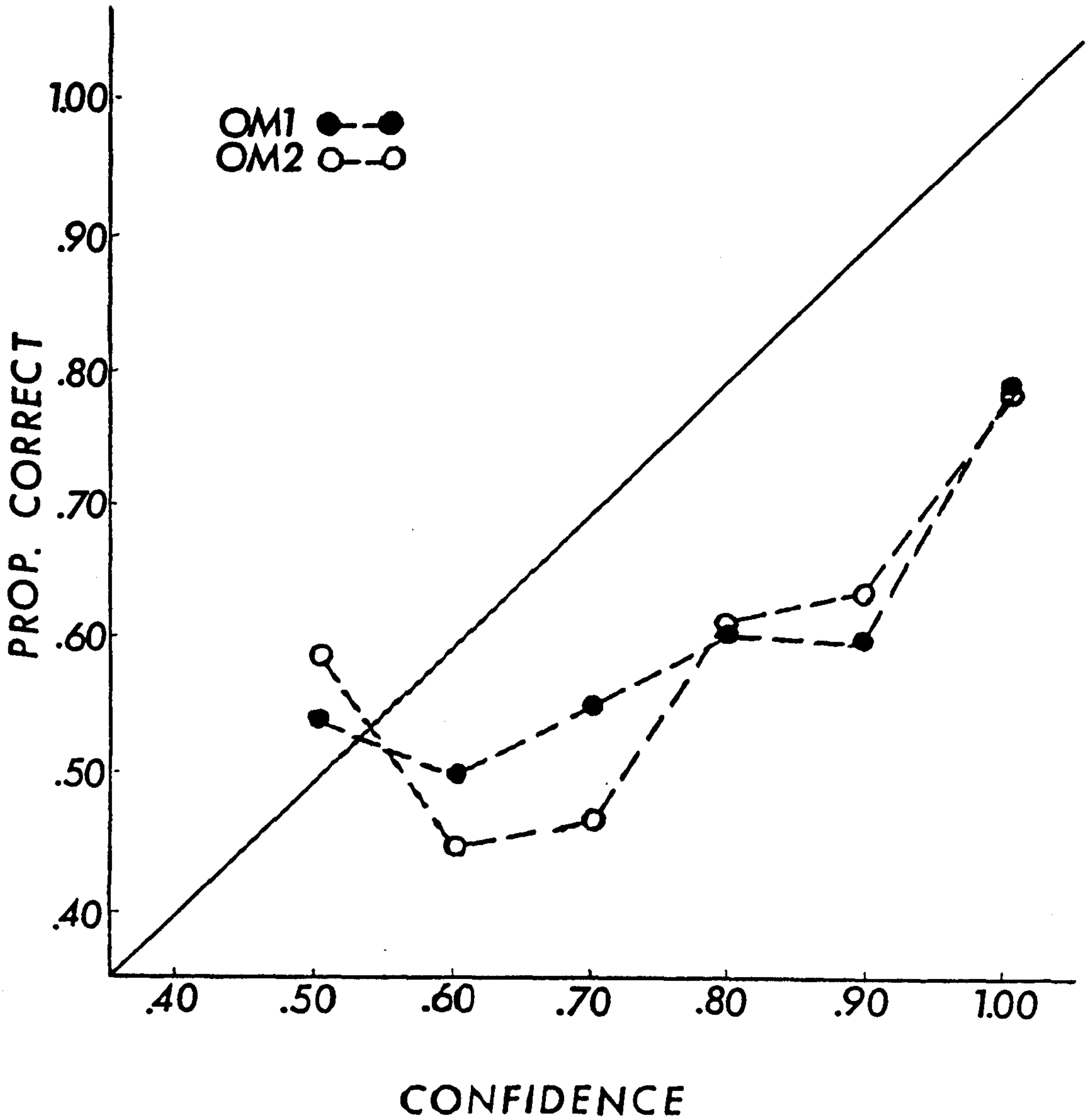

Figure 11. Calibration curves for older subjects between Trial 1 and Trial 2 of the money reinforcement condition. Subjects were matched for proportion correct so that the subjects on Trial 1 are the same subjects as those on Trial 2 (N=12).

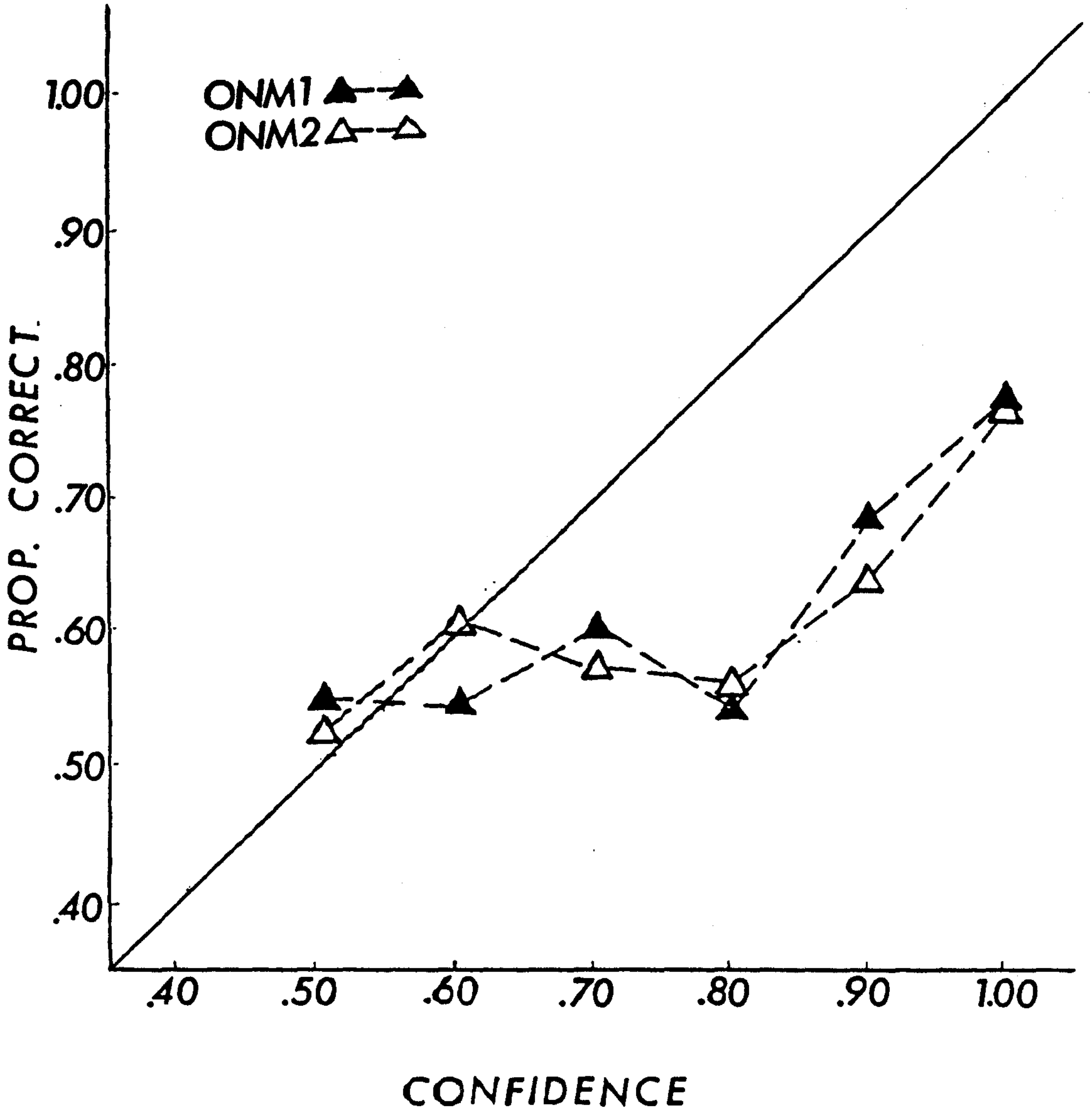

Figure 12. Calibration curves for older subjects between Trial 1 and Trial 2 of the no money reinforcement condition. Subjects were matched for proportion correct so that the subjects in Trial 1 are the same subjects as those in Trial 2 (N=12).

REFERENCES

Arbuckle, T. Y., & Cuddy, L. L. Discrimination of item strength at time of presentation. Journal of Experimental Pschology , 1969, 81 (1), 126-131.

Arkes, H. R., Lai, C., & Hacket, C. A. Two methods for reducing overconfidence. Paper presented at the meeting of the Midwestern Psychological Association, Minneapolis, MN. 1982, May.

Berch, B. B., & Evans, R. C. Decision processes in children's recognition memory. Journal of Experimental Child Psychology , 1973, 16 , 148-164.

Birkhill, W. R. & Schaie, K. W. The effect of differential reinforcement of cautiousness in intellectual performance among the elderly. Journal of Gerontology , 1975, 30 , 5, 578-583.

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

Blake, M. Prediction of recognition when recall fails : Exploring the feeling-of-knowing phenomenon. Journal of Verbal Learning and Verbal Behavior . 1973, 12 , 311-319.

Botwinick, J. Cautiousness in advanced age. Journal of Gerontology , 1966, 21 , 347-355.

Botwinick, J. Disinclination to venture responces vs. cautiousness in responding: Age differences. Journal of Genetic Psychology , 1969, 115 , 55-62.

Brown, A. L. The development of memory: Knowing, knowing about knowing, and knowing how to know. In H. W. Reese & L. P. Lipsett (Eds.) Advances in Child Development and Behavior , Vol. 10. New York: Academic Press, 1975. Pp. 103-152.

Brown, A. L., & Smiley, S. S. Rating the importance of structural units of prose passages: A problem of metacognitive development. Child Development , 1977, 48 , 1-8.

Brown, R. & McNeill, D. The "tip-of-the-tongue" phenomenon. Journal of Verbal Learning and Verbal Behavior , 1966, 5 , 325-337.

Cavanaugh, J. C., & Perlmutter, M. Metamemory: A critical examination. Child Development , 1982, 53 , 11-28.

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.

Flavell, J. H. Developmental studies of mediated memory. In H. W. Reese & L. P. Lipsett (Eds.) Advances in Child Development and Behavior , Vol. 5. New York: Academic Press, 1975. Pp. 181-211.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. Developmental changes in memorization processes. Cognitive Psychology , 197Ø, 1 , 324-34Ø.

Freedman, J. L., & Landauer, T. K. Retrieval of long-term memory: tip-of-the-tongue phenomenon. Psychonomic Science , 1966, 4 , 3Ø9-31Ø.

Gardiner, J. M., & Klee, H. Memory for remembered events: An assessment of output monitoring in free recall. Journal of Verbal Learning and Verbal Behavior , 1976, 15 , 227-233.

Groninger, L. D. Prediction recall : The "feeling-that-I-will-know" phenomenon. American Journal of Psychology , March 1979, 92 , 1, 45-58.

Hart, J. T. Memory and the felling-of-knowing experience. Journal of Educational Psychology , 1965, 56 , 2Ø8-216.

Hart, J. T. Second-try recall, recognition, and the memory-monitoring process. Journal of Educational Psychology , 1967, 58 , 193-197.

Hart, J. T. Memory and the memory-monitoring process. Journal of Verbal Learning and Verbal Behavior , 1967, 6 , 685-691.

Hintzman, D. L. Theoretical implications of the spacing effect. In R. L. Solso (Ed.), Theories in cognitive psychology: The Loyola symposium . Hillsdale, N.J.: Erlbaum, 1974.

James, W. The principles of psychology , Vol. 1. New York: Holt, 1893.

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

Kogan M., & Wallach, M. Age changes in values and attitudes. Journal of Gerontology , 1961, 16 , 272-280.

Kontos, S. Adult-child interaction and the origins of metacognition. Journal of Educational Research , Sept./Oct. 1983, 77 , 1, 43-53.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology : Human Learning and Memory , 1980, 6 , 2, 107-118.

Kreutzer, M. A., Leonard, C., & Flavell, J. H. An interveiw study of children's knowledge about memory. Monographs of the Society for Research in Child Development , 1975, 40 (1, Serial No. 159).

Lachman, J. L., Lachman R., & Thronesbery, C. Metamemory through the adult life span. Developmental Psychology , 1979, 15 , 543-551.

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. Junermann & 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 : Cambribge University Press, 1982

Maki, R. H., & Berry, S. L. Metacomprehension of text material. Journal of Experimental Psychology : Learning, Memory, and Cognition , 1984, 10 , 4, 663-679.

Masur, E. F., McIntyre, C. W., & Flavell, J. H. Developmental changes in apportionment of study time among items in multi-trial free recall task. *Journal of Experimental Child Psychology*, 1973, *15*, 237-246.

Murdock, B. B., Jr. The criterion problem in short term memory. *Journal of Experimental Psychology*. Sept. 1966, *72*, 3, 317-324.

Murphy, A. H. A new vector partition in the probability score. *Journal of Applied Meteorology*, 1973, *12*, 595-600.

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. Metamemory in the aged. *Journal of Gerontology*, 1980, *36*, 2, 185-193.

Nelson, T. O. & Narens, L. Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 338-368.

Nickerson, R. S. & McGoldrick, Jr., C. C. Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Perceptual and Motor Skills*, 1963, *17*, 159-167.

Nickerson, R. S. & McGoldrick, Jr., C. C. Confidence ratings and the level of performance on a judgmental task. *Perceptual and Motor Skills*, 1965, *20*, 311-316.

Okun, M. A., & Elias, C. S. Cautiousness in adulthood as a function of age and payoff structure. *Journal of Gerontology*, 1977, *32*, 4, 451-455.

Okun, M. A. & Di Vesta, F. J. Cautiousness in adulthood as a function of age and instructions. *Journal of Gerontology*, 1976, *31*, 5, 571-576.

Owings, R. A., Peterson, 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.

Perlmutter, M. What is memory aging the aging of?

Developmental Psychology , 1978, *14* , 4, 330-345

Pitz, G. F. Subjective probability distributions for imperfectly know quantities, In L. W. Gregg (Ed.), *Knowledge and Cognition* . New York: Wiley, 1974.

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

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior* , 1966, *5* , 156-163.

Smiley, S. S., Oakley, D. D., Worthen, D., Campione, J. C., & Brown, A. L. Recall of thematically relevant material by adolescent good and poor readers as a function of written versus oral presentation. *Journal of Educational Psychology* , 1977, *69* , 381-387.

Tulving, E., & Madigan, S. A. Memory and verbal learning. *Annual Review of Psychology* , 1970, *21* , 437-484.

Vroom, V. H. & Pahl, B. Relationship between age and risk taking. *Journal of Applied Psychology* , 1971, *55* , 5, 399-405.

Wellman, H. M. Tip of the tongue and feeling of knowing experiences : A developmental study of memory monitoring. *Child Development* , 1977, *48* , 13-21.

Winefield, A. H. & Mullins, G. P. Probability learning and aging. *Journal of Genetic Psychology* , 1980, *136* , 55-64.

Yarmey, A. D. I recognize your face but I can't remember your name : Further evidence on the tip-of-the-tongue phenomenon. *Memory & Cognition* , 1973, *1* , 3, 287-290.

Zechmeister, E. B., Christensen, J., & Rajkowski, B. What is known about what is known? : Predicting recall without prior test trials. Paper given at Fifty-second Annual Meeting of the Midwestern Psychological Association, St. Louis, MO. May 1980.

Zechmeister, E. B., Rusch, M. R.,& Markell, K. A.

Training college students to assess accurately what they know and don't know. Human Learning , In press.

Zechmeister, E. B., Shaughnessy, J. J., When you know that you know and when you think that you know but you don't. Bulletin of the Psychonomic Society , 198Ø, 15 ,(1), 41-44.

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.

Zivian, M. T. & Darjes, R. W. Free recall by in-school and out-of-school adults : Performance and metamemory. Developmental Psychology , 1983, 19 , 4, 513-52Ø.

APPENDIX A

The following are the statements that subjects responded true or false to. Correct answers are on the answer sheet (Appendix B):

1. The capitol of France is Paris.
2. The Bismarck is the name of Germany's largest battleship that was sunk in World War II.
3. The Hague is located in Belgium.
4. Popeye is the name of the cartoon character who eats spinach.
5. Raymond is the last name of the doctor who performed the first successful human heart transplant.
6. There are 2.54 centimeters in an inch (to the nearest hundreth).
7. Amigo is the name of the Lone Ranger's Indian side-kick.
8. Montgomery was the last name of the actor who portrayed the father on the television show "Father Knows Best".
9. Ibsen wrote the "Iceman Cometh".
10. Salk is the last name of the doctor who first developed a vaccine against polio.
11. Dormancy is the name of the long sleep that some animals go through during the entire winter.
12. Ravel composed "Claire de lune".
13. Thunder was the name of Roy Roger's dog.
14. The pancreas is the name of the organ that produces insulin.
15. The island of Sardinia is located in the Mediterranean sea.
16. Rockwell is the last name of the artist who painted "American Gothic".
17. Migraine is the name of the severe headache that returns periodically and often is accompanied by nausea.
18. The French Revolution began in 1730.
19. Orion is the name of the north star.
20. Hockey is the sport in which the Stanley Cup is awarded.
21. C6H12O6 is the chemical formula for dextrose (grape sugar).
22. A javelin is the name of the spear-like object that is thrown during a track meet.
23. Shakespeare is the last name of the man who wrote "Canterbury Tales".
24. Picasso painted "The Three Musicians".
25. Dillenger is the last name of the criminal who was killed by FBI agents outside of a Chicago movie theater.
26. The first air raid occurred in 1849.
27. The Magna Charta was signed in 1320.
28. The visual area of the brain is located in the temporal lobe.

29. Nightengale is the last name of the woman who founded the American Red Cross.
30. Descarte wrote the Dioptrice.
31. Backus is the last name of the man who was the voice of Mr. Magoo.
32. A sextant is the name of the navigation instrument used at sea to plot position relative to the magnetic North Pole.
33. Anthony is the last name of Flash's girlfriend in the comic strip "Flash Gordon".
34. In addition to the Kentucky Derby and the Belmont Stakes, the Preakness is the horse race that completes the triple crown.
35. Bush is the last name of the vice-president under the Reagan administration.
36. Gagarin is the last name of the cosmonaut who was the first person to orbit around the earth.
37. Granger was the last name of Billy the Kid.
38. A meteor is the name for the astronomical bodies that enter the earth's atmosphere.
39. Schultz is the last name of the man who created the comic strip "Li'l Abner".
40. The name of von Frisch is usually associated with the biological studies of bees.
41. Madison is the last name of the 4th U.S. president.
42. The drachma is the monetary unit in the country of Egypt.
43. Three fourths of the world's cacao comes from South America.
44. Angora is the breed of cat that has blue eyes.
45. Gantry is the last name of the football player known as "The Galloping Ghost".
46. Garland is the last name of the singer who made a hit recording of the song "Who's Sorry Now?".
47. Polo is the sport in which a rider on horseback hits a ball with his mallot.
48. Nebula is the name of the brightest star in the sky exclusing the sun.
49. Occur is the name of the substance derived from a whale that is used to make perfume.
50. A balk is the name of an illegal move by a baseball pitcher that results in all runners advancing one base.
51. Stone is the last name of the author of "The Agony and the Ecstasy".
52. Silver is the metal associated with a 50th wedding anniversary.
53. Mertz was the last name of Lucy's neighbors on the television show "I Love Lucy".
54. The Rhine is the name of the river that runs through Rome.
55. Communism is the most famous work written by Karl Marx.

56. Virgil wrote the "Aeneid".
57. Venezuela is the country in which Angel Falls is located.
58. Erhart is the last name of the first person to complete a solo flight across the Atlantic Ocean.
59. Carnegy is the last name of the man who invented the phonograph.
60. Bannister is the last name of the first man to run the mile in under four minutes.
61. Red is the color name given to a light of 650 millimicrons.
62. Corbett is the last name of the boxer who won the boxing title from John L. Sullivan.
63. Wings is the name of the first movie to receive the academy award for best picture.
64. Arthur is the last name of the twenty-first U.S. president.
65. The technical name for the collar bone is the scapular.
66. Eagle was the name of the Apollo lunar module that landed the first man on the moon.
67. An odometer is the name of the instument used to measure windspeed.
68. Potatoes are native to Ireland.
69. Sydney is the capitol of Austraila.
70. Dickens is the last name of the author who wrote "Oliver Twist".
71. Ford is the last name of the man who supposedly killed Jesse James.
72. Mozart is the last name of the composer who wrote "Don Giovanni".
73. Rothstein is the last name of the husband-wife spies who were electrocuted for passing atomic secrets.
74. Ollie was the name of the clownon the "Howdy Doody" television show.
75. E. Harriman founded the Standard Oil Company.
76. Adonis was the god of love.
77. In the Bible, Abraham had the coat of many colors.
78. Bergman was the last name of the female star in the movie "Casablanca".
79. Monticello is the name of the mansion in Virginia that was Thomas Jefferson's home.
80. Clampet is the name of the hillbilly family that had a famous feud with the McCoy family.
81. Ventnor is the name of the avenue that immediately follows Atlantic Avenue in the game of Monopoly.
82. Medusa is the name of the legendary one-eyed giant in Greek mythology.
83. Marlowe is the last name of the author who wrote "Romeo and Juliet".
84. Skinner is usually credited with developing the technique of operant conditioning.

85. Syria borders the Mediterranean sea.
86. Jason was the leader of the Argonauts.
87. The Yangtze is the longest river in Asia.
88. West is the last name of Batman's secret identity in the Batman comics.
89. Cody was the last name of Buffalo Bill.
90. Powers is the last name of the pilot of the U-2 spy plane shot down over Russia in 1960.
91. Pluto was the last planet to be discovered.
92. Kahlil Gibran was most inspired by the Buddhist religion.
93. Floyd is the last name of the criminal who was known as "Scarface".
94. Ruby is the last name of the man who assasinated President John F. Kennedy.
95. Fleming is the last name of the author of the James Bond novels.
96. The Dodgers won the 1959 World Series.
97. An ohm is the unit of electrical power that refers to a current of one ampere at one volt.
98. Grant is the last name of the union general who defeated the confederate army at the Civil War battle of Gettysburg.
99. Wells is the last name of the author of "The War of the Worlds".
100. Yahtze is the game which uses a doubling cube.

APPENDIX B

Answer Key to the General Information Statements (T=True, F=False):

1. T
2. T
3. F
4. T
5. F
6. T
7. F
8. F
9. F
10. T
11. F
12. F
13. F
14. T
15. T
16. F
17. T
18. F
19. F
20. T
21. T
22. T
23. F
24. T
25. T
26. T
27. F
28. F
29. F
30. F
31. T
32. F
33. F
34. T
35. T
36. T
37. F
38. T
39. F
40. T
41. T
42. F
43. F
44. F
45. F
46. F
47. T
48. F
49. F
50. T
51. T
52. F
53. T
54. F
55. F
56. T
57. T
58. F
59. F
60. T
61. T
62. T
63. T
64. T
65. F
66. T
67. F
68. F
69. F
70. T
71. T
72. T
73. F
74. F
75. F
76. F
77. F
78. T
79. T
80. F
81. T
82. F
83. F
84. T
85. T
86. T
87. T
88. F
89. T
90. T
91. T
92. F
93. F
94. F
95. T
96. T
97. F
98. F
99. T
100. F

APPENDIX C

Questions-Only Instructions

"In this study I am interested in what people "know about what they know". For instance, if I asked you a question and you gave me an answer, how sure are you that your answer is correct. We'll go through two examples: If I gave you the statement "The planet Mars is three light years from the Earth" and asked you whether it was true or false, you might say "if I have to give you an answer, I'll say its true, but I really have no idea whether the statement is true or false, so I would be taking a guess." However, if I gave you the statement "Mayor Byrne is the present mayor of Chicago" * you would probably tell me that the statement is true, and that in fact you are absolutely sure that your response is correct. You are positive that Mayor Byrne is the mayor of Chicago. So, when responding to statements as true or false, sometimes you have no idea whether your response is correct or not, sometimes you may be a little sure or fairly sure that your response is correct, and sometimes you may be positive that you know you have given the correct response.

I am going to give you a stack of cards and an answer sheet. On each card is a general information statement, with one word, name or date underlined. I want you to treat each statement as though it were a fill-in-the-blank question, where the blanks have already been filled in. I want you to decide two things for each statement. First of all, do you think the blank has been filled in correctly, or has it been filled in incorrectly? If you think it is correct, circle true. If you think it is wrong, circle false. Secondly, and most importantly, I want you to rate how sure you are that your decision is correct in terms of the scale, from Ø to 5, on the answer sheet. It is not important how many of your decisions are correct, but it is important that you be as accurate as you can in rating whether your decision is correct or not. If you really have no idea about whether or not a statement has been filled in correctly, you would just be taking a guess when making your decision, so you would probably want to circle Ø, which would indicate that you really have no idea whether your true/false response is correct or not. If you are somewhat sure that your decision is correct, you may want to circle a 1 or 2. If you are fairly sure that your decision is correct, you may want to circle a 3 or 4. If you have no doubt that your decision is correct, you are sure it is correct, then you may want to circle a 5. Only circle a 5 when you are absolutely sure that your decision is correct. Make sure that you give a decision and circle a rating for all 1ØØ statements. I do not expect that you will know the information in all the statements, but about half of the information in the statements will probably be familiar to you. It is not important how many of your responses are correct, but it is important that you be as accurate as you can be in your ratings. When you are done, you can take a 5 minute break and then we will begin the next part.

Second part

I want to give you some feedback about your ratings. The easiest way for me to tell how accurate you have been in rating the correctness of your responses is to correct your answers, and then, for each incorrect response you made, add up the number of points that you gave that response, in terms of how sure you were that it was correct. If you were very accurate in your ratings, the total rating score for all your incorrect responses should be low, indicating that you were not sure that these answers were correct. Your total score was ____. It is ,of course, difficult to tell what this number may mean about your performance, since you are not able to compare it to a number indicating average subject performance. However, in a general way, this number indicates that sometimes, when you thought an answer was correct, and gave it a high rating, it turned out to be incorrect. Please respond to and rate the questions again. Do not worry about what your responses were on the first trial. Try to approach the questions as if you were answering them for the first time. Also, try to be as accurate as you can in your confidence ratings for your responses, since this is the most important part of your response. " (Note: The feedback given to this group is purposely ambiguous. The number given to them, though an accurate measure of their total ratings for wrong answers, has very little meaning, since they are given no comparison values. The explanation given of the number they received is the important feedback. They are told that sometimes when they thought an answer was correct, it was not correct. Therefore they are given simple, brief feedback that they have been overconfident in some of their ratings.)

Monetary Incentive Instructions

In this study I am interested in what people "know about what they know". For instance, if I asked you a question and you gave me an answer, how sure are you that your answer is correct. We'll go through two examples: If I gave you the statement "The planet Mars is three light years from the Earth" and asked you whether it was true or false, you might say "If I have to give you an answer, I'll say its true, but I really have no idea whether the statement is true or false,so I would be taking a guess". However, if I gave you the statement "Mayor Byrne is the present mayor of Chicago" you would probably tell me that the statement is true, and that in fact you are absolutely sure that your response is correct. You are positive that Mayor Byrne is the mayor of Chicago. So, when responding to statements as true or false, sometimes you have no idea whether your response is correct or not, sometimes you may be a little sure or fairly sure that your response is correct, and sometimes you may be positive that you know you have given the correct response.

I am going to give you a stack of cards and an answer sheet. On each card is a general information statement, with one word, name or date underlined. I want you to treat each statement as though it were a fill-in-the-blank question, where the blanks have already been filled in. I want you to decide two things for each statement. First of all, do you think the blank has been filled in correctly, or has it been filled in incorrectly? If you think it is correct, circle true. If you think it is wrong, circle false. Secondly, and most importantly, I want you to rate how sure you are that your decision is correct. I want you to do this in terms of the amount of money (in pennies) that you would be willing to bet that your decision is correct. It is not important how many of your decisions are correct, but it is important that you be as accurate as you can in rating whether your decision is correct or not. This jar contains the maximum amount of pennies that you could win if all of your ratings, and therefore all of your bets are accurate. You also cannot lose more than this amount of money, so this does not at all involve having you use any of your own money. Think about placing your bets using the following procedure: imagine that the numbers from Ø to 5 by each of your decisions represents the number of pennies that you want to bet that your decision is correct. If you really have no idea about whether a statement has been filled in correctly or not, you would just be taking a guess when making your decisions, so you might not want to place any money on the probability that your decision is correct, and therefore you might want to circle Ø pennies. If you are somewhat sure that your decision is correct you may want to bet 1 or 2 pennies on the correctness of your decision. If you are fairly sure that your decision is correct, then you may want to bet 3 or 4 cents. If you have no doubt that your decision is correct, you are sure it is correct, then you may want to bet 5 cents. Only bet 5 cents when you are absolutely sure that your decision is correct. Make sure that you give a decision and circle the amount of pennies you want to place on the correctness of that decision for all 1ØØ statements. We do not expect that you will know the information in all the statements, but about half of the information in the statements will probably be familiar to you. It is not important how many of your decisions are correct, but it is important that you be as accurate as you can be in your ratings. When you are done, you can take a 5 minute break and then we will begin the next part.

<u>Second part</u>

I want to give you some feedback about your rating bets. The easiest way for me to tell how accurate you have been in rating the correctness of your response is to correct your answers, and then, for each incorrect response you made, add up the number of pennies that you bet on that response, in terms of how sure you were that it was correct. That is the number of pennies that you have lost. If you were very accurate in your bets, then the total number of pennies that you would have bet for all your incorrect responses should be low, indicating that you were not sure that these answers were correct. The total number of pennies you have lost is ____. It is difficult to tell exactly what this number means, since you don't know how much money you have won for your bets on your correct answers. However, this indicates that sometimes when you thought an answer was correct, and placed a higher number of pennies on it for your bet, it turned out to be incorrect, and you lost those pennies. Please answer and place bets on the questions again, and try to be as accurate as you can be with your bets. If you are more accurate in your rating bets this time, you can win back some or all of the money that you lost, since I will give you money from the trial you are most accurate on. Don't worry about the responses you gave last time. Try to approach the questions as though you are responding to them for the first time, and remember that it is important that you be as acccurate as you can be with your bets.(In actuality, due to limited funds,subjects were offered a maximum of $2 after the entire experiment was over, no matter how high their actual winnings were. Approximately 1/2 of the subjects of both age groups took their winnings, the others said they had enjoyed the study, and did not want to take their winnings, even though money won was offered to each of these subjects several times).

APPENDIX D

Subject Selection

Two forms of the general information questionnaire were originally planned to be administered to subjects; an "easy" form and a more "difficult" form. The difference between the two questionnaires was that every third question from questions 1 to 9Ø was an "easy" question on form A (selected so that it was answered correctly over 75% of the time on pilot studies) or a "difficult" question on form B (selected so that it was answered correctly less than 5Ø% of the time in pilot studies). The criterion for deciding which form to administer to each subject was the subject's score on the Quick Test. As stated earlier, subjects scoring 43 or above on the QT were given a form B, and subjects scoring less than 43 received form A.

However, after testing many of the older subjects, it was found that very few of the subjects who agreed to participate received less than a score of 43. Since it was difficult to recruit older subjects, form A was discarded from further analyses, and only those subjects who scored a 43 or above on the QT were included in the final analyses.

The subject groups included in the following analyses were matched for proportion of answers correct on the questionnaire to ensure that proportion correct would not be a factor. In order to match subjects as closely as possible, 5 young subjects that otherwise met the criterion for being included in the study were dropped because they could not be matched with subjects in the comparison group (4 in the young no money condition, 1 in the money condition). In addition, of the 17 young subjects used in each condition, 14 of these subjects were compared to the 14 older subjects in each condition. The best match comparing older subjects from Trial 1 to 2 was to drop 2 subjects each in the money and no money conditions.

Preliminary analysis of subject data before matching is shown on calibration curves in Figures 1D and 2D. The data pattern between these unmatched subject comparisons and the matched comparisons shown in the results section (Figures 1 -12) is similar, but since differences in proportion correct are known to have some influence on calibration (Lichtenstein, et al., 1982), all the following analyses are done on matched subjects.

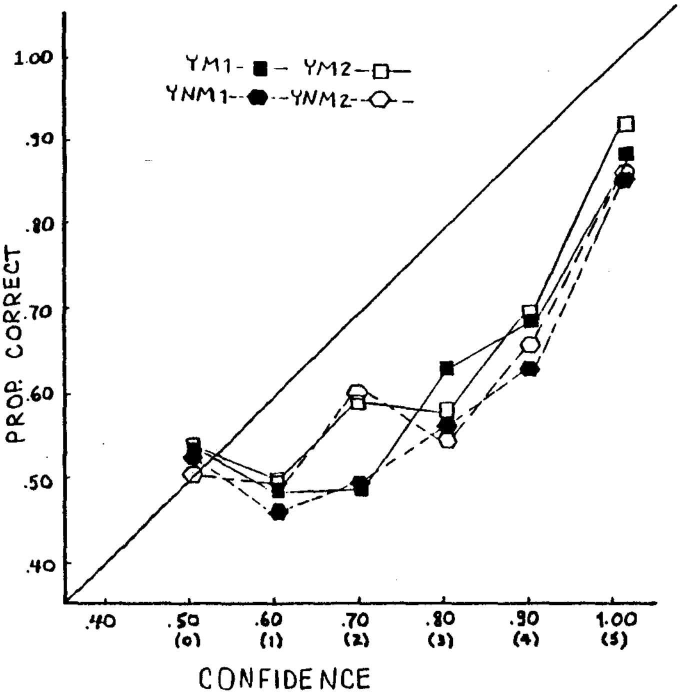

Figure 19. Calibration curves of young subject data for Trials 1 and 2 of the money (n=10) and no money (n=21) conditions.

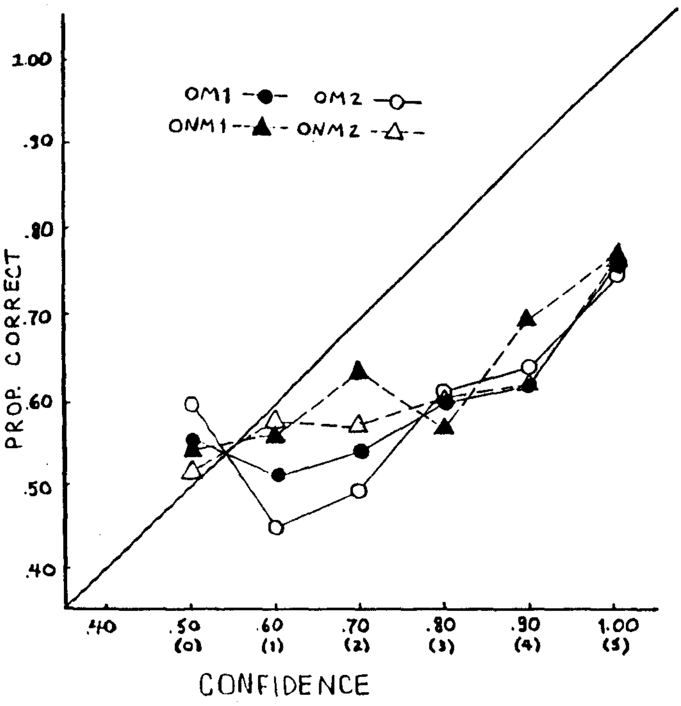

Figure 20. Calibration curves of older subject data for Trials 1 and 2 of the money (n=14) and no money (n=14) conditions.

APPROVAL SHEET

The dissertation submitted by Kathryn A. Markell has been read and approved by the following committee:

Dr. Eugene Zechmeister, Director
Professor, Psychology, Loyola

Dr. Daniel O'Connell
Professor, Psychology, Loyola

Dr. Jill Nagy Reich
Associate Professor, Psychology, Loyola

The final copies have beeen examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

8/18/88 Eugene B Zechmeister
Date Director's Signature