Age-Related Differences in Absolute but Not Relative Metamemory Accuracy

Lisa Tabor Connor
Boston University School of Medicine and Veterans Administration Medical Center, Boston

John Dunlosky
University of North Carolina at Greensboro

Christopher Hertzog
Georgia Institute of Technology

In 3 experiments, the effects of age on different kinds of metacognitive prediction accuracy were assessed. Participants made global memory predictions and item-by-item memory predictions in a single experimental task. Metacognitive accuracy was evaluated with correlational and more traditional difference-score measures. Difference-score measures were found, in some cases, to be sensitive to level of recall performance. Correlational techniques revealed that older adults monitored learning effectively. Relative to younger adults, they showed equally accurate immediate judgments of learning (JOLs), produced an equivalent delayed-JOL effect, and showed equivalent upgrading in the accuracy of their global prediction from before to after study of test materials.

A central topic of research on metamemory and aging is the degree to which individuals of different ages accurately predict their ability to remember recently studied materials (for reviews, see Hertzog & Dixon, 1994; Lovelace, 1990). The issue of age differences in the accuracy of monitoring learning and memory is important to the evaluation of the hypothesis that deficits in metacognition account for some—or even all—of the age differences in episodic memory performance (Light, 1991). Individuals with high levels of metacognitive accuracy ought to be able to use accurate monitoring to control and regulate their strategies for learning and retrieving information from memory (Schneider & Pressley, 1989). Optimal self-regulation should lead to higher levels of memory performance.

A common method for measuring monitoring accuracy has been to ask individuals to predict how well they will remember items during an upcoming memory test (e.g., Perlmutter, 1978). Although, as pointed out by several theorists (e.g., Cavanaugh & Green, 1990; Nelson & Narens, 1990), it is highly plausible to argue that memory predictions are based on monitoring the contents of memory, other variables could also influence memory predictions. For instance, Berry, West, and Dennehey (1989) and Bandura (1989) argued that memory predictions are based on memory self-efficacy within the context of a particular memory task (see Hertzog & Dixon, 1994).

Two different kinds of predictions have been frequently elicited in developmental research: global predictions, in which people judge how many items of an entire study list they will subsequently recall, and item-by-item predictions, in which people predict the likelihood of subsequent recall separately for each item.

Accuracy of Global Predictions

The accuracy of global memory predictions has been widely discussed in the aging literature (Hertzog & Dixon, 1994; Lovelace, 1990). Accuracy has been most frequently operationalized as either the signed or unsigned difference between the number of items predicted and the number of items correctly remembered. Data regarding the accuracy of older adults’ global predictions are somewhat contradictory. Some researchers have found that older adults’ predictions are accurate (e.g., McDonald-Miszczak, Hunter, & Hultsch, 1994, Experiment 1; Rebok & Balcerak, 1989, posttraining condition), whereas others have found substantial differences between older adults’ global predictions and their memory performance (e.g., Bruce, Coyne, & Botwinick, 1982; Perlmutter, 1978). Some studies have indicated that older adults underestimate their memory performance (e.g., McDonald-Miszczak et al., 1994, Experiment 2), but the predominant finding is that they overestimate it (Bruce et al., 1982; Coyne, 1985; Devolder, Brigham, & Pressley, 1990; Mur-
Table 1

Summary of Studies of Mean Global Predictions and Mean Recall Performance

<table>
<thead>
<tr>
<th>Study</th>
<th>Task</th>
<th>Young adults</th>
<th></th>
<th>Old adults</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prediction</td>
<td>Recall</td>
<td>Prediction</td>
<td>Recall</td>
</tr>
<tr>
<td>Bruce, Coyne, &amp; Botwinick (1982)</td>
<td>Free recall*</td>
<td>57</td>
<td>58</td>
<td>48</td>
<td>37</td>
</tr>
<tr>
<td>Coyne (1985)</td>
<td>Free recall*</td>
<td>51</td>
<td>49</td>
<td>48</td>
<td>34</td>
</tr>
<tr>
<td>Hertzog, Dixon, &amp; Hultsch (1990)</td>
<td>Categorized free recall*</td>
<td>52</td>
<td>67</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>Hertzog, Saylor, Fleece, &amp; Dixon (1994)</td>
<td>Free recall*</td>
<td>E1</td>
<td>52</td>
<td>66</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2 (no-information condition)</td>
<td>54</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2 (information condition)</td>
<td>55</td>
<td>65</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E3 (group written condition)</td>
<td>56</td>
<td>64</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E3 (individual written condition)</td>
<td>56</td>
<td>60</td>
<td>48</td>
</tr>
<tr>
<td>Lachman, Steinberg, &amp; Trotter (1987)</td>
<td>Free recall</td>
<td>-</td>
<td>-</td>
<td>70</td>
<td>64</td>
</tr>
<tr>
<td>Lineweaver (1994)</td>
<td>Free recall*</td>
<td>45</td>
<td>52</td>
<td>47</td>
<td>33</td>
</tr>
<tr>
<td>McDonald-Miszczak, Hunter, &amp; Hultsch (1994)</td>
<td>Free recall</td>
<td>E1*</td>
<td>58</td>
<td>83</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2* (15-word condition)</td>
<td>63</td>
<td>88</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2* (45-word condition)</td>
<td>53</td>
<td>65</td>
<td>45</td>
</tr>
<tr>
<td>Rabbit &amp; Abson (1991)</td>
<td>Free recall*</td>
<td>-</td>
<td>-</td>
<td>41</td>
<td>28</td>
</tr>
<tr>
<td>Rebok &amp; Balcerak (1989)</td>
<td>Serial recall</td>
<td>62</td>
<td>71</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>Pretraining</td>
<td></td>
<td>71</td>
<td>85</td>
<td>52</td>
<td>50</td>
</tr>
</tbody>
</table>

Note. All prediction and recall performance levels are expressed in percentages. E = Experiment.
\* Collapsed across imagery and frequency conditions. Results from the 60–69-year-old and 70–79-year-old groups are averaged to produce one older adult group. \( ^{1} \) Participants were told that the midpoint of the scale was how the ‘average’ person would do (50%) in the recall task. Predictions were averaged over gender. Only Trial 1 is reported. \( ^{2} \) Normative information (50%) was given to participants before predictions were made in Experiment 1. \( ^{3} \) Only older adults were included in this study. Average of first and second memory predictions. \( ^{4} \) No-information condition only. \( ^{5} \) Participants generated a prediction for a hypothetical person their age before giving their own memory prediction.

The literature also includes contradictory findings as to whether the accuracy of global predictions differs between older adults and younger adults. Older adults have often been characterized as being more prone to prediction errors, generally in the direction of overestimating memory performance (see Lovelace, 1990). However, two studies have actually indicated better global prediction accuracy for older adults (Hertzog, Dixon, & Hultsch, 1990; Hertzog, Saylor, Fleece, & Dixon, 1994).

We include in Table 1 only those studies that report separate means for predictions and recall. Studies reporting only a derived measure of prediction accuracy, such as the group mean of the absolute differences between predictions and recall (e.g., Devolder et al., 1990; Perlmutter, 1978), are not included. Predictions and recall performance are expressed in percentage of correct recall to make values comparable across studies.
younger adults' predictions would be relatively accurate and older adults would have overpredicted their performance.

Our argument is that this pattern does not necessarily indicate that younger adults are more accurate than older adults in monitoring the contents of their memory. Instead, patterns of age-related differences in predictive accuracy may be indirectly determined by patterns of age differences in the level of recall, in combination with predictions anchored at the midpoint of the possible range of performance for both age groups. Note that this account is consistent with the results of Hertzog et al. (1990) and Hertzog et al. (1994) shown in Table 1—older adults' performance was close to 50%, and their predictions were the more accurate of the two age groups. Although other factors such as task appraisal, memory self-efficacy, and memory monitoring may also affect global prediction accuracy (see Hertzog & Dixon, 1994, for a review), the midpoint-anchoring effect may account for more variance in prediction behavior than has been appreciated. One goal of the current research was to evaluate this possible dependence between global prediction accuracy (as measured by difference scores) and level of recall performance.

An alternative measure of accuracy is a correlation between global predictions and recall performance across individuals within each age group (e.g., Hertzog et al., 1994; Lachman, Steinberg, & Trotter, 1987). In contrast to accuracy measured by difference scores, correlations are (given normally distributed variables) statistically independent of mean levels of both prediction and recall. Instead, they depend on agreement of relative ordering of individuals (as indexed by deviations from the means of the two variables). Previous studies have shown age equivalence in correlations of initial predictions made before study of test materials with subsequent recall performance (Hertzog et al., 1990; Hertzog et al., 1994). These correlations are typically relatively low for the initial global prediction (ranging from 0 to .30 across different experiments) but increase in magnitude equally for younger and older adults with additional recall trials (see Hertzog et al., 1990; Hertzog et al., 1994; Lachman et al., 1987).

Hertzog et al. (1994) revealed a second disparity between indicators of prediction accuracy based on difference scores and those based on correlations. In their phased prediction task, global predictions were made both before and after study of the test materials. Whereas accuracy based on difference scores did not change significantly from before to after study, correlations of predictions with recall increased significantly. Both age groups manifested an increase in correlations from before to after study, but younger adults produced a larger increase (their correlations, aggregated over two experiments, increased from .15 to .49, whereas the older adults' correlations increased from .13 to .30). The disparate pattern of changes from before to after study between the difference scores and the correlations suggests that psychological processes associated with scaling the predictions, such as the anchoring effect, may adversely affect prediction accuracy as measured by difference scores.

The higher correlation between after-study global prediction and recall represents a beneficial effect of study on prediction accuracy. This outcome implies that monitoring learning during study may increase the accuracy of global memory predictions, although this hypothesis has yet to be tested directly. Likewise, the age difference in the magnitude of correlational increase may indicate an age difference in the use of metacognitive monitoring to adjust global performance predictions (see Hertzog et al., 1994). Another goal of the current investigation was to replicate this differential upgrade of accuracy between younger adults and older adults.

Accuracy of Item-by-Item Judgments of Learning

The experimental literature has emphasized item-by-item predictions in methods for measuring metacognitive monitoring at different phases of learning and remembering new materials (Nelson & Narens, 1990). Recently, a great deal of attention has been paid to understanding the bases and accuracy of judgments of learning (JOLs), which are item-by-item predictions of the likelihood of subsequent recollection of recently studied items. For example, Mazzoni and Nelson (1995) recently demonstrated that JOLs are sensitive to aspects of encoding behavior that are independent of the level of recall, suggesting that JOLs reflect monitoring of encoding processes.

In the developmental literature, several studies have examined the effect of age on JOLs. A common method for assessing JOL accuracy in the aging literature has been to compare the magnitudes of JOLs for items that are and are not subsequently recalled. A typical finding is that for both younger and older adults, JOLs are greater for items that are subsequently recalled (e.g., Bieman-Copland & Charness, 1994; Lovelace & Marsh, 1985; Rabinowitz, Ackerman, Craik, & Hinchley, 1982; Shaw & Craik, 1989). Lovelace and Marsh (1985) also demonstrated that for both younger and older adults, the probability of recall increased monotonically as a function of increases in JOLs.

Although such findings have been interpreted as showing age equivalence in the accuracy of adults' on-line monitoring of memory (e.g., Rabinowitz et al., 1982), this conclusion is not definitive. First, comparisons of mean predictions between items that are recalled and items that are not recalled are an unsatisfactory quantitative measure of predictive accuracy. Nelson (1984) showed that an analogous method for assessing the predictive accuracy of feeling-of-knowing judgments (Hart's difference score) is dependent on the overall level of subsequent recall performance (see Figure 1 in Nelson, 1984). In contrast, a Goodmann—Kruskal gamma correlation between an individual's JOLs for each item and the binary performance outcome (success or failure to remember the item at test) provides a quantitative measure of the degree to which that individual accurately predicted the relative likelihood of remembering different items. Gamma is also logically independent of the overall level of memory performance (Nelson, 1984). Moreover, gamma may be more sensitive to small gradations of JOL accuracy than is the mean difference between items that are recalled and those that are not. Unfortunately, analyses based on gamma correlations have not been used to examine age differences in JOL accuracy.

A second reason for not considering the inference of age equivalence in JOL accuracy to be a definitive conclusion involves the type of JOL studied. In all of the aforementioned investigations of aging and JOL accuracy, individuals made a JOL for an item immediately after the item had been presented for study. Immediate JOLs typically produce gammas that are significantly greater than 0 (indicating above-chance accuracy
in the judgments), with the magnitude varying as a function of a number of experimental factors. However, the accuracy of immediate JOLs is generally closer to 0 (which would indicate no association of predictions and subsequent item recall) than to 1 (which would indicate perfect ordinal association of the two variables). For example, a series of studies on JOLs for paired-associate learning by undergraduate students indicated a mean (across individuals) gamma correlation between JOLs and subsequent recall performance of approximately .40 (Dunlosky & Nelson, 1992, 1994; Nelson & Dunlosky, 1991). In these studies, immediate JOLs were compared with JOLs delayed about 30 s after an item had been studied (the participant studied other pairs during the delay). For example, if DOG–SPOON had been studied, DOG–? would be used as a prompt for a JOL either immediately after study or with a delay. When the students’ JOLs were delayed, the gamma correlations for these stimulus-alone delayed JOLs were nearly perfect (approximately .90 or greater). Dunlosky and Nelson (1992) showed that this dramatic increase in JOL accuracy, which they called the delayed-JOL effect, did not occur when both associates were used to prompt a delayed JOL (i.e., DOG–SPOON was presented when a delayed JOL was made). Instead, delayed JOLs for stimulus–response pairs were equivalent in accuracy to immediate JOLs.

A leading explanation for the close-to-perfect accuracy of stimulus-alone delayed JOLs is that they are based on monitoring the outcome of a covert retrieval attempt cued by the stimulus word (Nelson & Dunlosky, 1991; see Schwartz, 1994, for a review). That is, when presented with the cue after some delay (i.e., DOG–?), the individual first monitors retrieval of information about the response word of the pair from memory and then maps that information onto the scale imposed by the JOL (Dunlosky & Nelson, 1994). Delaying JOLs yields high predictive accuracy because information retrieved at the time of the delayed JOL is highly correlated with successful retrieval at test. Providing both the stimulus and response elements of the pair at the time of the delayed JOL may bypass the need for a covert retrieval of the response word from memory, reducing the accuracy of the delayed JOL to that of the immediate JOL (for other possible explanations, see Dunlosky & Nelson, in press).

According to this view, immediate and delayed JOLs differentially tap into multiple aspects of memory monitoring. Immediate JOLs may be principally affected by individuals’ monitoring of the encoding process, and hence their accuracy is limited because of the reduced diagnosticity of subjective judgments about encoding for subsequent recall (Mazzoni & Nelson, 1995). In contrast, stimulus-alone delayed JOLs may be principally affected by retrieval of the responses of items from memory. In theory, then, one may observe age differences in one or both kinds of JOL, depending on how aging affects information available to monitoring at the time of encoding and retrieval. For example, if older adults are less likely to base delayed JOLs on information about the response word retrieved from memory, they may show a reduced delayed-JOL effect.

Overview of the Present Experiments

One major goal of the present study was to use recent advances in the experimental literature on JOLs to examine more fully age differences in memory monitoring. Would age differences in memory monitoring emerge when gamma was used as the index for JOL accuracy? Would we find differential patterns of age differences, depending on the timing and nature of the JOL? A second major goal was to relate and integrate findings on age differences in the accuracy of global and item-by-item predictions, which has typically not been addressed in the experimental literature (but see Mazzoni & Nelson, 1995, for an initial step in this direction). Would we replicate the correlational upgrade for global predictions found by Hertzog et al. (1994)? Would we replicate age differences in the magnitude of this upgrading, and if so, would this effect be associated with age differences in the accuracy of immediate or delayed JOLs? Demonstrating concurrent age differences in JOL accuracy and reduced magnitude of correlational upgrading for global predictions would provide support for the hypothesis that deficient memory monitoring is a source of age differences in global prediction accuracy. Finally, is the effect of study on global prediction accuracy attributable to the benefits of monitoring learning? If it is, individuals’ after-study global prediction should have a higher correlation with JOLs than their before-study prediction.

In summary, the present three experiments were designed to accomplish a variety of interrelated goals: (a) to demonstrate that the accuracy of global judgments is at least partially dependent on the level of recall performance, (b) to replicate the age-differential upgrade of the correlational accuracy of global predictions, (c) to evaluate the possibility of age-related differences in the accuracy of immediate and delayed JOLs, and (d) to determine whether the delayed-JOL effect occurs for older adults. To accomplish these goals, we used both kinds of judgments (global and item by item) in a single experimental task, integrating methods developed in earlier investigations of global predictions (e.g., Hertzog et al., 1994) and in investigations of immediate and delayed JOLs (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991).

Experiment 1

We used the procedure and materials used by Nelson and Dunlosky (1991) to extend the delayed-JOL effect to older adults. We also sought to determine the degree to which the kind of cue given for a JOL modulates any age-related differences in JOL accuracy. Some individuals made JOLs that were cued by the stimulus alone, whereas others made JOLs that were cued by the stimulus–response pair. To foreshadow, use of the procedures from Nelson and Dunlosky closely replicated their previous findings, including recall performance near 50% for younger adults. Mean recall performance of older adults was considerably lower, and several older adults recalled few (or even no) words. Although the relatively poor performance of the older adults had several disadvantages, one fortuitous advantage concerned implications for demonstrating recall-level sensitivity of difference-score measures of global prediction accuracy.

Method

Participants

Thirty-four community-dwelling older adults from the Atlanta, Georgia, metropolitan area received $10 each for participation. The mean
age of these participants was 71.1 years (SD = 4.3), they had an average of 15.4 years of education (SD = 3.0), and they reported being in good health. They reported taking an average of 2.4 medications (SD = 2.4). Their mean vocabulary score, as measured by Advanced Vocabulary Test V4 from the Educational Testing Service Reference Kit (Ekstrom, French, Harman, & Dermen, 1976), was 21.9 (SD = 8.7) out of a possible 36 items.

Sixty younger adult participants, recruited from the Georgia Institute of Technology psychology participant pool, received extra course credit. They averaged 20.2 years of age (SD = 1.9), had thus far attained 13.3 years of education (SD = 1.3), and reported being in good health. They were taking an average of 0.40 medications (SD = 1.0), and their mean vocabulary score was 16.4 (SD = 4.1).

Older adults took significantly more medications,\(^2\) F(1, 92) = 30.60, MSE = 2.79, p < .001, and had significantly higher vocabulary scores, F(1, 92) = 17.07, MSE = 38.17, p < .001, than did younger adults.

**Apparatus and Materials**

The experimental task was controlled by Apple IIe microcomputers fitted with Digitry CTS cards and button boxes that enabled the measurement of response latencies to the nearest millisecond.

Items for the paired-associate recall task were 76 noun–noun pairs. The nouns were high in concreteness (above 6.0) by the Paivio, Yuille, and Madigan, 1968, norms for those items listed and were high in frequency of usage in the language (M = 48.5 per million, range - 29–83) as measured by the Kucera and Francis (1965) word norms. Paired words were unrelated but were roughly matched in frequency and concreteness. Word pairs were divided into a block of 10 pairs that were used for practice, 6 buffer pairs that appeared at the beginning of the first experimental block, and 60 pairs used for experimental trials (see the Appendix for the stimuli).

**Design**

The experimental design consisted of a 2 (age: young vs. old) × 2 (JOL type: cued by the stimulus word alone vs. cued by the stimulus–response pair) × 2 (time of JOL: immediate vs. delayed) mixed design. Age and JOL type were between-subject factors, and time of JOL was a within-subject factor. The basic experimental design and procedure closely followed those of Dunlosky and Nelson (1992). Participants studied paired associates for later recall and made both immediate and delayed JOLs within each of two blocks of trials. Randomly selected halves of the 30 pairs in each block were assigned to either an immediate-JOL trial or a delayed-JOL trial. If the item was to receive an immediate JOL, the cue for the JOL appeared again immediately after study and the participant made the rating. If a pair of words was assigned to a delayed-JOL trial, at least 10 other study/JOL trials intervened before the stimulus word was presented to cue the judgment. In this manner, each item in the block was studied and received a JOL within the same block.

For recall, items were separated into blocks of 10 according to the order in which they had been studied. The first 10 items studied were then presented in a new random order for recall, followed by the second 10 items studied, and so forth.

**Procedure**

Individuals participated in small groups of 1 to 5 in an experimental session that lasted 60–90 min. After reporting some demographic information, participants filled out two questionnaires concerning their memory performance in everyday situations. Both the short form of the Memory Functioning Questionnaire (Gilewski, Zelinski, & Schaie, 1990) and the long form of the Metamemory in Adulthood Instrument (Dixon, Hultsch, & Hertzog, 1988) were administered. Data from these questionnaires were generally consistent with previous findings regarding age differences in scale means and did not add materially to the study questions. Hence they are not reported.

Participants were then seated in a partitioned booth in front of a computer workstation. They were instructed both verbally by the experimenter and through written instructions on the computer screen how to do the memory task and item-by-item memory prediction task. In particular, participants were informed that they would be studying pairs of items for a later recall test and would be making predictions about the likelihood that each item would be recalled. These JOLs were requested in the following format: "How confident are you that in about ten minutes you will be able to recall the second word of the item when prompted with the first (0 = definitely won’t recall, 20 = 20% sure, 40 = 40% sure, 60 = 60% sure, 80 = 80% sure, 100 = definitely will recall?)" A practice block of 10 items was given, for which participants made immediate JOLs for each item and the experimenter answered any questions.

After completing the practice block but before studying the items in the experimental blocks, participants wrote down in their recall booklets the number of responses out of 60 they predicted they would recall at the final paired-associate recall test. After making this first global prediction, participants began the item-by-item learning/judgment task. Each item appeared on the computer screen for 10 s. Participants were instructed to study the item for as long as it remained on the computer screen so as to be able to recall on a later test the second member of the pair when it was prompted with the first member of the pair: For participants whose JOLs were cued by the stimulus alone, the stimulus word appeared followed by a question mark; for example, if OCEAN–TREE had been studied, OCEAN–? would be the cue. For participants in the stimulus–response condition, both words appeared on the screen (i.e., OCEAN–TREE). Participants made a JOL by pressing one of six keys on a button box that corresponded to the memory prediction that the individual was making with the buttons labeled from left to right as 0%, 20%, 40%, 60%, 80%, and 100%. Participants had up to 20 s to make their rating for each item. If they exceeded 20 s, an error message appeared, and the trial was considered an error trial and was not included in the analyses. On each trial, both the rating made and the latency to make the rating were recorded. The next item for study appeared 2 s after the JOL was made. If the studied item was slated for a delayed JOL, no rating was made for that item until the end of the block and the next item appeared for study. After all items had been studied and all immediate JOLs had been made, the 15 items in the block that were slated for delayed judgments reappeared one at a time and participants made a JOL for each of those items. After the last delayed JOL in the first block, the whole process of studying items and making immediate and delayed JOLs was repeated for the 30 pairs presented in the second block.

After they had studied and rated all of the items, the participants were asked to again predict the number of items (out of 60) that they would recall in the paired-associate recall test. After this second global prediction, the recall task was administered. The first member of each item appeared on the computer screen followed by a question mark. Participants were instructed to write down the second word of the pair, if possible, in the recall booklet. Recall trials were paced by participants. They pressed a button on the button box to go on to the next item. After the recall task, participants were asked to make a global postdiction by writing on the next page of the recall booklet the total number of paired associates they had correctly recalled.

After the recall task, participants were given a questionnaire that asked what strategy (if any) they thought that they had used to learn the pairs of words. The first question asked them to describe the strategy they thought they had used. The second question described some strategies—

---

\(^2\) The number of medications failed to correlate significantly with any of our performance measures in any of the three experiments.
forming a sentence containing the two items of the pair, using interactive visual imagery, or using a combination of the two—and asked if any of them had been used. Finally, participants were given the Advanced Vocabulary Test V4 with a 4-min time limit.

**Results**

Ten of the 34 older adults who were tested performed poorly on paired-associate recall, remembering less than 5% correctly. Computation of item-by-item gamma correlations requires sufficient numbers of correct and incorrect responses for each participant to generate a stable marginal distribution of JOL ratings for each outcome. Hence, in Experiment 1 as well as in Experiments 2 and 3, we excluded from analysis any participant who performed at or below 5% correct or at or above 95% correct. Therefore, data from 24 older adults and 60 younger adults were included in the analyses.

**Overall Memory and Metamemory Performance**

**Mean level of recall.** The mean proportions of response words recalled by young and old adults in the JOL type and time of JOL conditions are shown in Table 2. A 2 (age) × 2 (JOL type) × 2 (time of JOL) mixed factor analysis of variance (ANOVA) revealed a main effect of age, with younger adults recalling a higher proportion of associated response words than did older adults (M = .49 vs. .24), F(1, 78) = 29.61, MSE = 0.07, p < .0001. A main effect of time of JOL, F(1, 78) = 44.60, MSE = 0.01, p < .0001, indicated that recall was higher for items receiving delayed JOLs (M = .48) than for items receiving immediate JOLs (M = .36). In addition, the interaction of JOL type and time of JOL was statistically significant, F(1, 78) = 11.34, MSE = 0.01, p < .01. The improvement in recall due to delaying JOLs was larger for participants who received stimulus–response cues than for those who received the stimulus-alone cues. There were no significant interactions involving age (all Fs < 2.25).

**Mean global predictions.** Although participants' predictions and postdictions were given as the number of words that would be recalled, comparison across experiments was facilitated by rescaling the values into estimated proportions of words recalled. Table 3 presents the mean proportion of predicted and postdicted recall by age group and JOL type. In general, older participants predicted lower recall than did younger participants (M = .40 vs. .49), F(1, 76) = 18.19, MSE = 0.06, p < .0001. There was a main effect of prediction phase, F(2, 152) = 8.56, MSE = 0.02, p < .001, which was qualified by a significant Age × Prediction Phase interaction, F(2, 152) = 4.69, MSE = 0.02, p < .05. Predictions made after study were lower than those made before study, more so for older adults than for younger adults. For younger adults, the postdiction was higher than the second prediction, whereas for older adults the postdiction was lower than the second prediction. In general, this pattern of age differences in predictions roughly corresponded to age differences in recall (cf. accuracy analyses below).

**Mean JOL ratings.** Mean JOL ratings for young and old adults by JOL type and time of JOL are reported in Table 4. On average, younger adults predicted a higher proportion correct than did older adults (M = .53 vs. .42), F(1, 78) = 10.06, MSE = 0.05, p < .01. This finding was qualified by two interactions. First, a significant Age × JOL Type interaction was detected, F(1, 78) = 5.59, MSE = 0.05, p < .05. Older individuals manifested little difference in the mean proportion predicted by JOL type, whereas younger individuals cued by the stimulus alone predicted a higher proportion correct than did younger individuals cued by both items of the pair. Second, a significant disordinal Age × Time of JOL interaction, F(1, 78) = 10.91, MSE = 0.10, p < .01, indicated that mean immediate JOLs differed relatively little between the two age groups, but mean delayed JOLs conformed more closely to actual probability of recall (low for older adults and higher for younger adults).

**Accuracy of Global Predictions**

The accuracy of global predictions was assessed using differences between predictions and performance as well as Pearson product–moment correlations of prediction and recall. In each age group, one individual omitted a global prediction. Those 2 participants were excluded from this analysis, leaving 23 older adults and 59 younger adults.

**Difference scores.** Table 5 contains the proportion predicted minus the proportion recalled. Younger adults were more accurate in making global predictions than were older adults (M = −.03 vs. −.10), F(1, 79) = 8.13, MSE = 0.08, p < .01, with older adults predicting more than they recalled and younger adults recalling slightly more than they predicted. There was

---

3 Excluding these participants did not change the magnitude of other performance measures (i.e., global predictions). In Experiment 1, no one performed at or above 95% correct.
Table 3
Means and Standard Deviations of Proportion of Predicted Recall at Each of Three Prediction Phases

<table>
<thead>
<tr>
<th>Age group and cue condition</th>
<th>Before-study prediction</th>
<th>After-study prediction</th>
<th>Postdiction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.40</td>
<td>.24</td>
<td>.25</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.40</td>
<td>.17</td>
<td>.32</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.49</td>
<td>.14</td>
<td>.44</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.49</td>
<td>.17</td>
<td>.40</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.37</td>
<td>.22</td>
<td>.37</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.35</td>
<td>.23</td>
<td>.41</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.43</td>
<td>.18</td>
<td>.52</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.48</td>
<td>.22</td>
<td>.52</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.62</td>
<td>.19</td>
<td>.47</td>
</tr>
<tr>
<td>Young</td>
<td>.58</td>
<td>.19</td>
<td>.58</td>
</tr>
</tbody>
</table>

Note. Only stimulus-alone cues were given in Experiment 3.

Table 4
Means and Standard Deviations of Judgment of Learning Ratings

<table>
<thead>
<tr>
<th>Age group and cue condition</th>
<th>Time of judgment of learning</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Immediate</td>
<td>SD</td>
<td>Delayed</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.44</td>
<td>.24</td>
<td>.35</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.47</td>
<td>.13</td>
<td>.43</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.56</td>
<td>.12</td>
<td>.63</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.46</td>
<td>.13</td>
<td>.48</td>
<td>.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.43</td>
<td>.29</td>
<td>.49</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.44</td>
<td>.22</td>
<td>.55</td>
<td>.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.60</td>
<td>.21</td>
<td>.70</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.48</td>
<td>.18</td>
<td>.51</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td>.57</td>
<td>.19</td>
<td>.60</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td>.61</td>
<td>.18</td>
<td>.72</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Only stimulus-alone cues were given in Experiment 3.

also a main effect of prediction phase, \( F(2, 158) = 9.53, MSE = 0.02, p < .001 \), and a significant Age × Prediction Phase interaction, \( F(2, 158) = 5.84, MSE = 0.02, p < .01 \). For older adults, the difference between prediction and recall decreased across prediction phase. For younger adults, a more complicated pattern emerged. Their first prediction was relatively accurate, the second prediction underpredicted recall, and the postdiction was as accurate as the first prediction.

Correlations. In order to obtain better statistical power, the correlational analysis of global prediction accuracy was collapsed across JOL type. Table 6 contains the Pearson correlations between predictions and recall for young and old adults. The first global prediction, before study of the word pairs, did not correlate significantly with recall for either young or old adults. The negative sample correlation between before-study prediction and recall for older adults, although not reliably different from 0, was a surprise. Inspection of the data revealed that 3 older adults with relatively low recall levels had initially predicted perfect performance (60 items recalled). Predicting the maximum possible performance could reflect goal-setting behavior instead of an attempt to estimate future performance accurately. When these individuals' data were eliminated, the correlation of the older adults' initial prediction with recall was \( -0.05 \).

The second (after-study) global prediction correlated significantly with paired-associate recall only for the young adults, although both age groups' correlations became more positive after study. The correlation between the postdiction and recall was statistically significant for both age groups. A significant
Table 5
Mean Differences and Standard Deviations Between Proportion Predicted and Proportion Recalled at Each Prediction Phase

<table>
<thead>
<tr>
<th>Age group and cue condition</th>
<th>Before-study prediction - recall</th>
<th>After-study prediction - recall</th>
<th>Postdiction - recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>0.20</td>
<td>0.30</td>
<td>0.05</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>0.17</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>-0.01</td>
<td>0.24</td>
<td>-0.07</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>0.03</td>
<td>0.24</td>
<td>-0.07</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>-0.09</td>
<td>0.29</td>
<td>-0.08</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>-0.20</td>
<td>0.25</td>
<td>-0.14</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>-0.24</td>
<td>0.25</td>
<td>-0.15</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>-0.21</td>
<td>0.24</td>
<td>-0.17</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction 1</td>
<td>-.06</td>
<td>.23</td>
<td>.04</td>
</tr>
<tr>
<td>Prediction 2</td>
<td>.10</td>
<td>.68**</td>
<td>.55***</td>
</tr>
<tr>
<td>Postdiction</td>
<td>.68**</td>
<td>.95***</td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction 1</td>
<td>-.06</td>
<td>.23</td>
<td>.04</td>
</tr>
<tr>
<td>Prediction 2</td>
<td>.39</td>
<td>.66***</td>
<td>.53***</td>
</tr>
<tr>
<td>Postdiction</td>
<td>.92***</td>
<td>.92***</td>
<td></td>
</tr>
</tbody>
</table>

Note. Only stimulus-alone cues were given in Experiment 3.

likelihood ratio chi-square test (see Hertzog et al., 1994) showed that the younger adults’ after-study prediction correlated more highly than their before-study prediction with recall, \( \chi^2(1, N = 59) = 17.92, p < .001 \). The same effect was also significant in the older group, \( \chi^2(1, N = 23) = 4.11, p < .05 \). Younger adults had a significantly higher correlation between after-study prediction and performance than did older adults (Fisher’s \( r \) to \( z = 2.01, p < .05 \), one-tailed).

Accuracy of Item-by-Item Predictions

Calibration curves. The accuracy of item-by-item predictions has been evaluated in some studies by visual inspection of calibration curves—plots of the mean proportion of recalled items as a function of each JOL rating scale point (e.g., Lovelace & Marsh, 1985). The closer each age group’s curve is to the main diagonal, the better calibrated are different JOL ratings to the observed mean probability of item recall. For descriptive purposes and for continuity with earlier literature, these calibration curves are provided in Figure 1 for the stimulus-alone and stimulus–response conditions for immediate and delayed JOLs. Delayed JOLs led to more accurate item-by-item judgments by both young and old adults. In addition, calibration appeared to be less accurate for older than for younger adults.

Gammas. A gamma correlation was computed for each participant (and for each within-subject condition) and then treated as a dependent variable measuring the relative accuracy of item-by-item predictions. Table 7 presents mean gamma correlations for young and old adults as a function of JOL type and time of JOL. As expected, there was a substantial delayed-JOL effect, as evidenced by the significant interaction between JOL type and time of JOL, \( F(1, 78) = 22.58, MSE = 0.09, p < .001 \). There was little change in item-by-item accuracy between immediate and delayed JOLs when individuals were cued by both the stimulus and response (\( M = .46 \) and .43, respectively). However, a substantial improvement in accuracy between imme-
diate and delayed time of JOL was obtained when individuals were cued with just the stimulus item of a pair (M = .33 and .88, respectively). Although the variability for the older adults was large, it may be seen from Table 6 that the delayed-JOL effect was virtually the same for older and younger individuals. There was no main effect of age or any interaction involving age (all Fs < 1).

Relations Between JOLs and Global Predictions

To examine relations between the two types of predictions, we computed Pearson correlations among each participant’s mean JOL (both immediate and delayed), global predictions and postdictions, and recall. Table 8 provides the correlations for both age groups and JOL types. The small sample size for older adults limits the usefulness of these data, but they are provided for making comparisons across experiments. In the stimulus-alone condition, younger adults’ delayed JOLs correlated highly with recall and with the after-study global prediction, whereas their mean immediate JOLs had virtually no relation to their predictions or performance. In the stimulus-response condition, both types of JOL correlated significantly with performance. However, mean JOLs tended to correlate more highly with the global prediction made after study than with the prediction made before study. Note that the same pattern occurred in the correlations for older adults.

Because separating recall into recall for immediate and delayed trials did not alter the findings, correlations were computed with total items recalled.
Means and Standard Deviations of Gamma Correlations

<table>
<thead>
<tr>
<th>Age group and cue condition</th>
<th>Time of judgment of learning</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate</td>
<td>Delayed</td>
<td>Immediate</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.44</td>
<td>.58</td>
<td>.88</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.50</td>
<td>.36</td>
<td>.39</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.29</td>
<td>.26</td>
<td>.88</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.44</td>
<td>.30</td>
<td>.44</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.50</td>
<td>.38</td>
<td>.83</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.33</td>
<td>.40</td>
<td>.59</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus alone</td>
<td>.51</td>
<td>.33</td>
<td>.82</td>
</tr>
<tr>
<td>Stimulus–response</td>
<td>.56</td>
<td>.29</td>
<td>.55</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.49</td>
<td>.37</td>
<td>.78</td>
</tr>
<tr>
<td>Young</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.55</td>
<td>.24</td>
<td>.82</td>
</tr>
</tbody>
</table>

*Note.* Only stimulus-alone cues were given in Experiment 3.

Discussion

We replicated the delayed-JOL effect—higher relation of delayed JOLs, relative to immediate JOLs, to probability of item recall—both in calibration curves and in gamma correlations (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991). We also demonstrated a new distinction between immediate and delayed JOLs, namely, different patterns of correlations with global predictions. In the stimulus-alone condition, individual differences in younger adults’ mean delayed JOLs correlated highly with individual differences in both recall and after-study global prediction. In the stimulus–response condition, younger adults’ mean immediate and delayed JOLs correlated with recall and with both before- and after-study global predictions. This increase in correlations of JOLs with after-study prediction suggests that the second global prediction was influenced by outcomes of the memory monitoring process that were tapped by JOLs.

With respect to age differences in memory monitoring, Experiment 1 produced several intriguing findings. Most important, there was little indication of an age difference in predictive accuracy as measured by gamma coefficients. Like younger adults, older adults showed a robust delayed-JOL effect. The magnitudes of the gamma correlations were comparable for the two age groups in all experimental conditions, suggesting that monitoring of encoding and retrieval processes during paired-associate learning was not adversely affected by age. Second, this pattern of rough equivalence in gamma correlations occurred in data showing age differences in calibration curves. This finding indicates that Lovelace and Marsh’s (1985) results regarding age differences in calibration should not be taken as evidence of poor memory monitoring by older adults but may instead reflect age differences in how monitoring is transformed into a JOL rating. Third, the experiment replicated the upgrading effect identified by Hertzog et al. (1994) for correlations of free recall with global predictions made before and after study, demonstrating that this phenomenon extends to a paired-associate recall task.

The salient limitation of Experiment 1 was the unexpected poor performance of older adults in the standard paired-associate task used by Nelson and Dunlosky (1991). A significant number of older individuals performed near the floor in recall. The older adults’ self-reported encoding strategies provide one possible explanation of this effect. Thirty-two percent of them, including 5 of the 10 older individuals who performed very poorly on the task (less than 10% recall), reported that they had not used any specific strategy to learn the paired associates. None of the younger individuals reported that they had not used some strategy to encode the pairs. This outcome is consistent with other studies (e.g., Hulicka & Grossman, 1967) indicating that older adults are less likely to use mediational strategies spontaneously in paired-associate learning (see Kausler, 1994, for a review). Given the little intrinsic association between the words in each of our pairs, failure to use a mediational strategy would suppress performance.

The poor performance by some older adults probably accounts for the limited global postdiction accuracy found for older adults. Other studies (Devolder et al., 1990; Hertzog et al., 1994) suggest age equivalence in postdiction accuracy. The correlation reported in Table 6 is lower than analogous correlations between global postdictions and free recall in the .80 to .95 range found by Hertzog et al. (1994). The postdiction–recall correlation in Experiment 1 may have been attenuated because of restriction of range on performance. This restriction-of-range problem limits the inferences that may be drawn about age differences in memory monitoring on the basis of Experiment 1. A probable exception is the inference about comparable memory monitoring accuracy based on the gamma statistic, which is known to be a robust estimator of monitoring accuracy across wide ranges of overall task performance (Nelson, 1984).

Fortuitously, the effect of using younger and older adults in an experimental task that had been designed and piloted on younger adults produced data about possible level sensitivity of global and item-by-item predictions. Note that the pattern of prediction accuracy scores mirrors the result found in many other studies (Lovelace, 1990), namely, that younger adults’ mean global predictions are much closer to their mean recall than older adults’ predictions are to their recall. With reference to the summary presented in Table 1, this effect could be due to younger and older adults’ predictions being anchored near 50% performance, with younger adults performing at about that level but older adults performing below it. Experiment 1, therefore, is consistent with the argument that age differences in prediction accuracy, as measured by difference scores, are attributable at least in part to anchoring.

Experiment 2

To provide a better evaluation of age differences in memory monitoring, we designed Experiment 2 to ensure that older
Table 8
Correlations of Mean Immediate and Delayed Judgments of Learning (JOLs) With Recall, Global Predictions, and Global Postdiction

<table>
<thead>
<tr>
<th>JOL condition</th>
<th>Stimulus alone</th>
<th>Stimulus–response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate JOL</td>
<td>Delayed JOL</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>Recall</td>
</tr>
<tr>
<td>Old</td>
<td>11</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>.43</td>
</tr>
<tr>
<td>Young</td>
<td>29</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>.80**</td>
</tr>
<tr>
<td>Old</td>
<td>33</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>.77**</td>
</tr>
<tr>
<td>Young</td>
<td>28</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>.38</td>
</tr>
<tr>
<td>Old</td>
<td>28</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>.68**</td>
</tr>
<tr>
<td>Young</td>
<td>29</td>
<td>.27</td>
</tr>
</tbody>
</table>

Note. Only stimulus-alone cues were given in Experiment 3.
*p < .01. **p < .001.

adults’ performance would be well above floor. To improve the overall levels of recall, we included an initial study trial of the pairs in which no item-by-item predictions were made (e.g., Mazzoni & Nelson, 1995), and participants were explicitly instructed to use a strategy, such as visual imagery or formation of a sentence with the items, to learn the word pairs. Increasing the level of recall should remove the restriction-of-range effect, and therefore we expected improvement in older adults’ global prediction accuracy and the calibration of their JOLs.

Method

Participants

Seventy-eight community-dwelling older adults from the Atlanta metropolitan area received $10 each for participation. They were an average of 69.4 years old (SD = 5.0), had 14.7 years of education (SD = 3.0), and reported being in good health. They reported taking an average of 1.7 medications (SD = 1.8). Their mean vocabulary score was 22.2 (SD = 7.7) out of a possible 36 items correct.
Eighty younger adults, recruited from the Georgia Institute of Technology psychology participant pool, received extra course credit for participation. Their mean age was 21.0 years (SD = 2.6), they had attained 14.0 years of education (SD = 1.4), and they reported being in good health. They reported taking an average of 0.31 medications (SD = 0.69). Their mean vocabulary score was 15.8 (SD = 4.2).

The older adults were taking significantly more medications than the younger adults, F(1, 156) = 38.64, MSE = 1.18, p < .001, and had significantly higher vocabulary scores than did the younger adults, F(1, 156) = 42.98, MSE = 38.6, p < .001.

Procedure

Participants completed the same questionnaires used in Experiment 1 and in the same order as in that experiment. The procedure for memory prediction was identical to the procedure in Experiment 1, except for the following two changes. First, every participant was exposed twice to all of the paired associates that would later appear on the recall test. In the first exposure, participants studied each pair for 10 s. No JOLs were made during this trial. Participants also did not practice making JOLs until after this trial was completed. The second exposure to the pairs was identical in format to the procedure used in Experiment 1, including practice making JOLs, global predictions, and the final recall task.

Second, participants were explicitly instructed to try to use a strategy to help them link the stimulus and response words of the pair. They were told that two ways this could be accomplished were to make sentences that contained both words of the pair and to generate a mental image in which the two words of the pair interacted. An example was made of the pair DOG-TABLE. "If you saw the pair DOG-TABLE you could form the sentence, 'The dog was standing on the table,' or you could picture a dog standing on a table. When you were later given the word DOG as a cue to recall the pair you could recall your sentence or mental image to aid in producing the word TABLE.'"

Results

As in Experiment 1, we excluded from analyses data from individuals who performed above 95% or below 5% on the recall task. Eleven older adults (7 of whom were at ceiling and 4 of whom were at floor) and 26 younger adults (1 individual at floor and 25 at ceiling) were excluded on the basis of these criteria. The number of participants remaining was 67 older adults and 54 younger adults.

Overall Memory and Metamemory Performance

Mean level of recall. The mean proportion of response words recalled by each age group in each experimental condition is presented in Table 2. A 2 (age) × 2 (JOL type) × 2 (time of JOL) ANOVA demonstrated that older adults recalled less than younger adults (M = .56 vs. .68), F(1, 117) = 14.55, MSE = 0.13, p < .001, and that recall was higher in the delayed JOL condition (M = .62) than in the immediate JOL condition (M = .54), F(1, 117) = 44.86, MSE = 0.01, p < .0001. In addition, there was a significant interaction of time of JOL and cue condition, F(1, 117) = 8.30, MSE = 0.01, p < .01, such that for participants cued by both the stimulus and response items of the pair there was better recall performance in the delayed JOL condition (M = .66) than in the immediate JOL condition (M = .56), but for participants cued by the stimulus alone there was little difference between recall levels in the immediate (M = .53) and delayed (M = .57) JOL conditions. There were no significant interactions with age (all Fs < 1). This pattern replicates effects obtained in Experiment 1.

Mean global predictions. Table 3 reports the mean proportion predicted at each phase of global prediction—before study, after study, and after recall. Older adults predicted lower recall (M = .41) than did younger adults (M = .55), F(1, 117) = 12.78, MSE = 0.12, p < .001. Predictions increased from before study to after study to after recall, F(2, 234) = 47.65, MSE = 0.02, p < .0001. There was also a marginal Age × Prediction Phase interaction, F(2, 234) = 2.92, MSE = 0.02, p = .056, which indicated that younger adults increased their predictions across phases more than older adults did.

Mean JOL ratings. The mean JOL ratings are reported in Table 4. The ANOVA revealed a significant main effect of age, F(1, 110) = 5.58, MSE = 0.10, p < .05, with younger adults predicting higher recall (M = .58) than older adults (M = .48), and a main effect of time of JOL, F(1, 110) = 21.14, MSE = 0.002, p < .0001, with delayed JOLs producing higher item-by-item predictions (M = .56) than immediate JOLs (M = .49). There was also a significant Age × JOL Type interaction, F(1, 110) = 5.25, MSE = 0.10, p < .05. Younger adults showed a greater difference between stimulus-alone and stimulus–response JOLs, with higher mean probability of recall reported in the stimulus-alone condition.

Accuracy of Global Predictions

Difference scores. Table 5 contains the proportion predicted minus the proportion recalled for older and younger adults in both cue conditions. There was no significant difference between younger and older adults in the differences between predictions and recall or between the cue types (all Fs < 2.17). There was, however, a significant main effect of prediction phase, F(2, 236) = 47.71, MSE = 0.02, p < .0001, such that predictions became more accurate across phases. 5

Correlations. The Pearson correlations between global predictions and paired-associate recall are reported in Table 6. Both groups showed an equivalent, significant correlation of the first prediction with recall. Note also that this correlation was significantly different from the analogous correlation in Experiment 1 for older adults (see Table 6; Fisher's r to z = 2.63, p < .01, one-tailed). The same r to z test for the correlational difference in younger adults approached but did not achieve statistical significance (z = 1.24, p = .11, one-tailed). 6

1 The identical analysis including participants who had been excluded on the basis of ceiling or floor performance produced a different pattern. That is, a significant Age × Prediction Phase interaction was found such that younger adults were less accurate in both before- and after-study predictions (difference scores of −17.2 and −13.1, respectively) than were older adults (−7.6 and −5.3, respectively), but both groups were equally accurate at postdiction (young = −1.11 and old = −7.6). This finding is not inconsistent with the conclusions that we draw from the more selective sample.

5 The post hoc estimate of power to detect the z-transformed difference in younger adults' correlations of .24, given the associated standard error of .194, was only .34. Hence, the failure to reject the null hypothesis may well be a function of limited power.
The likelihood ratio chi-square test rejected the null hypothesis of equal correlations of before- and after-study predictions with recall, $\chi^2(2, N = 121) = 25.32, p < .001$. After-study predictions correlated higher with recall than did before-study predictions. The most interesting aspect of the data from Experiment 2 is the lack of an age difference in the upgrading of correlations of global predictions with recall (see Table 6). If anything, older adults showed a greater increase in correlation from before to after study than did younger adults, although the null hypothesis of age equivalence in correlations of after-study predictions with recall could not be rejected (Fisher’s r to z = 1.12, $p > .10$).

**Accuracy of Item-by-Item Predictions**

*Calibration curves.* Calibration curves are presented in Figure 2. First, note that the slopes of the functions relating item-by-item predictions with recall are similar between young and old adults in all four panels. Second, for the groups cued by the stimulus alone, delayed JOLs produced better calibrated item-by-item predictions than did immediate JOLs. Third, older adults appeared to be somewhat more accurate than younger adults, in that their calibration curves were closer to the main diagonal, particularly in the immediate-JOL conditions.

*Gammas.* Table 7 reports the mean gamma correlations. Results of the ANOVA with respect to the delayed-JOL effect mirrored those in Experiment 1, with the critical JOL Type × Time of JOL interaction achieving statistical significance, $F(1, 110) = 5.28, MSE = 0.10, p < .05$. As expected, the gamma correlation was higher for the delayed-JOL condition than for the immediate-JOL condition, but only for participants cued by the stimulus alone. There were no age differences in this effect (interaction $F < 1$).

**Relations Between JOLs and Global Predictions**

Table 8 presents the correlations of immediate and delayed JOLs with predictions, postdictions, and paired-associate recall. A striking finding was the high correlations of both types of JOL with the first prediction for both younger and older adults.

---

**Figure 2.** Experiment 2 calibration curves. I-SA = immediate stimulus-alone judgment of learning (JOL) condition; D-SA = delayed stimulus-alone JOL condition; I-SR = immediate stimulus–response JOL condition; D-SR = delayed stimulus–response JOL condition. Filled circles represent older adults, and filled squares represent younger adults.
Although in all four cases sample correlations between delayed JOLs and Prediction 2 (after study) were higher than those between delayed JOLs and Prediction 1 (before study), the difference appeared to be attenuated relative to that found in Experiment 1. A second interesting feature of the correlational data was the higher correlations involving immediate JOLs for older adults, relative to younger adults, in the stimulus-alone JOL condition.

**Discussion**

Raising the recall levels of the older and younger adults had a subtle but important effect on relations among global predictions, JOLs, and recall. Namely, the age differences detected in Experiment 1 in calibration curves and in the correlations of global predictions with recall disappeared in Experiment 2. These results have implications regarding the sensitivity of predictive accuracy to levels of recall. First, despite the increase in recall from Experiment 1 to Experiment 2, global predictions made before study differed negligibly from those found in Experiment 1. In fact, the increase in recall across experiments was contrasted by a slight decrease in global predictions made before study. Accordingly, such boosts in recall without concomitant changes in before-study global predictions resulted in a different pattern of findings with regard to accuracy as measured by difference scores. In Experiment 2, difference scores showed (a) substantial underprediction by younger adults, in contrast to their nearly perfect predictive accuracy in Experiment 1, and (b) relatively accurate prediction to underprediction by older adults, in contrast to their overprediction in Experiment 1.

The data from Experiment 2 demonstrate that calibration curves for immediate JOLs may also be sensitive to levels of recall. Whereas in Experiment 1 recall was closer to 50% for younger adults than for older adults and calibration curves for younger adults were closer to the line of perfect calibration, in Experiment 2 recall was closer to 50% for older adults and their calibration of immediate JOLs appeared closer to the line of perfect calibration. The same pattern was evident for delayed JOLs cued by the stimulus–response pair but not for delayed JOLs cued by the stimulus alone, which were typically highly calibrated regardless of the level of recall.

In contrast to the difference scores or measures of calibration, there was no age difference in the correlation between the before-study prediction and recall. The overall pattern suggests that correlational measures of predictive accuracy may be less sensitive to the level of recall than are measures that have been more frequently used in the aging literature (e.g., difference scores).

It was important that the initial study trial did not create age differences in the correlations between before-study predictions and recall (as may be predicted by a resource limitation hypothesis; see Bieman-Copland & Charness, 1994). On the contrary, no reliable age differences were found in initial correlations, and the age difference in correlational upgrading was eliminated in Experiment 2. Moreover, correlations between initial predictions and recall were higher in Experiment 2 than in Experiment 1 for both age groups, suggesting that the initial study trial enhanced the accuracy of the first prediction, possibly as a function of the beneficial effect of monitoring study before making the prediction. Rather than simply the initial study trial, another factor that may have influenced the initial predictions in Experiment 2 was the instruction to use mediators during study. Although no evidence currently indicates that using mediators increases predictive accuracy (but for inconsistent evidence see Dunlosky & Nelson, 1994), Lovelace (1984) showed that extra study trials boost the accuracy of younger adults' immediate JOLs (cf. analogous results from the present Experiments 1 and 2, reported in Table 7). Perhaps on-line monitoring during any given study trial benefits both item-by-item and global predictions.

The results from Experiment 2 are limited to a degree by the tendency for a ceiling effect in younger adults. Even though we excluded participants with extremely high recall, the restriction of range could have limited the obtained correlations. In Experiment 3 we used an alternative method to constrain levels of paired-associate recall to avoid ceiling and floor effects.

**Experiment 3**

Ideally, one would like to obtain intermediate memory performance for both younger and older adults in a single experimental task that did not rely on extra study trials to produce greater performance. Experiment 3 was designed to accomplish this goal. That is, the experimental design of Experiment 1 was once again used, but the degree of intrinsic semantic association between the two words of a pair was manipulated to produce intermediate performance for both younger and older individuals.

Several studies have examined the sensitivity of older and younger adults’ item-by-item predictions to manipulations of experimental variables, such as the sensitivity of JOLs to different study activities (Rabinowitz et al., 1982) or to different kinds of cue–target relations (Bieman-Copland & Charness, 1994; Shaw & Craik, 1989). Sensitivity in these studies has generally been measured in two ways: mean differences in item-by-item predictions between experimental conditions and differences between predictions and recall, when the former are re-scaled in the metric of recall performance. Both older and younger adults’ predictions are typically sensitive to experimental manipulations in the weaker sense of showing mean differences across experimental conditions; however, in at least some cases older adults’ predictions are more discrepant from the probability of successful remembering than are younger adults’ predictions (Bieman-Copland & Charness, 1994; Lovelace & Marsh, 1985). Lovelace (1990) suggested that older adults’ predictions are more sensitive to variations in observable stimulus properties (e.g., the degree of association between word pairs) than to processing differences induced by manipulations of instructions, orienting tasks, and the like.

The manipulation of association strength enabled us to examine Lovelace’s (1990) hypothesis that both old and young adults’ item-by-item predictions are sensitive to variations in stimulus properties. We created a mixed list of high- and low-association pairs and were then able to evaluate whether there were age differences in the sensitivity of predictions to the level of relatedness. Previous work suggests that age and relatedness interact in producing recall performance, with older adults performing more poorly on low-association pairs, relative to high-
association pairs, than younger adults (e.g., Zaretsky & Halberstam, 1968; see Kausler, 1994, for a review). Would we find a similar Age \times Association interaction for the item-by-item predictions?

**Method**

**Participants**

Thirty community-dwelling volunteers from the Atlanta metropolitan area received $10 each for participation. The mean age of the older individuals was 65.5 years (SD = 12.3); they had an average of 15.1 years of education (SD = 3.5), reported themselves to be in good to excellent health, reported taking an average of 1.55 medications (SD = 2.20), and scored 21.0 on the vocabulary test (SD = 6.43).

Thirty younger adults recruited from the Georgia Institute of Technology psychology participant pool received extra course credit for participating in this study. They were an average of 20.1 years old (SD = 1.3), had attained 13.6 years of education (SD = 1.0), were in good to excellent health, took an average of 0.27 medications (SD = 0.52), and scored 17.5 on the vocabulary test (SD = 3.9).

Older adults took more medications, F(1, 58) = 9.70, MSE = 2.5, p < .01, and had higher vocabulary scores, F(1, 59) = 6.59, MSE = 28.42, p < .01, than did the younger adults.

**Materials**

Items for the recall task were 10 practice noun–noun pairs, 6 buffer pairs, and 60 pairs for the experimental task. One third of the list of experimental pairs was taken from the pairs generated for Experiment 1 and had low levels of association between the stimulus and response (see the Appendix). Words in two thirds of the pairs were considered highly associated on the basis of the University of South Florida associability norms. However, we minimized the possibility of guessing the target word by eliminating the primary associates of the stimulus word (see Bieman-Copland & Charness, 1994). Each of these pairs was selected so that not more than 10% of the normative sample produced the target word as a first associate to the stimulus item. The 2:1 ratio of high- and low-association pairs was selected to ensure that older adults’ performance would be well above floor.

**Procedure**

Procedures followed the pattern established in Experiment 1, with the exception of the following modifications: (a) Participants did not fill out any metamemory questionnaires in this experiment; (b) all participants received only stimulus-alone cues when making their JOLs; (c) as in Experiment 2, all participants were instructed to use a strategy to learn the pairs and were given some sample strategies; and (d) postdictions were made after the strategy questionnaire was answered in order to prevent participants from scanning recent memory to arrive at their postdictions. Otherwise, the procedure was identical to that followed in Experiment 1. The experiment lasted approximately 60 min.

**Results**

The implementation of the present procedure was successful in that only 1 younger adult performed above 95% correct recall and had to be excluded from the analyses. Therefore, 29 young adults and 30 older adults were included in the analyses.

**Overall Memory and Metamemory Performance**

**Mean level of recall.** Table 2 contains the mean proportion recalled by each age group. Older adults again recalled less (M = .44) than did younger adults (M = .66), F(1, 56) = 33.63, MSE = 0.04, p < .001. Note that both young and older individuals performed near the middle of the scale and that the approximately 20% difference between their proportions that had been found in the previous two experiments was obtained here as well. Neither the main effect of time of JOL nor the Age \times Time interaction was significant (both Fs < 2.13). These findings, too, were consistent with the previous two experiments’ results for the stimulus-alone JOL condition (compare values for Experiments 2 and 3 in Table 2).

When decomposed by degree of association, the data revealed the expected effect of association on recall (Kausler, 1994). As can be seen in Table 9, young adults recalled more high-association words (M = .79) than low-association words (M = .47). This difference was similar but slightly larger for older adults (high-association M = .59; low-association M = .15). This Age \times Association interaction was statistically reliable, F(1, 58) = 7.51, MSE = 0.03, p < .01.

**Mean global predictions.** Table 3 presents the mean proportions predicted at recall before study, after study, and after recall for both age groups. The ANOVA revealed a significant main effect of age, F(1, 57) = 5.01, MSE = 0.06, p < .05, with the older adults making lower global predictions (M = .51) than the younger adults (M = .61). There was also a significant main effect of prediction phase, F(2, 114) = 3.72, MSE = 0.02, p < .05, such that the first global prediction was the highest (M = .60) and the last two were approximately equal (M = .53 and .55, respectively). The main effects of age and prediction phase were, however, qualified by a significant interaction of these variables, F(2, 114) = 10.33, MSE = 0.02, p < .001. This interaction can best be understood by contrasting the before-study global predictions of the older and younger adults. Older adults predicted higher recall before they began studying the list and then revised their predictions downward after study. The younger adults predicted the same mean level of recall before and after study of the paired-associate list. Only after the recall test did they raise their postdictions to match the level of recall they had actually achieved.

**Mean JOL ratings.** The mean JOL ratings are shown in Table 4. The ANOVA yielded significant main effects of both age, F(1, 58) = 4.22, MSE = 0.04, p < .05, and time of JOL, F(1, 58) = 10.63, MSE = 0.01, p < .01, as well as a significant interaction, F(1, 58) = 4.61, MSE = 0.01, p < .05. For the older adults, the mean JOL rating was only slightly higher for delayed JOLs than for immediate JOLs. For the younger adults, however, the mean delayed JOL ratings were moderately higher than the immediate JOL ratings.

Table 9 presents mean JOLs by association level. Association had a highly reliable effect on mean JOL, F(1, 58) = 206.16, MSE = 0.02, p < .001, and there was a Time of JOL \times Association interaction, F(1, 58) = 5.54, MSE = 0.01, p < .05. The difference between mean JOLs for the immediate and delayed conditions was slightly higher for the low-association pairs. However, neither the two-way nor the three-way interactions involving age and association were statistically significant.

**Accuracy of Global Predictions**

**Difference scores.** The differences between the proportion of predicted recall and actual proportion recalled at each predic-
AGING AND METAMEMORY ACCURACY

Table 9
Effect of Association Strength and Time of Judgment of Learning (JOL) on Recall, Mean JOL, Mean JOL/Recall Differences, and Gamma Coefficients in Experiment 3

<table>
<thead>
<tr>
<th>Age group and measure</th>
<th>High</th>
<th></th>
<th>Low</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Immediate JOL</td>
<td>Delayed JOL</td>
<td>Immediate JOL</td>
<td>Delayed JOL</td>
</tr>
<tr>
<td></td>
<td>M  SD</td>
<td></td>
<td>M  SD</td>
<td></td>
</tr>
<tr>
<td>Older adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>.56 .16</td>
<td>.61 .18</td>
<td>.12 .12</td>
<td>.16 .12</td>
</tr>
<tr>
<td>Mean JOL</td>
<td>.68 .14</td>
<td>.69 .18</td>
<td>.36 .21</td>
<td>.40 .12</td>
</tr>
<tr>
<td>JOL - recall</td>
<td>.12 .08</td>
<td>.58 .45</td>
<td>.24 .24</td>
<td></td>
</tr>
<tr>
<td>Gamma*</td>
<td>.32 .51</td>
<td>.52 .45</td>
<td>.56 .45</td>
<td></td>
</tr>
<tr>
<td>Younger adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>.79 .14</td>
<td>.79 .18</td>
<td>.44 .23</td>
<td>.49 .25</td>
</tr>
<tr>
<td>Mean JOL</td>
<td>.72 .15</td>
<td>.79 .14</td>
<td>.40 .20</td>
<td>.57 .19</td>
</tr>
<tr>
<td>JOL - recall</td>
<td>-.07 .00</td>
<td>.00 .54</td>
<td>-.04 .08</td>
<td></td>
</tr>
<tr>
<td>Gamma*</td>
<td>.36 .48</td>
<td>.69 .54</td>
<td>.39 .60</td>
<td>.63 .54</td>
</tr>
</tbody>
</table>

* Because of low levels of recall for low-association pairs, valid data for gamma coefficients are based on 24 older and 28 younger adults.

Conclusion phase are reported in Table 5. Before study, older adults overpredicted their performance and younger adults underpredicted their performance. Significant main effects of age, $F(1, 57) = 19.09, MSE = 0.06, p < .001$, and prediction phase, $F(2, 114) = 3.72, MSE = 0.02, p < .05$, and a significant interaction between age and prediction phase, $F(2, 114) = 10.33, MSE = 0.02, p < .001$, were obtained. The older adults adjusted their overprediction at the after-study global prediction, whereas the younger adults corrected their underprediction only at postdiction.

Correlation. The Pearson correlations between global predictions and recall are shown in Table 6. Like Experiment 1, the initial (before-study) global prediction did not correlate significantly with recall. The after-study prediction correlated significantly with recall for both age groups. The change in correlation from before to after study was statistically reliable, likelihood ratio $\chi^2(2, N = 59) = 15.29, p < .001$. Although the older adults' sample correlation between after-study predictions and recall was lower than the younger adults' correlation, this difference was not statistically reliable.

The correlations of postdictions with recall were higher than the correlations of after-study predictions with recall but were lower in Experiment 3 than they had been in Experiment 2. This was most likely due to the change in the experimental procedure, in which participants made postdictions after answering the strategy questionnaire rather than immediately after recall. Nevertheless, the correlation between postdiction and recall was high and was equivalent for older and younger individuals.

Accuracy of Item-by-Item Predictions

Calibration curves. Figure 3 presents the calibration curves for the immediate and delayed JOLs. In contrast to Experiments 1 and 2, no consistent age differences were apparent in the calibration curves for immediate JOLs. Although calibration for delayed JOLs was somewhat better for younger adults, the more impressive outcome was the relatively accurate calibration of delayed JOLs for both age groups.

Gammas. Table 7 lists the mean gamma correlations. The ANOVA yielded a main effect of time of JOL, $F(1, 56) = 47.14, MSE = 0.05, p < .001$, such that delayed JOLs were more accurate than immediate JOLs. There was no significant main effect or interaction with age ($Fs < 1$). The small age differences that were present in the sample means were less than 0.20 standard deviations in magnitude.

The mean gamma correlations for immediate JOLs were significantly ($p < .001$) greater than 0 for both age groups, indicating greater than chance accuracy. Some of this effect could be attributed to JOLs being sensitive to the level of association of the stimulus–response pairs. That is, participants may routinely give high-association pairs a higher immediate JOL than they give low-association pairs. This would be consistent with the robust effect of association on mean JOLs. This effect could act to mask any age differences in accuracy of item-by-item judgments. To evaluate this issue, we computed gamma correlations separately for high- and low-association pairs (see Table 9). The gamma correlations did not differ greatly as a function of level of association and, more important, did not vary materially by age group. The data from Experiment 3 suggest age equivalence in magnitude of item-by-item gamma correlations.

Difference scores. As noted above, both age groups showed salient effects of association on mean JOL. However, there were age differences in sensitivity as measured by the difference score between mean JOL and recall performance. There were reliable main effects of age, $F(1, 58) = 23.24, MSE = 0.08, p < .001$, and association, $F(1, 58) = 24.59, MSE = 0.02, p < .001$. Younger adults were more accurate overall, and mean JOLs for high-association pairs were also more accurate. There was also an Age × Association interaction, $F(1, 58) = 4.52, MSE = 0.02, p < .05$. The interaction reflected the fact that older adults' low-association JOLs were relatively more discrepant from their low levels of recall.
Relations Between JOLs and Global Predictions

Table 8 provides the correlations of mean immediate and delayed JOLs with recall, global predictions, and global postdictions. Both age groups evinced higher correlations between mean delayed JOLs and recall than between mean immediate JOLs and recall and an increase in the correlation between the after-study prediction and delayed JOLs relative to that between the before-study prediction and delayed JOLs.

Discussion

The results from Experiment 3 provide additional evidence that older adults’ memory monitoring is not impaired. As in Experiments 1 and 2, there was little evidence of an age difference in JOL accuracy as measured by gamma correlations. It may be premature, however, to accept the null hypothesis of age equivalence in gamma on the basis of Experiment 3. The small age difference in sample means (effect sizes of approximately 0.20 standard deviations) may reflect small population differences in gamma. At minimum, one can safely argue that the effect of age on monitoring accuracy was small to nonexistent and paled in comparison with the robust age differences in paired-associate recall (in which effect sizes were greater than 1.50 standard deviations, averaging over conditions). Experiment 3 differs substantially from the other two experiments in that there was little evidence of age differences in the shape of calibration curves.

The results from Experiment 3 are not fully consistent with the hypothesis that participants would anchor their predictions at the midpoint of the performance scale, however. Younger adults did underpredict performance, and their global predictions were closer to 50% than were their recall levels. Older adults, however, predicted on average higher levels of recall than did younger adults before studying the materials. This led to substantial overprediction of performance initially, which was corrected in the after-study prediction. One possible explanation is that the presence of high-association items in the practice list (in 2:1 proportion, as in the experimental list) may have lured some older adults into making higher initial global predictions, overriding the tendency to anchor their predictions at the midpoint. An inspection of the univariate distributions did reveal that 4 older adults initially predicted they would recall 50 stimulus words (83% of the 60 words).

It is interesting that the older adults in Experiment 3 showed greater changes in mean global predictions after study than had been found in Experiments 1 and 2 and in the previous study by Hertzog et al. (1994). Older adults’ after-study predictions were highly correlated with the mean immediate JOLs and the mean delayed JOLs given during study. It may be the case—despite the fact that older adults’ mean JOLs overestimated probability of recall—that monitoring the lower likelihood of recall for low-association pairs caused the older adults to return their prediction to a level near the midpoint of the scale (.47). In any event, this second global prediction was highly accurate for older adults. In contrast, younger adults did not shift their mean global prediction after study and hence continued to underestimate their subsequent level of recall.

Experiment 3 produced what some would consider classic findings of overprediction by older adults, both in the initial global prediction and in the item-by-item JOLs (Bieman-Copland & Charness, 1994). It is interesting to note that older adults did differentiate high- and low-association pairs, producing lower mean JOLs for the latter. Despite this effect, older adults overestimated the chance of subsequent recall of low-association items.

Once again, Experiment 3 demonstrates that correlational measures of memory monitoring produce different outcomes from difference-score measures regarding age differences in monitoring accuracy. There were minimal age differences in gammas as a measure of accuracy of immediate and delayed JOLs, despite substantial differences in the accuracy of mean JOLs between the age groups. Younger adults’ initial global predictions were closer to actual recall levels than were older adults’ predictions, yet neither group showed a significant correlation of predictions with recall. Older adults’ second global prediction was slightly more accurate than younger adults’ sec-
and prediction, yet they showed a tendency for a smaller upgrade in their sample correlation between after-study prediction and recall.

General Discussion

The three experiments reported here assessed the effects of aging on different kinds of metamemory judgments. We organize discussion of the results around four topics: (a) the effects of aging on item-by-item JOLs, (b) the sensitivity of measures of absolute accuracy to levels of recall, (c) the effects of aging on global predictions, and (d) the implications of our findings for adult learning.

Effects of Aging on Item-by-Item JOLs

All three experiments replicated the delayed-JOL effect first reported by Nelson and Dunlosky (1991). Immediate JOLs were above chance in all experiments (a mean gamma across experiments of approximately .45 for both age groups), and both young and old adults were much more accurate in predicting recall when their JOLs were both delayed and cued by the stimulus alone (a mean gamma of approximately .85 across experiments) than when immediate and cued by the stimulus alone. Previous studies in aging only examined age differences in immediate JOLs, and none used gamma correlations to evaluate predictive accuracy. Thus, our results agree with the conclusions of prior studies that monitoring of memory as reflected by immediate JOLs is relatively spared by aging (see Lovelace, 1990) and extends this inference to delayed JOLs as well.

These findings are inconsistent with what Hertzog and Dixon (1994) characterized as a pure monitoring deficit hypothesis—the notion that age differences in the ability to monitor learning account for age differences in memory performance (see also Light, 1991). On the contrary, negligible age differences in the accuracy of monitoring were coupled with robust age differences in recall.

Sensitivity of Measures of Absolute Accuracy to Levels of Recall

Our emphasis on the relative sparing of metacognitive monitoring may seem puzzling, given the variability we found in the age-related differences in absolute accuracy of immediate JOLs. In Experiments 1 and 3, for example, older adults’ mean immediate JOLs substantially overestimated the likelihood of subsequent recall, whereas younger adults showed better predictive accuracy (compare the immediate JOLs shown in Table 4 with corresponding values of recall in Table 2). In Experiment 2, however, older adults’ mean immediate JOLs accurately estimated recall, whereas younger adults underestimated recall, particularly in the stimulus—response condition.

We propose that absolute accuracy as measured by the differences between predicted and actual recall is influenced by factors other than on-line monitoring of memory. Any factor that influences probability of recall after JOLs are made may limit these measures independently of the quality of monitoring that is reflected by the JOL. Consider immediate JOLs. A participant may accurately monitor the quality of his or her encoding of an item during study and yet fail to anticipate important variables that will affect later recall, such as within-list interference effects due to similarity of stimulus words, the retention interval, and so forth. Thus, age-related differences in absolute accuracy may be determined largely by variables other than monitoring that have an effect on probability of item recall.

This idea also illustrates that a person’s absolute accuracy may partly rely on how well he or she anticipates the degree to which various factors affect recall. Younger and older adults may be equally poor at anticipating these effects and hence anchor immediate JOLs near the middle of the scale (see Table 4). Accordingly, the pattern of age-related differences will be partially a function of which age group happens to recall about 50% of the items, an outcome that is often fixed by the demands of the task. Given that numerous factors are relevant to performance on any specific memory task, it is unlikely that younger and older adults will consistently predict the absolute level of recall performance for that task without significant training.

Analyses involving calibration curves also suggest that the absolute accuracy of immediate JOLs depends partially on the overall level of recall. Namely, calibration curves for immediate JOLs seem to be influenced by the overall level of recall, such that high or low levels of recall distort calibrations. For instance, consider the relatively poor calibration of immediate JOLs for older adults in Experiment 1 (in which their recall performance was extremely low) and for younger adults in Experiment 2 (in which their recall performance was relatively high). Absolute accuracy as measured by calibration curves (at least for immediate JOLs) may be determined less by on-line monitoring than by other factors that may reduce or enhance recall and thus may reflect how well participants anticipate effects of these other factors.

These factors also limit the relative accuracy of immediate JOLs as measured by gamma to the extent that they cause inversions in recall across items. In contrast to the changes in the pattern of age differences in absolute accuracy across experiments, however, age invariance was the norm for accuracy of immediate JOLs for predicting the recall of one item relative to another.

These findings and our rationale regarding item-by-item JOL accuracy have two important implications. First, measures of item-by-item accuracy based on calibration curves and difference scores and measures of item-by-item accuracy based on gamma correlations reflect different kinds of accuracy (i.e., absolute vs. relative aspects of accuracy, as discussed by Nelson, 1996) and thus result in different patterns of age-related effects. Therefore, neither measure should be used exclusively to assess the accuracy of people’s item-by-item predictions. Second, because calibration curves and difference scores (standard measures of absolute accuracy) appear to be partially dependent on level of recall, age differences in calibration curves should be interpreted cautiously if age differences exist in recall.

Effects of Aging on Global Predictions

The partial dependence of difference scores on the level of recall was also evident in before-study global predictions. Older adults in Experiment 1 overestimated performance, with predicted recall of 40% despite mean recall of about 20%. Likewise, younger adults in Experiment 2 predicted close to 50%
recall and hence substantially underpredicted their recall of nearly 70%. Results from Experiment 3, however, are inconsistent with the hypothesis that these discrepancies between global predictions and recall are solely caused by anchoring of the global predictions at 50% recall. Mean global predictions before study by both age groups were approximately 60%. Thus, any inaccuracy of global predictions is not completely dependent on anchoring at 50%. Two possible inferences are that multiple influences other than anchoring affect global predictions (e.g., memory self-efficacy; see Hertzog et al., 1994) and that the presence of high-association pairs in Experiment 3 override any tendency of older adults to limit predictions to 50% recall or lower. Nevertheless, given the partial dependence of difference scores on the level of recall, examining differences scores will be most informative about possible age differences in the accuracy of metacognitive monitoring when recall is matched for the two age groups.

Present findings for correlations between global predictions and recall replicate and extend various findings and hypotheses from Hertzog et al. (1994) regarding the accuracy of global predictions. We replicated the correlational upgrade effect in all experiments; that is, after-study predictions correlated more highly with recall than did before-study predictions. Moreover, in Experiments 1 and 3, the initial global predictions made before study correlated weakly (and often not significantly) with recall (see also Hertzog et al., 1994). However, as demonstrated in Experiment 2, the magnitude of this initial correlation can be boosted by providing a prior study trial (with no JOLs and no actual recall). This finding converges with the correlational upgrade effect for after-study predictions. That is, given on-line monitoring of study, the correlations of predictions with recall are boosted. Prior work without JOLs (Hertzog et al., 1994) and the results from Experiment 2 show that this effect does not rely on requiring JOLs for each item. This suggests that adults in both age groups spontaneously monitor learning and then use that monitoring to influence their predictions after study.

These experiments provide additional evidence of the link between monitoring at the level of item-by-item JOLs and global predictions made after study. In all experiments, the after-study predictions were more highly associated with mean immediate and delayed JOLs than were before-study global predictions. This pattern of relations also supports the argument that after-study prediction is based on monitoring, such as monitoring the ease of processing an item during study (as perhaps tapped by immediate JOLs) or monitoring the relative ease of covertly retrieving an item during study (as tapped by stimulus-alone delayed JOLs). Of course, other factors may influence these correlations, and it is interesting to note that the correlations between JOLs and both global predictions were generally higher for older than for younger adults. Such an effect may reflect how the scales are used (e.g., anchoring effects), influencing both global and item-by-item predictions, more so for older adults.

Results from the present research fail to support Hertzog et al.'s (1994) argument for an age deficit in the upgrading of the correlation between after-study predictions and recall. Although this effect was observed in Experiment 1, older adults in Experiment 2 showed equivalent upgrading. Results in Experiment 3 were ambiguous, in part because of the relatively small sample size (low power) for testing differences in correlations. The disappearance of the effect may be attributed to requiring people to make JOLs; with the initial study trial and JOLs, the salience of monitoring is increased and overrides any tendency of older adults to ignore evidence from monitoring in upgrading the after-study prediction. Requiring JOLs and providing extra study trials do not appear to be necessary, however, because Lineaweaver (1994) recently reported age-equivalent upgrading in correlations using a 40-item free-recall task in a large sample.

Taken together, these outcomes reinforce the argument that older adults are able to monitor encoding and retrieval processes during learning. When age differences in correlational upgrading do occur, they may reflect difficulties in applying information gained from monitoring to the production of a global prediction, rather than difficulties with on-line monitoring per se (see Hertzog et al., 1994, for a discussion of alternative hypotheses).

Implications for Adult Learning

Regardless of why older adults’ global and item-by-item predictions often inaccurately estimate the absolute level of their memory performance, such inaccuracy may influence subsequent learning. For instance, in everyday situations in which a person pairs his or her learning, older adults who overestimate their memory may prematurely terminate study (Murphy et al., 1981). In other situations where older adults underestimate their performance (as in our Experiment 2), they may spend too much time studying an item for a negligible return in recall. Accordingly, if such over- and underestimation is determined by factors that commonly change with each new task (e.g., retention interval), global predictions may often yield poor absolute accuracy and ineffective regulation of study. A challenge will be to discover techniques that ensure that the absolute accuracy of global predictions is consistently high across a variety of tasks.

The high degree of accuracy (both absolute and relative) we found for stimulus-alone delayed JOLs across all three experiments indicates that older adults can, in principle, benefit from on-line monitoring of study. That is, training individuals to use this kind of metacognitive monitoring to regulate their studying may increase the effectiveness of their learning (Bjork, 1994; Dunlosky & Nelson, 1992). Consistent with this possibility, Nelson, Dunlosky, Graf, and Narens (1994) demonstrated that younger adults more effectively allocated their study time when they used their own delayed JOLs to allocate study than when their study was allocated by the normative difficulty of the items. Furthermore, Dunlosky and Hertzog (1996) found that older adults spontaneously used their delayed JOLs to select items for restudy in the same manner as younger adults, indicating that older adults can also benefit from using this kind of monitoring to regulate learning.

This investigation presents evidence that older adults are able to monitor their learning effectively. Relative to younger adults they show equally accurate immediate JOLs, produce an equivalent delayed-JOL effect, and under some conditions show an equivalent upgrading of global after-study predictions. This research thus sets the stage for additional work to elucidate the conditions under which older adults do and do not use their...
relatively intact monitoring capabilities to regulate their learning and retention.

References


Dunlosky, J., & Nelson, T. O. (in press). Similarity between the cue for judgments of learning (JOLs) and the cue for test is not the primary determinant of JOL accuracy. Journal of Memory and Language.


Appendix

Paired-Associate Stimuli

Paired-Associates Used in Experiments 1 and 2

1. COTTON–SNAKE
2. LUNCH–MOTOR
3. SUGAR–STREAM
4. FLOWER–HORN
5. DIRT–QUEEN
6. ANIMAL–LIBRARY
7. CIRCLE–BOAT
8. TARGET–SEED
9. SHEET–LESSON
10. COFFEE–JURY
11. GARDEN–SISTER
12. PRISON–TISSUE
13. LUMBER–SUIT
14. FACTORY–STONE
15. BREAD–WEAPON
16. BEEF–SEAT
17. MUSEUM–TRUCK
18. TONGUE–PICTURE
19. TUBE–SNOW
20. THROAT–PRINCE
21. BAKER–WAGON
22. BABY–FOREST
23. GATE–MEAT
24. HILL–GAME
25. BAND–CEILING
26. MESSAGE–SOLDIER
27. SALT–MAYOR
28. DRESS–MOVIE
29. SONG–KNEE
30. PIANO–COAT
31. INCH–PORCH
32. BEER–GRASS
33. MILK–CLERK
34. UNCLE–COAL
35. FISH–CHAMBER
36. MISSILE–STOMACH
37. NOISE–POCKET
38. PENCIL–BEAR
39. COLUMN–KNIFE
40. TEXT–MAID
41. CHICKEN–STEEL
42. PLASTIC–NOTE
43. COMPOSER–WOOD
44. TOOL–COAST
45. ROAD–DANCER
46. JACKET–FINGER
47. DOLLAR–BLANKET
48. CLOTH–ATOM
49. WHEEL–SENATOR
50. WIRE–FORT
51. TRAFFIC–BEACH
52. ROOM–BIRD
53. CHAIN–FRUIT
54. NOSE–BENCH
55. CROWD–RICE
56. ROOF–POPE
57. HANDLE–CORN
58. CAMERA–BRAIN
59. LAKE–TEMPLE
60. BONE–WINE

Paired-Associates Used in Experiment 3

1. CHAIR–ARM
2. PAINT–ARTIST
3. FLUTE–BAND
4. PANTS–BELT
5. SKIN–BODY
6. PAPER–BOOK
7. LEAF–BRANCH
8. CAB–BUS
9. BICYCLE–CAR
10. FENCE–CHAIN
11. HEART–CHEST
12. FEATHER–CHICKEN
13. PIPE–CIGARETTE
14. TIE–COAT
15. SUGAR–COFFEE
16. PIG–COW
17. SIDEWALK–CONCRETE
18. LAMP–DESK
19. WALL–DOOR
20. CRAYON–DRAWING
21. SNOW–RAIN
22. NOSE–EAR
23. MILL–FACTORY
24. MIRROR–FACE
25. OYSTER–FISH
26. WEED–GARDEN
27. PLASTIC–GLASS
28. BEAR–HONEY
29. SUIT–JACKET
30. STOVE–KITCHEN
31. ANKLE–KNEE
32. SANDWICH–LUNCH
33. FLAME–MATCH
34. POTATO–MEAT
35. COLLAR–NECK
36. SALT–OCEAN
37. TISSUE–ORGAN
38. CASTLE–PALACE


AGING AND METAMEMORY ACCURACY

39. CAKE-PARTY
40. DOLLAR-PENNY
41. BONE-WINE
42. ANIMAL-LIBRARY
43. THROAT-PRINCE
44. BABY-FOREST
45. MESSAGE-SOLDIER
46. MILK-CLERK
47. UNCLE-COAL
48. COLUMN-KNIFE
49. TEXT-MAID
50. TOOL-COAST
51. ROAD-DANCER
52. FLOWER-HORN
53. WHEEL-SENATOR
54. WIRE-FORT
55. ROOM-BIRD
56. CROWD-RICE
57. TONGUE-PICTURE
58. HANDLE-CORN
59. CAMERA-BRAIN
60. TARGET-WOOD

Received January 21, 1996
Revision received June 10, 1996
Accepted June 10, 1996

Instructions to Authors

Authors should prepare manuscripts according to the Publication Manual of the American Psychological Association (4th ed.). All manuscripts must include an abstract containing a maximum of 960 characters and spaces (which is approximately 120 words) typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the Manual. Also, all manuscripts are subject to editing for sexist language.

Masked reviews are optional, and authors who wish masked reviews must specifically request them when submitting their manuscripts. Each copy of a manuscript to be mask reviewed should include a separate title page with authors’ names and affiliations, and these should not appear anywhere else on the manuscript. Footnotes that identify the authors should be typed on a separate page. Authors should make every effort to see that the manuscript itself contains no clues to their identities.

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 6.24). As this journal is a primary journal that publishes original material only, APA policy prohibits as well publication of any manuscript that has already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data upon which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 6.25). APA expects authors submitting to this journal to adhere to these standards. Specifically, authors of manuscripts submitted to APA journals are expected to have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242.

Authors should submit manuscripts in quadruplicate. All copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Dittoed or mimeographed copies are not acceptable and will not be considered. Authors should keep a copy of the manuscript to guard against loss. In addition to addresses and phone numbers, authors should supply electronic mail addresses and fax numbers, if available, for potential use by the editorial office and later by the production office. Effective in January 1997, the Incoming Editor is receiving all submissions to the journal. Submissions that are accepted will be published beginning with the 1998 volume. Mail manuscripts to the Incoming Editor, Leah L. Light, Psychology Department, Pitzer College, Claremont, California 91711.