

Age-Related Equivalence and Deficit in Knowledge Updating of Cue Effectiveness

Greg Matvey and John Dunlosky
University of North Carolina at Greensboro

Raymond J. Shaw
Merrimack College

Colleen Parks and Christopher Hertzog
Georgia Institute of Technology

Knowledge updating involves learning about cue effectiveness based on task experience. Prior research has yielded inconsistent conclusions regarding age and knowledge updating. To resolve this inconsistency, the authors analyzed the effects of aging within a single paradigm. Participants studied cue–target associates during 2 study–test trials. Cues included rhyme cues and highly effective category cues. On each study–test trial, different items were presented, and participants predicted recall performance, received a cued recall test, and postdicted performance. Knowledge updating was operationalized as an improvement in the accuracy of predictive judgments across trials. An age deficit was evident in improvements in absolute accuracy, whereas age equivalence was evident in relative accuracy. Evidence suggested that deficient inferential processes contributed to the age deficit in knowledge updating.

A central component of metamemory is knowledge about the effectiveness of cues and strategies for memory (Wellman, 1977), such as knowing that a semantic cue (e.g., slippery as a cue for “ice”) is more effective than is a nonsemantic cue (e.g., splice for “ice”). Although this knowledge tends to increase throughout childhood (Schneider & Pressley, 1997), adults of all ages do not necessarily possess accurate knowledge about the effectiveness of many cues and strategies (e.g., Brigham & Pressley, 1988; see also Hertzog & Hulstsch, 2000). One way adults may gain accurate knowledge involves using feedback from task experience to update existing knowledge about the differential effectiveness of cues and strategies for memory. Just as important, age-related deficits in such “knowledge updating” could reduce the likelihood that older adults will subsequently use effective cues and strategies (Bieman-Copland & Charness, 1994; Brigham & Pressley, 1988).¹ A major goal of the present research was to evaluate the degree to which aging influences knowledge updating after task experience.

Do Age-Related Deficits Occur in Knowledge Updating?

Although some evidence from previous research suggests that age-related differences in knowledge about cue and strategy effectiveness are minimal (Perlmutter, 1978; Bieman-Copland & Charness, 1994), reports of age differences in knowledge updating have

been inconsistent (Bieman-Copland & Charness, 1994; Dunlosky & Hertzog, 2000). Younger adults often update their knowledge of the relative effectiveness of various cues and strategies after task experience (Bieman-Copland & Charness, 1994; A. L. Brown, Smiley, & Lawton, 1978; Pressley, Levin, & Ghatala, 1984). However, different conclusions arose from two studies (Bieman-Copland & Charness, 1994; Dunlosky & Hertzog, 2000) that compared younger versus older adults in knowledge updating.

Bieman-Copland and Charness (1994) suggested that age-related deficits exist in knowledge updating. They had participants study a list of cue–target items with letter (e.g., “ic–ice”), rhyme (e.g., “hurt–dirt”), or category cues (e.g., “a diagram–chart”). These cues represent a levels-of-processing manipulation that results in higher recall for category-cued items than for letter-cued or rhyme-cued items (Craik & Tulving, 1975; Shaw & Craik, 1989). After studying a given item, participants made a judgment of learning (JOL) about the likelihood of recalling the response when later shown the cue. After all items had been studied on Trial 1, cued recall occurred. Participants then received a second trial in which new items were studied and judged. If knowledge about cue effectiveness was gained during the first trial, then JOLs on the second trial were expected to more accurately reflect the effect of each cue condition on recall. Younger adults differentially adjusted

Greg Matvey and John Dunlosky, Department of Psychology, University of North Carolina at Greensboro; Raymond J. Shaw, Department of Psychology, Merrimack College; Colleen Parks and Christopher Hertzog, Department of Psychology, Georgia Institute of Technology.

Correspondence concerning this article should be addressed to John Dunlosky, P.O. Box 26170, Department of Psychology, University of North Carolina at Greensboro, Greensboro, North Carolina, 27402-6170. E-mail: dunlosky@uncg.edu

¹ We use the term *knowledge updating* (as opposed to *belief updating*) to provide a connection to our previous use of this term (i.e., Dunlosky & Hertzog, 2000). The term *belief* has often been used in the aging literature in specific reference to a person’s beliefs about memory self-efficacy. Because knowledge updating may be affected by one’s memory self-efficacy, we further consider this specific kind of belief within the Discussion section. However, a comprehensive discussion of whether this kind of research examines knowledge (as justified true belief) or belief is beyond the scope of this article.

their JOLs across trials for each cue condition, reflecting the differential effectiveness of the cue conditions for memory. Older adults lowered their JOLs across trials, regardless of cue condition. Bieman-Copland and Charness (1994) concluded that older adults did not update their knowledge about the differential effectiveness of the cues but instead adjusted judgments across trials on the basis of more global aspects of performance.

By contrast, Dunlosky and Hertzog (2000) did not find age-related differences in knowledge updating. Participants studied items during two study-test trials, with different items presented for study during each trial. They were instructed to study half of the items with interactive imagery and the other half with rote repetition. Prior to each study trial, they predicted the total number of imagery items and the total number of repetition items that would later be recalled. These are called *global-differentiated predictions* because participants are asked to predict performance for an entire group of items (as with the more standard global judgments) but must differentiate among subsets of the items. These predictions are presumed to partially tap knowledge about the effectiveness of the strategies. During study, participants also made a JOL immediately after studying each item. Although both kinds of prediction presumably tap knowledge, JOLs also may be influenced by the encoding of individual items during study (Mazzoni & Nelson, 1995). By contrast, global-differentiated predictions occur prior to a study trial, and hence they provide a better measure of knowledge per se because they will be influenced less by on-line monitoring during study (Dunlosky & Hertzog, 2000). Accordingly, as compared to JOLs, global-differentiated predictions may be more sensitive to age-related changes in knowledge updating.

Dunlosky and Hertzog (2000) operationalized knowledge updating as increases in the accuracy of judgments at predicting recall performance across trials, and they reported several measures of predictive accuracy for each trial. The measure that was most closely related to the analyses reported by Bieman-Copland and Charness (1994) was *absolute accuracy*, which refers to the discrepancy between mean JOLs and mean recall (cf. calibration, Keren, 1991). Absolute accuracy was computed separately for imagery items and for repetition items. If participants learned about the effectiveness of these strategies on the first trial, then the absolute accuracy of the predictions was expected to improve on the second trial. Dunlosky and Hertzog (2000) found negligible age differences in the changes across trials of absolute accuracy both for global-differentiated predictions and for JOLs. A different measure of accuracy did demonstrate knowledge updating. This measure involved computing correlations across participants' mean judgments and mean recall performance for both older and younger adults. These between-persons correlations reflect the degree to which individual differences in predictions accurately reflect individual differences in recall performance. For both age groups, the magnitude of these correlations increased across trials.² This effect was interpreted as evidence for knowledge updating by both older adults and younger adults.

In summary, these two studies provide inconsistent conclusions regarding the existence of age-related deficits in knowledge updating. Although the task procedures differed in many ways, two differences may be critical in resolving the apparent inconsistency. First, the manipulations differentially affected recall. The levels-of-processing manipulation resulted in a 40% to 60% boost in

recall from letter-cued to category-cued items, whereas the strategy manipulation resulted in about a 20% boost from the repetition to imagery strategy. Thus, Dunlosky and Hertzog (2000) may have failed to discover age-related differences in knowledge updating (measured by the changes across trials in absolute accuracy), because a larger difference in recall across conditions may be necessary for younger adults to consistently update knowledge from task experience (Kruschke & Johansen, 1999).

Second, Bieman-Copland and Charness's (1994) conclusion of age deficits was based on only one aspect of knowledge updating. In particular, when learning about the effectiveness of a set of cues, a person can gain at least two different kinds of knowledge. One kind refers to learning about the overall effectiveness of each cue condition for memory (e.g., knowledge of the mean effectiveness of category cues). This kind of knowledge is reflected in the absolute accuracy of judgments in predicting recall for a particular cue condition. Another kind of knowledge refers to the relative effectiveness of different kinds of cue conditions for memory (e.g., knowledge that category cues yield relatively greater memory performance than do letter cues). Both kinds of knowledge are important in accurately estimating future performance (N. R. Brown & Siegler, 1993). Given that Bieman-Copland and Charness (1994) focused on changes in the mean JOLs across trials, their conclusion of an age-related deficit in knowledge updating pertains specifically to learning about the absolute effects of the different cue conditions. They did not examine knowledge updating of relative cue effectiveness.

Knowledge updating of relative cue effectiveness can be assessed by examining the changes across trials in a measure of relative accuracy of JOLs, such as the Goodman-Kruskal gamma correlation (Nelson, 1984). To assess knowledge updating of relative cue effectiveness, we collapsed across the different cue conditions on each trial before computing the gamma correlation. When the data on each trial are aggregated in this way, this *aggregate* gamma correlation between JOLs and recall will be influenced by the degree to which participants adjust their JOLs across trials to reflect the relative effects of the different cue conditions. More specifically, if participants learn about the relative effectiveness of different cue conditions during the first trial, the aggregate gamma correlation between JOLs and recall will increase across trials.³ An analysis of the aggregate gamma correlation by Dunlosky and Hertzog (2000) suggested that neither

² As we discuss in the remainder of the introductory section, our primary focus was on knowledge updating about the absolute and relative effects of the cue conditions. Because knowledge updating about these effects is not directly relevant to changes across trials in the magnitude of between-persons correlations, we do not discuss them in the main body of the article. For completeness, these correlations are reported in the Appendix, along with further rationale for using them as a measure of knowledge updating.

³ Relative accuracy may also improve across trials because individuals merely gain experience at discriminating between the relative recallability of individual items; thus, increases may not necessarily reflect updating knowledge about the relative effectiveness of the cue conditions per se. The effects of these two potential influences on Trial 2 accuracy—knowledge about cue conditions versus knowledge about item recallability—can be analytically separated (Dunlosky & Hertzog, 2000). A treatment of this issue is provided in the Results section.

older nor younger adults updated their knowledge about the relative effects of imagery versus rote repetition on recall performance. Thus, the question of whether age-related deficits occur in knowledge updating about the relative effectiveness of cues remains open.

In the present research, we investigated knowledge updating by combining methods used by Bieman-Copland and Charness (1994) with those used by Dunlosky and Hertzog (2000). We used the cue manipulation from Bieman-Copland and Charness (1994), because it results in a relatively large effect on recall. As in Dunlosky and Hertzog (2000), we examined changes across trials in both absolute accuracy and relative accuracy.⁴ Thus, we were able to evaluate knowledge updating about both the absolute and relative effects of the cue conditions by using a design that should provide sufficient experimental effect sizes to enable knowledge updating at least in younger adults.

Isolating Age Deficits in Knowledge Updating

Another critical issue we examined was whether any age-related differences found in knowledge updating can be attributed to inaccurate performance monitoring and/or deficient inferential processes regarding cue effectiveness. Accurate performance monitoring during recall on the first trial is presumably necessary for knowledge updating because participants must be able to determine which items are being correctly recalled. Performance monitoring itself is not sufficient for knowledge updating because an individual must infer that differences in recall among items are related to the different cue conditions. Thus, any age-related deficits that arise in knowledge updating (as measured by changes in predictive accuracy across trials) may result from deficits in performance monitoring and/or in inferential processes that are based on such monitoring.

Accordingly, we collected measures that tapped these component processes of knowledge updating. We assessed performance monitoring by having participants rate confidence in each recall response, called *retrospective confidence* (RC) judgments. The hypothesis that older adults have greater difficulties monitoring performance in this task (Bieman-Copland & Charness, 1994) predicts that age-related differences will be present in the accuracy of RC judgments (e.g., Kelley & Sahakyan, 1999; but see Dunlosky & Hertzog, 2000, for one task in which age differences were negligible in performance monitoring). If substantive age-related deficits in performance monitoring are not found, then another possibility is that age-related deficits in knowledge updating result from concurrent deficits in the inferences made about the different cue conditions. We assessed this possibility by having participants estimate the number of items that they had successfully recalled for each cue condition. That is, participants made global-differentiated postdictions after the recall trials. If these postdictions are highly accurate, then participants presumably made appropriate inferences about the differential effectiveness of the cue conditions based on their performance monitoring.

To overview, we had two primary goals. We first searched for age-related differences in knowledge updating by examining changes across trials in the absolute and relative accuracy of predictive judgments. Next, we isolated the causes of any age-related deficits with measures of postdictive accuracy, which tap components of knowledge updating.

Method

Design and Participants

The design was a 2 (age: younger vs. older adults) \times 2 (trial: first or second) \times 3 (cue: letter, rhyme, or category) mixed-model factorial with trial and cue as within-subject factors. Fifty-one younger adults (mean age = 19.2, *SD* = 1.9) and 51 older adults (mean age = 70.4, *SD* = 5.9) participated. Younger adults were students at the University of North Carolina at Greensboro who participated for course credit. Older adults were normal community-dwelling adults from the Atlanta area who received a nominal fee. As usual, mean performance on a standard vocabulary test was reliably better for older ($M = 18.1$, *SD* = 7.3) than for younger ($M = 14.9$, *SD* = 4.5) adults, $t(99) = 2.64$. Both age groups reported being in good health— $M_s = 1.7$, *SD*s = 0.6, $t < 1.0$, for both groups, on a scale ranging from 1 (*excellent health*) to 4 (*poor health*)—and both groups were well educated [mean years of education for older adults was 14.6, *SD* = 3.1, and for younger adults, it was 13.1, *SD* = 1.3, $t(98) = 3.06$]. Thus, the group of older adults was relatively select, being well educated and in good health.

Materials

We selected 90 paired associates from the larger pool of paired associates used by Bieman-Copland and Charness (1994). Each list consisted of 45 items. Within each list, 15 items were randomly assigned to receive letter cues (e.g., “ic-ice”), 15 items were randomly assigned to receive rhyme cues (e.g., “hurt-dirt”), and 15 items were randomly assigned to receive category cues (e.g., “a diagram-chart”). Within a list, no two items shared identical letter, rhyme, or category cues. Also, all targets within a list differed in the spelling of their first three letters. Within each age group, 26 participants received List A during the first study trial, and 25 participants received List B during the first study trial.

Procedure

There were no practice trials. Participants completed two study-test trials with a different list (List A or List B) of items on each trial. On each study-test trial, participants completed tasks in the order that they are described below.

Global-differentiated predictions. Prior to study, participants made global-differentiated predictions. One global-differentiated prediction was made for items with letter cues, one for items with rhyme cues, and one for items with category cues. The order of the global-differentiated predictions was counterbalanced across participants. For each global-differentiated prediction, participants were asked to indicate the percentage of the 15 items with a particular cue that they believed they would remember on the upcoming cued recall test. Participants were instructed to use any integer from 0 to 100 when making this judgment.

Study. Each item (e.g., “a diagram-chart”) was presented in its entirety for 10 s. Participants were instructed to study the item for an upcoming test of memory where they would be shown the cue (i.e., “a diagram-”) and would have to recall the response (i.e., “chart”).

Immediately following offset of the presentation of an item for study, participants made a JOL. The JOL was prompted by the cue of the item

⁴ Because we were assessing knowledge updating, our main focus was on changes across trials in measures of absolute and relative accuracy. Although previous research has shown that age-related deficits do occur in absolute accuracy (but usually not in relative accuracy), such dissociations have been demonstrated only within a study-test trial (e.g., Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002). Less is known about changes in accuracy across trials that are relevant to investigating knowledge updating.

(e.g., “a diagram—”) along with the query, “How confident are you that in about ten minutes from now you will be able to recall the second part of the item when prompted by the first part of the item? (use any number from 0 to 100: 0 = *definitely won’t recall*, 1 = *1% sure I will be able to recall*, 2 = *2% sure I will be able to recall*, 50 = *50% sure I will be able to recall*, 99 = *99% sure I will be able to recall*, and 100 = *definitely will recall*).” After making a JOL, participants were immediately presented with the next item.

Cued recall. Following the JOL for the last item in a list, the order of items was randomized anew for recall. The cues for each item were individually presented (e.g., “a diagram—” would be presented at test for “a diagram—chart”). Participants were instructed to type in the word that was paired with that particular cue during study. Recall trials were self-paced, and omissions were not allowed. To minimize spelling errors, we scored a response correct if the first three letters of a word were typed correctly.

For each item, immediately after entering a response, participants made an RC judgment for each response. Participants were asked, “How confident are you that the response you just gave is correct? (use any number from 0 to 100: 0 = *definitely was not correct*, 1–99 = *1%–99% sure answer was correct*, and 100 = *answer is definitely correct*).” After making an RC judgment for a given item, participants were immediately presented with the cue for the next item.

Global-differentiated postdictions. Following the test trials, participants made global-differentiated postdictions. One postdiction was made for letter-cued items, for rhyme-cued items, and for category-cued items. The order of these postdictions was counterbalanced across participants. For each postdiction, participants indicated the percentage of items (on a scale from 0 to 100) that they correctly recalled on the test for that cue condition.

Results

In the analyses that follow, we first describe cued-recall performance to demonstrate that the cue conditions were differentially effective for memory. Next, we examine measures of knowledge

updating, which entails describing changes across trials in the accuracy of global-differentiated predictions and JOLs. Finally, to isolate any possible causes of age deficits in knowledge updating, we report the accuracy of RC judgments and global-differentiated postdictions. All differences declared as reliable had $p < .05$.

Cued-Recall Performance

For knowledge updating of cue effectiveness to occur, the cue conditions must be differentially effective for memory performance. Included in Table 1 are the mean values of recall for each cue condition. These values were analyzed by a 2 (age: younger or older adults) \times 2 (trial: first or second) \times 3 (cue: letter, rhyme, or category) analysis of variance (ANOVA). Reliable main effects were found for age, $F(1, 100) = 59.35$, $MSE = 0.13$, and for cue, $F(2, 200) = 361.97$, $MSE = 0.03$. The Age \times Cue interaction was also reliable, $F(2, 200) = 16.66$, $MSE = 0.03$. Age differences in recall were present for each cue condition, all $t_s > 3.00$. On the first trial, recall performance between the rhyme-cued and letter-cued items did not differ reliably for either age group, $t_s < 1$. On the second trial, recall performance was reliably greater for rhyme-cued than letter-cued items for the younger adults, $t(50) = 3.05$, but not for the older adults, $t < 1$. For both age groups and on both trials, recall performance was reliably greater for category-cued items than either for the letter-cued or rhyme-cued items, all $t_s > 3.00$. Thus, for older and younger adults, the opportunity existed for knowledge updating about the effectiveness of category cues relative to the other cues.

Absolute Accuracy

If participants update their knowledge about the average effectiveness of each cue condition for recall, then differences between

Table 1
Magnitude of Recall, Predictions, and Postdictions

Measure	Trial					
	1			2		
	L	R	C	L	R	C
	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>
Younger adults						
Recall	61 (3)	65 (3)	94 (1)	57 (3)	65 (3)	91 (1)
Global predictions	64 (3)	71 (3)	63 (3)	50 (3)	58 (3)	76 (3)
JOLs	60 (3)	60 (3)	65 (2)	56 (3)	56 (3)	72 (3)
Confidence	57 (3)	65 (3)	94 (1)	58 (3)	66 (3)	92 (1)
Global postdictions	46 (4)	56 (4)	81 (3)	50 (4)	58 (3)	82 (3)
Older adults						
Recall	37 (3)	32 (3)	79 (3)	36 (3)	35 (3)	81 (2)
Global predictions	59 (3)	65 (4)	61 (3)	34 (3)	43 (3)	48 (3)
JOLs	64 (3)	63 (3)	71 (3)	47 (3)	50 (3)	67 (3)
Confidence	32 (3)	38 (3)	78 (3)	35 (3)	43 (3)	80 (3)
Global postdictions	31 (3)	41 (3)	48 (3)	33 (3)	40 (3)	52 (3)

Note. All global judgments refer to global-differentiated judgments. Confidence refers to RC judgments made for responses during paired-associate recall. Entries are mean percentages for the corresponding measures. For each cue condition, 15 items were presented at study and test on both trials. L = letter cue; R = rhyme cue; C = category cue; JOLs = judgments of learning.

predictive judgments and recall will approach 0 across trials. Accordingly, we assessed changes in absolute accuracy across trials by computing, for each participant, signed difference scores between predictions and recall (i.e., mean prediction minus mean recall) separately for each cue condition.⁵ As discussed earlier, changes across trials in the accuracy of these predictions were most critical to our examination of knowledge updating; however, because the magnitude of the predictions are a component of accuracy, we report them in Table 1. In Table 2, we present the mean values across individuals' difference scores for global-differentiated predictions and JOLs. An inspection of changes in prediction magnitudes across trials yields the same conclusions drawn from the analyses of absolute accuracy described next.

For global-differentiated predictions, the difference scores were analyzed with a 2 (trial) \times 2 (age) \times 3 (cue) ANOVA. Reliable main effects occurred for age, $F(1, 100) = 7.67$, $MSE = 1937.38$; for trial, $F(1, 100) = 23.75$, $MSE = 951.8$; and for cue, $F(2, 200) = 209.38$, $MSE = 325.23$. Reliable two-way interactions were also revealed between age and trial, $F(1, 100) = 14.60$, $MSE = 951.8$; age and cue, $F(2, 200) = 21.62$, $MSE = 325.23$; and trial and cue, $F(2, 200) = 19.42$, $MSE = 300.81$. These main effects and interactions were qualified by a reliable three-way interaction, $F(2, 200) = 5.00$, $MSE = 300.81$. Younger adults' difference scores changed across trials for all three cue conditions, $t_s > 2.50$. For letter-cued and rhyme-cued items, these changes were indicative of little improvement across trials in absolute accuracy; note, however, that the difference scores were near zero on both trials, which indicates younger adults' predictions for these cue conditions were always highly accurate. For the category-cued items, difference scores converged toward zero across trials, which reflects an improvement in absolute accuracy. Older adults' difference scores also changed across trials for all three cue conditions, all $t_s > 3.3$. Absolute accuracy improved across trials

(values converging on zero) for letter-cued and rhyme-cued items. Most important and in contrast to the younger adults, a decline across trials in absolute accuracy (values diverging from zero) occurred for older adults' category-cued items. This pattern emerged because older adults shifted predictions downward across trials for each cue condition.

For JOLs, reliable main effects were obtained for age, $F(1, 100) = 22.87$, $MSE = 2984.15$; for trial, $F(1, 100) = 11.39$, $MSE = 384.04$; and for cue, $F(2, 200) = 134.53$, $MSE = 340.51$. Reliable two-way interactions were also found between age and trial, $F(1, 100) = 18.38$, $MSE = 384.04$; age and cue, $F(2, 200) = 11.53$, $MSE = 340.51$; and trial and cue, $F(2, 200) = 12.06$, $MSE = 176.91$. Although the three-way interaction was not reliable, $F < 1$, the pattern of changes in the difference scores for JOLs across trials was similar to the pattern described above for the global-differentiated predictions (as in Bieman-Copland & Charness, 1994). Younger adults' difference scores did not reliably change from Trial 1 to Trial 2 for the letter-cued or rhyme-cued items, $t_s < 1.50$, but for category-cued items, difference scores were indicative of improvements in absolute accuracy across trials, $t(50) = 4.06$. Older adults' difference scores changed across trials for each cue condition, all $t_s > 2$, and were indicative of improved absolute accuracy across trials for the letter-cued and rhyme-cued items but decreased accuracy for the category-cued items.

In summary, age-related differences were found in the changes across trials in the absolute accuracy of global-differentiated predictions. These findings conceptually replicate those from Bieman-Copland and Charness (1994), and the same pattern was also evident for JOLs, providing further evidence for age differences in knowledge updating of the absolute effectiveness of each cue condition. For younger adults, increases across trials in absolute accuracy were apparently due to a shift of predictions across trials that reflected the differential effectiveness of the cue conditions for memory. For older adults, however, absolute accuracy decreased across trials for category-cued items, with the magnitude of predictions shifting away from the mean level of category-cued recall (Table 1). Accordingly, even the improvements in absolute accuracy across trials for rhyme-cued and letter-cued items may have been less indicative of updating knowledge about the absolute effectiveness of each cue condition and more indicative of a global decline in older adults' beliefs about their memory ability.

Relative Accuracy of the Judgments of Learning

Our primary assessment of relative accuracy involved aggregating the items on each trial, without regard to cue condition, and then computing a gamma correlation between JOLs and recall across this aggregated set of items. If participants updated their knowledge about the relative effectiveness of the cue conditions after task experience on the first trial, then these aggregate gamma correlations will increase across trials.

Table 2
Absolute Accuracy of Global-Differentiated Predictions and Judgments of Learning (JOLs)

Group	Cue condition		
	Letter	Rhyme	Category
	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)
Global-differentiated predictions			
Younger adults			
Trial 1	3.51 (4.17)	6.15 (4.49)	-30.90 (3.12)
Trial 2	-7.31 (3.40)	-6.87 (3.59)	-14.94 (3.19)
Older adults			
Trial 1	21.92 (3.73)	32.83 (4.41)	-17.82 (3.60)
Trial 2	-2.19 (3.28)	7.52 (3.62)	-33.46 (3.19)
JOLs			
Younger adults			
Trial 1	-0.34 (3.67)	-4.98 (4.59)	-28.65 (2.67)
Trial 2	-1.52 (3.39)	-9.19 (3.86)	-18.93 (2.83)
Older adults			
Trial 1	26.60 (4.24)	31.22 (4.71)	-8.05 (3.15)
Trial 2	11.71 (3.98)	15.34 (4.51)	-13.71 (3.19)

Note. Entries are means of individuals' difference scores between judgments and recall.

⁵ An alternative is to compute the unsigned difference scores between the mean predictions and mean recall performance. In contrast to these unsigned scores, the signed difference scores reported here indicate the degree the predictions reflect participants' underconfidence or overconfidence in recall performance and hence were preferable for analyses of knowledge updating. Moreover, analyses of the unsigned difference scores yielded the same conclusions, so we report only the signed scores.

The mean values of these correlations are presented in Table 3. A 2 (age) × 2 (trial) ANOVA revealed a main effect of trial, $F(1, 92) = 16.39, MSE = 0.08$, indicating that the aggregate accuracy for both age groups increased across trials. The main effect of age approached reliability, $F(1, 92) = 3.37, MSE = 0.09, p < .10$, with the direction of this effect toward greater accuracy for older versus younger adults. The interaction between age and trial was not reliable, $F < 2$.

Improvements across trials in the relative accuracy of JOLs suggest that both older and younger adults updated knowledge about the relative effectiveness of the cue conditions. An alternative interpretation is that experience on the first trial improves an individual’s skill at discriminating between the differential recallability of individual items. Thus, the increases in aggregate accuracy may be less indicative of knowledge updating than a general improvement in monitoring. If this interpretation is correct, then JOL accuracy computed separately for each cue condition will also increase across trials. The gamma correlations between JOLs and recall for each cue condition are also presented in Table 3. A 2 (age) × 2 (trial) × 3 (cue) ANOVA revealed no reliable main effects or interactions, all $F_s < 3$. Thus, increases in the aggregate accuracy of JOLs across trials cannot be explained by a general improvement in monitoring across trials.

Accuracy of Retrospective Confidence Judgments and Global-Differentiated Postdictions

Knowledge updating of cue effectiveness presumably requires both accurate monitoring of performance during test and valid inferences about the differential effectiveness of the cue conditions. We analyzed the accuracy of RC judgments to assess performance monitoring and the accuracy of global-differentiated postdictions to assess the quality of inferences drawn from performance monitoring. We investigated the degree to which these two components contributed to the age-related deficits in absolute accuracy (Table 2). Hence, only the corresponding analyses of absolute accuracy for the postdictions are reported.

For each participant, absolute accuracy was assessed by computing the signed difference scores between mean postdictions and mean recall for each cue condition. Because the magnitude of the postdictions is a component of accuracy, we report them in Ta-

ble 1. Mean values across individuals’ difference scores are reported in Table 4.

For RC judgments, a 2 (age) × 2 (trial) × 3 (cue) ANOVA revealed main effects for trial, $F(1, 100) = 4.50, MSE = 108.16$, and for cue, $F(2, 200) = 14.39, MSE = 142.44$. The Age × Cue interaction was also reliable, $F(2, 200) = 8.30, MSE = 142.44$. This interaction was due primarily to the greater accuracy of the younger adults’ RC judgments (compared to those of older adults) for rhyme-cued items both on the first trial, $t(100) = 2.14$, and second trial, $t(100) = 2.57$.

Age-related differences were evident in absolute accuracy of the global-differentiated postdictions. A 2 (age) × 2 (trial) × 3 (cue) ANOVA revealed both a main effect of cue, $F(2, 200) = 46.25, MSE = 474.44$, and an Age × Cue interaction, $F(2, 200) = 32.11, MSE = 474.44$. Older and younger adults were equally accurate at estimating their performance for letter-cued items, $t < 2$. For rhyme-cued items, reliable age-related differences emerged on the first, $t(100) = 4.25$, and second trial, $t(100) = 2.28$. Note, however, that the absolute values of the difference scores for rhyme-cued items for older and younger adults were nearly identical. The most critical age difference occurred for the category-cued items. Although both groups underestimated performance, older adults were reliably less confident than were younger adults at estimating their performance on the category-cued items on both the first trial, $t(100) = 4.25$, and on the second trial, $t(100) = 5.03$. This particular effect suggests that after experience with the task, older adults made less accurate inferences about the specific influence that the category cues had on recall.

Discussion

Previous research generated inconsistent conclusions regarding age differences in knowledge updating. Results of the present study, however, indicate that the conclusions of age-related deficit (Bieman-Copland & Charness, 1994) and age equivalence (Dunlosky & Hertzog, 2000) in knowledge updating may both be appropriate. Age differences occurred in the changes across trials in absolute accuracy (suggesting age-related deficits), whereas both older and younger adults showed improvements across trials in relative accuracy (suggesting age equivalence). We describe this dissociation in more detail next and then consider some possible interpretations.

Table 3
Relative Accuracy of Judgments of Learning

Group	Cue condition			
	Aggregate	Letter	Rhyme	Category
	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>	<i>M (SE)</i>
Younger adults				
Trial 1	0.24 (0.05)	0.16 (0.56)	0.20 (0.06)	0.36 (0.11)
Trial 2	0.45 (0.04)	0.16 (0.56)	0.37 (0.05)	0.30 (0.13)
Older adults				
Trial 1	0.35 (0.04)	0.19 (0.07)	0.27 (0.09)	0.31 (0.09)
Trial 2	0.48 (0.04)	0.27 (0.07)	0.25 (0.07)	0.15 (0.10)

Note. Entries are means of individuals’ gamma correlations between judgments of learning and recall. “Aggregate” refers to relative accuracy computed across all items and without respect to cue condition.

Table 4
Absolute Accuracy of Retrospective Confidence (RC) Judgments and Global-Differentiated Postdictions

Group	Cue condition		
	Letter	Rhyme	Category
	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)	<i>M</i> (<i>SE</i>)
RC judgments			
Younger adults			
Trial 1	-3.43 (1.92)	0.37 (1.37)	0.27 (0.64)
Trial 2	0.89 (1.83)	1.16 (1.53)	1.22 (0.71)
Older adults			
Trial 1	-4.81 (2.74)	6.65 (2.60)	-0.70 (1.48)
Trial 2	-1.23 (2.75)	8.13 (2.24)	-1.15 (1.84)
Global-differentiated postdictions			
Younger adults			
Trial 1	-14.23 (3.21)	-8.18 (2.70)	-13.31 (3.07)
Trial 2	-7.03 (3.04)	-6.34 (3.06)	-8.98 (2.48)
Older adults			
Trial 1	-6.29 (3.08)	9.67 (3.22)	-30.69 (2.69)
Trial 2	-3.03 (3.31)	4.71 (3.77)	-29.22 (3.16)

Note. Entries are means of individuals' difference scores between postdictions and recall.

As knowledge about cue effectiveness is updated, predictions (global-differentiated predictions and JOLs) should be adjusted across trials to come into closer correspondence with the actual effectiveness of the cues. In the present study, younger adults differentially adjusted predictions across trials as a function of cue condition, whereas older adults adjusted their predictions downward regardless of cue condition. Most important, for the category cues, these adjustments resulted in increased accuracy across trials for the younger adults' predictions and decreased accuracy across trials for the older adults' predictions. These outcomes conceptually replicate those from Bieman-Copland and Charness (1994). In their study, older adults' recall on Trial 1 for all cue conditions was lower than their corresponding JOLs, so that the global shift downward in JOLs for all cue conditions was not necessarily indicative of inaccurate knowledge updating. By contrast, consider the values for older adults in Table 1: On Trial 1, recall was lower than JOLs for rhyme cues and letter cues, but recall was higher than JOLs for category cues. Thus, the global shift downward in JOLs on Trial 2 may be viewed as indicative of inaccurate updating. An alternative interpretation of these effects is that the shift downward in older adults' predictions for the rhyme-cued and letter-cued items does represent accurate knowledge updating and is not indicative of a global deficit. That is, older adults' improvements across trials in absolute accuracy for these two classes of items may represent partial knowledge updating. Both interpretations (i.e., partial updating or global deficit) are plausible, and further research will be required to competitively evaluate them. Regardless of the correct interpretation, however, the critical point is that the absolute accuracy of older adults' predictions for the category-cued items did decrease across trials. Put differently, the intriguing age-related effect on absolute accuracy concerns the deficit for the category-cued items, not the less interpretable

changes in absolute accuracy for rhyme-cued and letter-cued items.

Moreover, older adults' predictions for rhyme-cued and letter-cued items tended to shift downward more across trials as compared with predictions for category-cued items, which is indicative of accurate knowledge updating that was captured by improvements in relative accuracy across trials. That is, in contrast to the aforementioned age-related deficit in absolute accuracy, when we examined changes across trials in the relative accuracy of JOLs aggregated across cue conditions, both age groups showed evidence of knowledge updating.

One interpretation for the dissociation between the changes across trials in absolute and relative accuracy is apparent when one considers the task a person must perform when estimating the effectiveness of a particular cue condition. When making a prediction, people presumably apply knowledge they have about the effectiveness of a cue condition and map that knowledge onto a rating scale. Relative accuracy, aggregated across cue conditions, reflects the knowledge people have about the relative effectiveness of the cue conditions on memory performance. In contrast to relative accuracy, absolute accuracy is also influenced by the process of mapping this knowledge onto specific estimates (e.g., 20%) on the rating scale and pertains to knowing the absolute effectiveness of each cue condition on memory performance. In these terms, the dissociation may have occurred for older adults because they had learned about the relative effectiveness of the cue conditions but did not accurately translate this knowledge into point estimates representing the absolute effectiveness of the cues, especially in relation to the category-cued items.

Even so, the question remains as to why older adults were deficient at estimating the degree to which the category cues influenced recall performance. One explanation lies in the differential effects of aging on the components of knowledge updating. Two central components are monitoring performance and making inferences based on that monitoring (for a detailed discussion, see Dunlosky & Hertzog, 2000). Products from performance monitoring (e.g., knowing which items are correct vs. incorrect on the test) and subsequent inferences (e.g., judging how many category-cued items had been correctly recalled) are critical for accurate knowledge updating. Accordingly, either component may have mediated the age-related differences observed in the changes across trials in absolute accuracy. In the remaining discussion, we discuss the role of each component through a consideration of people's RC judgments and global-differentiated postdictions.

In previous research, performance monitoring appears to be relatively unaffected by age (Devolder, Brigham, & Pressley, 1990; Dunlosky & Hertzog, 2000; Hertzog, Saylor, Fleece, & Dixon, 1994). Most important, inspection of Table 4 reveals that the absolute accuracy of older adults' RC judgments was excellent, especially for the category-cued items. Thus, age-related declines in performance monitoring per se apparently contribute little to the aforementioned age deficiencies in knowledge updating.

Analyses of global-differentiated postdictions suggest that age-related deficits in inferential processing contribute to the poor absolute accuracy of older adults' predictions for category-cued items. To understand why, consider again two kinds of knowledge that may be inferred from performance monitoring: knowledge about the relative effectiveness of the different cue conditions at test (e.g., category cues are most effective), and knowledge about

the absolute effectiveness of each cue condition (e.g., the number of category-cued items recalled). Both older and younger adults showed evidence of accurate inferences about the relative effectiveness of the cues, as indicated by a reliable levels-of-processing effect on their global-differentiated postdictions. That is, these postdictions were consistently higher for category-cued items than for the rhyme-cued and letter-cued items (Table 1). However, older adults appeared to be particularly poor at inferring the absolute effectiveness of the category cues: Whereas their mean level of recall for category-cued items on Trial 1 was 79%, the corresponding postdiction was only 48%. Why may older adults inaccurately postdict the absolute effectiveness of category cues? We offer three possible answers to this question.

First, older adults may forget which items had been correctly recalled on the test trial when later making the judgments. Second, they may not accurately estimate the frequency with which items were correctly recalled as a function of cue condition. The first hypothesis cannot readily account for the excellent absolute accuracy of older adults' postdictions for rhyme-cued and letter-cued items. However, evidence from a recent investigation is consistent with the second hypothesis that deficiencies in judging frequency may contribute. Mutter and Goedert's (1997) participants studied words that varied in the number of repetitions (from one to six). After presentation, they were shown pairs of words and either circled the word that was presented most frequently or estimated the number of times each word had been presented. Both older and younger adults discriminated equally well between items that were presented more versus less frequently. Note that this outcome is analogous to the present one, indicating older adults obtain knowledge about the relative effectiveness of the different cue conditions. By contrast, age differences arose in the absolute estimates of how often a particular item was presented, with older adults showing the most extreme levels of underestimation for items that had the highest frequency of presentation. An analogous pattern is also evident in the present data: Older adults underestimated the frequency of correct recall for items with category cues, which had the highest frequency of correct recall.

A third hypothesis is that older adults base their global-differentiated postdictions solely on the average level of recall across all cue conditions. Although previous studies have reported that older adults' postdictions can be quite accurate (e.g., Devolder et al., 1990), these studies have typically used a single, undifferentiated postdiction for an entire list of items that included no within-participant manipulation of item variables (such as cue condition). In the present research, the overall level of correct recall (averaged across all cue conditions) for older adults was 49%, which was lower than their initial estimates about how well they would perform (i.e., global-differentiated predictions made on Trial 1 were around 60%). Thus, in adjusting their beliefs about task performance, older adults may have been unduly influenced by their overall level of performance. That is, they may anchor their postdictions on overall recall instead of basing each global-differentiated postdiction solely on an analysis of how many items were recalled per cue condition. One plausible contributing factor is older adults' possible adverse reaction to their overconfidence on Trial 1 global-differentiated predictions and JOLs. On monitoring low levels of overall recall performance, they may overcompensate for their initial level of overconfidence (Hertzog & Hultsch, 2000), perhaps as a result of the activation of beliefs

regarding low-memory self-efficacy. Older adults' lowered memory self-efficacy could in turn cause them to postdict low levels of recall for all cue conditions.

The three explanations described above provide somewhat different perspectives on why older adults show deficits in knowledge updating about the absolute effectiveness of the category cues for recall. Each one can be empirically evaluated by using relatively minor variations of the standard methods used to investigate knowledge updating that have been developed here and elsewhere (e.g., Brigham & Pressley, 1988). Because the explanations are not mutually exclusive, the challenge will be to estimate the degree to which the various factors (i.e., forgetting, poor frequency estimation, and anchoring due to poor memory self-efficacy) jointly explain the age-related differences in the inferential component of knowledge updating.

In summary, the present research provided a critical extension of prior research on the effects of age on knowledge updating. Apparent contradictory conclusions from the literature regarding age deficits in knowledge updating (Bieman-Copland & Charness, 1994; Dunlosky & Hertzog, 2000) were both supported and explained within a multicomponent framework that indicates how age deficit and age equivalence in knowledge updating can coexist. We also established that age-related deficits in knowledge updating were partly a function of an age-related deficit in inferential processes associated with global-postdiction accuracy. Discovering why older adults are deficient in this component of knowledge updating awaits future research.

References

- Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology and Aging, 9*, 287–302.
- Brigham, M. C., & Pressley, M. (1988). Cognitive monitoring and strategy choice in older and younger adults. *Psychology and Aging, 3*, 249–257.
- Brown, A. L., Smiley, S. S., & Lawton, S. Q. C. (1978). The effects of experience on the selection of suitable retrieval cues. *Child Development, 49*, 829–835.
- Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*, 511–534.
- Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.
- Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging, 5*, 291–303.
- Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*, 462–474.
- Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 417–466). Mahwah, NJ: Erlbaum.
- Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17*, 209–225.
- Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1994). Metamemory and aging: Relations between predicted, actual, and perceived memory task performance. *Aging and Cognition, 1*, 203–237.

- Kelley, C. M., & Sahakyan, L. (1999). *Age-related changes in monitoring and control of cued recall*. Paper presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, CA.
- Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.
- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1083–1119.
- Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 1263–1274.
- Mutter, S. A., & Goedert, K. M. (1997). Frequency discrimination vs frequency estimation: Adult age differences and the effect of divided attention. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, 52B, 319–328.
- Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin*, 95, 109–133.
- Perlmutter, M. (1978). What is memory aging the aging of? *Developmental Psychology*, 14, 330–345.
- Pressley, M., Levin, J. R., & Ghatala, E. S. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior*, 23, 270–288.
- Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd ed.). Mahwah, NJ: Erlbaum.
- Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging*, 4, 131–135.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87, 245–251.
- Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development*, 48, 13–21.

Appendix

Between-Persons Correlations Between Predictions and Performance

After task experience on Trial 1, an individual may adjust predictions on Trial 2 to more accurately reflect his or her own levels of recall yielded by different cue conditions. If so, individual differences in predictions for items studied with each kind of cue will be more highly related to individual differences in recall performance on Trial 2 than on Trial 1. To evaluate this possibility, we computed for each cue condition Pearson r correlations between individuals' mean predictions and recall. The Pearson r correlation covaries standard scores for predictions and performance, removing the mean levels from the two variables. It focuses instead on the agreement between prediction and performance in the relative ordering of individuals. As a result, the Pearson correlation is less influenced than difference scores (signed or unsigned) by factors that may affect a person's mean level of prediction, such as his or her appraisal of task difficulty (Hertzog et al., 1994). An increase in these correlations across trials is evidence that individuals adjusted their predictions on Trial 2 to more accurately reflect individual differences in recall performance.

The Pearson r correlations between participants' mean global predictions and recall are presented in Table A1. Although correlations tended to increase across trials for both age groups, these increases were not consistently reliable. Simple-effect tests (Steiger, 1980) were calculated for the difference in correlations between the first trial and the second trial for each cue condition. Reliable effects were found for younger adults for letter-cued items ($z = 2.74$) and for rhyme-cued items ($z = 2.76$) but not for category-cued items ($z = 0.44$). For older adults, the increase across trials for rhyme-cued items approached reliability ($z = 1.78$), but neither the increase for letter-cued items ($z = 1.19$) or category-cued items ($z = 1.32$) was reliable.

The Pearson r correlations between participants' mean JOLs and recall are also presented in Table A1. These correlations tended to increase across trials for both age groups. Simple-effect tests revealed reliable increases across trials only for the younger adults for rhyme-cued items ($z = 2.59$) and for the older adults for letter-cued items ($z = 2.49$). In summary,

Table A1
Between-Persons Correlations

Group	Cue condition		
	Letter	Rhyme	Category
Global-differentiated predictions			
Younger adults			
Trial 1	–0.04 _a	–0.03 _b	0.26
Trial 2	0.44 _a	0.44 _b	0.18
Older adults			
Trial 1	0.12	0.08	0.26
Trial 2	0.34	0.37	0.47
Judgments of learning			
Younger adults			
Trial 1	0.14	–0.03 _c	0.06
Trial 2	0.35	0.32 _c	0.27
Older adults			
Trial 1	–0.19 _d	–0.09	0.34
Trial 2	0.21 _d	0.10	0.39

Note. Entries are the Pearson r correlation between mean judgment and mean recall. Cell entries sharing a common subscript are reliably different at $p < .05$.

although between-persons correlations tended to increase across trials, the effects were not consistently reliable for either younger or older adults.

Received January 22, 2001

Revision received February 26, 2002

Accepted February 28, 2002 ■