Published in final edited form as:

Appl Cogn Psychol. 2009 ; 24(1): 134. doi:10.1002/acp.1553.

How Accurately Can Older Adults Evaluate the Quality of Their Text Recall? The Effect of Providing Standards on Judgment Accuracy

Julie Baker,

Department of Psychology, Wake Forest University

John Dunlosky, and

Department of Psychology, Kent State University

Christopher Hertzog

School of Psychology, Georgia Institute of Technology

Abstract

Adults have difficulties accurately judging how well they have learned text materials; unfortunately, such low levels of accuracy may obscure age-related deficits. Higher levels of accuracy have been obtained when younger adults make postdictions about which test questions they answered correctly. Accordingly, we focus on the accuracy of postdictive judgments to evaluate whether age deficits would emerge with higher levels of accuracy and whether people's postdictive accuracy would benefit from providing an appropriate standard of evaluation. Participants read texts with definitions embedded in them, attempted to recall each definition, and then made a postdictive judgment about the quality of their recall. When making these judgments, participants either received no standard or were presented the correct definition as a standard for evaluation. Age-related equivalence was found in the relative accuracy of these term-specific judgments, and older adults' absolute accuracy benefited from providing standards to the same degree as did younger adults.

Research on metacomprehension concerns people's ability to judge their learning and understanding of text materials. A major focus of this research is the relative accuracy of the judgments that people make over texts. To measure relative accuracy, participants read a series of texts and are asked to make a judgment about how well they will perform on a test for each text. Relative accuracy is computed by a within-person gamma correlation of each participant's judgments with test performance. Higher mean correlations, averaging over individuals, indicate that variation in predictions across texts are associated with higher levels of test performance for those texts. College undergraduates often have low relative accuracy, indicating that they are poor at judging which texts have been learned well and which texts have not been learned well (for reviews, see Dunlosky & Lipko, 2007; Maki, 1998; Weaver, Bryant, & Burns, 1995; for exceptions to poor accuracy using this method, see Thiede, Dunlosky, Griffin, & Wiley, 2005).

Despite extensive work employing undergraduate samples, only four studies have investigated older adults' judgment accuracy (Dunlosky, Baker, Rawson, & Hertzog, 2006; Miles & Stine-Morrow, 2004; Olin & Zelinski, 1997; Stine-Morrow, Shake, Miles, & Noh, 2006). As shown in Table 1, three of the four studies reported that older adults judge their learning of text as

well as younger adults do (Dunlosky, et al., 2006; Olin & Zelinski, 1997; Stine-Morrow, et al., 2006). Miles and Stine-Morrow (2004), however, reported lower relative accuracy for older adults than for younger adults. Although these studies largely support the conclusion that there are minimal age-related differences in relative accuracy, the absence of age effects may be less noteworthy considering that both groups show strikingly low levels of relative accuracy. In particular, the mean relative accuracy presented in Table 1 is only 0.22, which shows that people of all ages have considerable room for improvement in their ability to evaluate text learning. More important, low relative accuracy may conceal differences in the ability of older and younger adults to accurately judge their learning of text content.

Accordingly, in the present research, we explored age-related effects on the accuracy of a new kind of metacomprehension judgment – the term-specific judgment (discussed in detail below). These judgments are postdictions in which people evaluate how well they have just performed on tests over specific information from the text materials. They are similar to retrospective confidence judgments (or postdictions) that have been extensively explored using simple materials (e.g., paired associates) in the literature on metacognitive judgments (Dunlosky & Metcalfe, 2009). In previous text research, judgment accuracy is higher for postdictions of test performance over text materials than for standard predictions of test performance (Pierce & Smith, 2001), which has been referred to as the *postdiction superiority effect*. Likewise, the term-specific postdictions investigated in the present research produce high levels of relative accuracy for younger adults (Dunlosky, Rawson, & Middleton, 2005; Rawson & Dunlosky, 2007). Rawson and Dunlosky (2007) also reported that accuracy for these term-specific judgments was further improved through the use of providing a standard of evaluation.

A major goal of the present research was evaluating whether age-related differences arise in the accuracy of term-specific judgments. We also investigated whether providing a standard of evaluation (described in detail below) would improve the accuracy of older adults' judgments. The latter issue focuses on the degree to which older adults can attain high levels of judgment accuracy, which is invaluable for supporting efficient learning of text materials (Thiede, Anderson, & Theriault, 2003). In the remainder of this Introduction, we explain term-specific judgments in more detail and then describe the experiment reported in this study.

Term-Specific Judgments

Rawson and Dunlosky (2007) described conditions that supported higher levels of relative accuracy than are typically found in the metacomprehension literature, which involved using term-specific judgments (see also, Dunlosky et al., 2005). In particular, college students read short texts that included four highlighted definitions that were embedded within each text. After studying each text, they were asked to make a standard metacomprehension prediction about how well they would do on a test over the text material. More important, they were tested on each of the key-term definitions, with each answer scored as receiving “no credit,” “partial credit,” or “full credit.” By having participants make postdictive judgments about their responses (versus predicting future performance), one can more directly investigate how well they can evaluate the *quality* of their memory for text without being influenced by their theory of retention (Rawson, Dunlosky, & McDonald, 2002; Koriat, Bjork, Sheffer, & Bar, 2004). As expected from previous research, Rawson and Dunlosky (2007) showed that the relative accuracy of the term-specific judgments (M gamma correlation = .68) was considerably higher than for the standard predictions (M = .21). One reason for this postdiction superiority effect is that the participants must attempt to retrieve the correct answers on the test, and they can use the products of such retrieval attempts (e.g., how much is recalled) to evaluate how well they have learned a concept (cf. the retrieval hypothesis discussed by Pierce & Smith, 2001).

To date, no research has examined age differences in the accuracy of term-specific judgments. Will older adults' accuracy also benefit from making term-specific judgments? Given previous

research on aging and judgments of text learning (Table 1), one might expect that age-related deficits would not occur. Nevertheless, some evidence suggests that age-related deficits may occur in the accuracy of term-specific judgments. In particular, Miles and Stine-Morrow (2004) found an age-related deficit in accuracy (Table 1) for judgments about how well sentences had been learned. Asking participants to rate their memory for sentences is quite similar to the term-specific postdictions used here, except that they had participants make predictions soon after a sentence had been studied (e.g., with one intervening sentence between reading the target sentence and making the prediction), whereas the term-specific postdictions occur at a delay and can be based on *retrieval* of the definitions from long-term memory. Note, however, that Stine-Morrow et al. (2006) used a nearly identical procedure to Miles and Stine-Morrow (2004) in which older and younger adults predicted their memory for sentences. For the former study (Table 1), age-related differences did not arise, which suggests that age-related deficits in prediction accuracy may not be robust or may be moderated by currently unknown factors. Thus, the issue remains open as to whether accuracy for term-specific judgments will be as high for older adults as for younger adults.

Reducing Overconfidence By Providing a Standard of Evaluation

The relative accuracy for term-specific judgments is high. Unfortunately, despite good relative accuracy, the absolute accuracy of people's term-specific judgments (whether one's average confidence corresponds to the same absolute level of performance) can at the same time be rather low. Student's term-specific judgments have demonstrated substantial overconfidence (Dunlosky et al., 2005). Specifically, when younger adults make an incorrect response during the recall test (i.e., a commission error), they often believe they recalled the correct answer. For instance, Rawson and Dunlosky (2007) found that students said that they should receive "partial credit" or "full credit" 83 percent of the time after making a commission error. This overconfidence was reduced when participants were also shown the correct answers when scoring their own recall responses. That is, when a student was evaluating the quality of her recall attempt for the definition of a given term (e.g., What is a "devitalized marriage"?), she would have the correct definition of "devitalized marriage" to compare to her own recall. If overconfidence was due to an inability to evaluate the quality of an answer, providing this standard of evaluation (i.e., the correct answer) at the time of scoring should reduce overconfidence. Providing the standard did not entirely eliminate younger adults' overconfidence, but it did reliably reduce it.

Although providing a standard reduced younger adults' overconfidence in commission errors, it did not influence the relative accuracy of their postdictions (Rawson & Dunlosky, 2007), partly because relative accuracy for term-specific judgments was already high. Thus, our main interest in standards here was whether it reduced people's overconfidence. In particular, assuming older adults also produce high-confidence commission errors, whether the standard would also benefit their term-specific postdictions is an open question, and one that is important for potential interventions focused on improving older adults' self evaluation and self-regulated study. One might expect that older adults' overconfidence will not benefit as much from the use of a standard. In particular, note that even younger adults' overconfidence was not entirely eliminated by presenting correct definitions as standards for evaluation, which was likely due to the fact that even comparing one's answer to an external standard exceeds the capacity of the limited working-memory system required to process the definitions while making the comparison (for details, see Rawson & Dunlosky, 2007). Given that working memory capacity is reduced as we age (Salthouse, 1994; for an exception, see May, Hasher, & Kane, 1999), older adults may have greater difficulties comparing the external standard to their answer when making term-specific judgments. If so, presenting the correct definition may reduce overconfidence in commission errors less for older adults than for younger adults.

In summary, a main goal of the current experiment was to evaluate whether older adults can attain high levels of relative accuracy using term-specific judgments and whether the overconfidence of their judgments for commission errors would be reduced by providing the correct definition. To do so, we had participants: (a) read texts with embedded definitions, (b) make a standard metacomprehension prediction about how well they would perform on a test over the text content, (c) attempt to recall each definition, and (d) make a term-specific judgment after each recall attempt (which either did or did not include providing a standard of evaluation). Because of the differences between standard predictions and postdictive term-specific judgments¹, we do not directly compare their accuracy. Instead, we focused on age-related differences in the relative accuracy of term-specific judgments and whether providing standards reduced older adults' overconfidence for commission errors.

Method

Design and Participants

The experiment involved a 2 (Age: Older Adults vs. Younger Adults) X 2 (Standard: Presence vs. Absence of Standard) factorial design. Fifty-nine younger adults (M age = 19.6 years, SE = 0.5) and 55 older adults (M age = 68.9 years, SE = 0.9) participated in this experiment. Younger adults participated as part of a course requirement in General Psychology. Older adults were recruited from the greater Akron area via newspaper ads and personal referrals, and they received \$20 for participating. Providing a standard was a between-subjects variable; 28 older adults and 31 younger adults were randomly assigned to receive the correct definition (standard groups) when judging their recall responses, whereas the rest of the participants did not receive the standard (no standard groups).

Vocabulary performance was collected using a modified version of the Educational Testing Service Advanced Vocabulary Test, V-4, which contained 36 questions (Ekstrom, French, Harman, & Dermen, 1976). Performance was higher for older adults (M score = 22.8, SE = 1.0) than younger adults (M score = 12.1, SE = 0.5), $t(105) = 9.55, p < .001$, which is a standard outcome in the field. Older adults and younger adults reported being in good health, as reported on a scale of 1 to 5 (older M score = 3.9, SE = 0.1; younger M score = 4.1, SE = 0.2), $t(80) = 1.01$. Finally, to ensure that differences did not exist between the experimental groups, we conducted t-tests to compare the experimental groups within each age group on age, self-rated health, and vocabulary scores. The younger adult groups did not significantly differ on age, $t(57) = 0.76$, health, $t(57) = 0.23$, or vocabulary, $t(57) = 0.10$. The older adult groups did not significantly differ on age, $t(53) = 0.90$, or vocabulary, $t(46) = 1.78$. Self-reported health did differ between the older adult groups, $t(52) = 2.24, p < .05$, with those in the no-standard group reporting being in slightly better health (M = 4.15, SE = 0.14) than those in the standard group (M = 3.68, SE = 0.16).

Materials and Procedure

We used the same seven texts (1 sample text and 6 experimental texts) as Dunlosky et al. (2005) and Rawson and Dunlosky (2007). Each of the experimental texts included four embedded terms, along with their definitions (each of which included 2–4 idea units). For instance, for the text on “marriage,” key definitions were presented for “devitalized marriage,” “passive-congenial marriage,” “vital marriage,” and “total marriage.” Additional information about the texts is available in Appendix A and an example of one experimental text is presented

¹In the present study as well as in previous research, the two kinds of judgment not only differ on whether one is predicting eventual performance versus evaluating previous performance but also on knowledge of the specific test items and the proximity of the judgment to the actual test. Thus, although our results largely replicate the postdiction superiority effect, we do not discuss it further because our research was not designed to isolate the source(s) of it.

in Appendix B. Participants were instructed to read to “understand and remember” the material for an upcoming test. They were further instructed to focus on the key definitions, which appeared in all capital letters in the texts (for an example, see Appendix B).

Each participant completed the experiment individually via computer. After reading all instructions, participants read each text, one at a time. After reading each text, participants made a global judgment based on the following question, which was used in our previous research (e.g., Dunlosky et al., 2005): “How well will you be able to complete a test over this material?” Participants were allowed to type any value from 0 to 100, where 0 = “will not be able to complete a test over the materials,” 10 = “10% sure I will be able to complete a test over the materials,” and 100 = “certain I will be able to complete a test over the materials.”

After making a global judgment for a text, participants attempted to recall each definition from it. Participants were prompted to recall each definition using the following prompt: “Please recall the definition for {term inserted here}”. The order of terms from each text was randomized anew for each participant. The procedure was ideal for examining the accuracy of term-specific judgments, because testing immediately after study yields recall protocols that include correct recalls, commission errors and omission errors. Such variability in recall is critical for evaluating the accuracy of term-specific judgments, which were the focus of the present article. These judgments were collected as follows: Once participants had recalled as much as they could about a given definition, they made a term-specific judgment for the definition based on the following question: “Please score the definition you just typed by pressing the corresponding button. Would you grade your response as receiving ‘No Credit’, ‘Partial Credit’, or ‘Full Credit’?” For all participants, their recall response was presented below this prompt. For those in the standard group, this screen also included the correct definition. After making a term-specific judgment for a given term, participants repeated the same recall-judgment procedure for each of the remaining terms from the text. After all definitions had been judged for a given text, participants then read the next text and completed the judgment-recall cycle for each text until the end of the experiment.

Scoring

Each recall protocol was scored for the number of idea units correctly recalled. Moreover, each recall response was categorized into one of four categories: (a) omission errors (participant recalled nothing); (b) commission error (participant recalled some information, which was entirely incorrect); (c) partially correct (participant recalled some information, which was correct but did not include all the ideas in the sought-after definition); and (d) correct (participants recalled all the idea units in the sought-after definitions). Recall protocols were scored by two raters, who initially scored a subset of the protocols and were 92 percent consistent in their scoring of the categories described above (Cohen’s kappa = .89). As important, specific inconsistencies in this subset were minor – e.g., a difference in identifying a single idea unit – and were corrected through discussion.

Results

All effects declared as reliable have an associated *p*-value that is less than .05.

Relative Accuracy

Relative accuracy was calculated using Goodman-Kruskal gamma correlations between judgments and performance on the recall trials (for detailed discussion, see Nelson, 1984). All correlations were computed over a set of text items for each person and the mean within-person correlations were computed for all individuals within each age X standard group.

Term-specific judgments—To compute the correlations, we converted the categorical term-specific judgments to the following scale (cf. Dunlosky et al., 2005; Rawson & Dunlosky, 2007): “no credit” = 0, “partial credit” = 50, and “full credit” = 100. For each individual, this rating scale was correlated (across definitions) with the proportion of correct idea units recalled. Given that the gamma correlation is a nonparametric statistic, the conversion above will not influence the magnitude of the correlation – any monotonic conversion (where full credit is higher than partial credit, and both are higher than no credit) would yield the same value. Four of the older participants (3 from the no-standard group and 1 from the standard group) had indeterminate relative accuracy for term-specific judgments because they gave the same judgment magnitude to all items and had no variance in judgments. Accordingly, the analysis was conducted on the remaining 110 participants.

As shown in Figure 1, both older and younger adults obtained relatively high levels of accuracy when they made term-specific judgments. A 2 (age) X 2 (standard) analysis of variance (ANOVA) showed no main effect of age, $F(1,106) = 0.54$, partial $\eta^2 = .005$. The main effect of standard was not reliable, $F(1,106) = 2.26$, partial $\eta^2 = .021$. The reliable Age X Standard interaction, $F(1,106) = 4.77$, partial $\eta^2 = .043$, indicated that younger adults’ relative accuracy benefited from the standard whereas older adults’ relative accuracy did not. However, the effect of providing a standard improved younger adults’ relative accuracy to the level already achieved by the older adults without the standard. Given that previous research focusing on younger adults did not demonstrate an effect of providing a standard on relative accuracy (Rawson & Dunlosky, 2007), this slight increase in younger adults’ relative accuracy should be interpreted with caution. In any case, the main finding is that, like younger adults, older adults demonstrate high levels of relative accuracy when evaluating their recall of definitions for specific terms.

Because older adults had higher vocabulary scores, and given well-known effects of verbal ability on text comprehension (e.g., Hulstsch, Hertzog, & Dixon, 1984) and on accuracy of definitions repeated from texts (Veenman & Beishuizen, 2004), we conducted analyses controlling for the effects of vocabulary performance. Controlling on verbal ability still did not reveal a main effect of age, $F(1,100) = 1.81$, partial $\eta^2 = .018$, or of standard, $F(1,100) = 1.95$, partial $\eta^2 = .019$. The Age X Standard interaction, $F(1,100) = 4.75$, partial $\eta^2 = .045$, was still reliable.

Global judgments—Although global judgments were less central to the main aims of this research, we report their relative accuracy because they have been extensively investigated in previous research (for reviews, see Dunlosky & Lipko, 2007; Maki, 1998; Weaver et al., 1995). Accuracy for global judgments was calculated by correlating an individual’s global judgments for each text with recall performance across the texts (with recall performance for each text being computed as the mean proportion of idea units recalled across the four definitions). Eleven participants (2 from the older adult no-standard group, 7 from the older adult standard group, and 2 from the younger adult standard group) had indeterminate accuracy for global judgments because they gave the same global judgment to all texts. The following analyses were conducted on the remaining 103 participants.

The relative accuracy for global judgments was reliably lower for older adults ($M = 0.01$, $SE = 0.09$) than for younger adults ($M = 0.43$, $SE = 0.08$), $F(1,99) = 12.18$, partial $\eta^2 = .11$. Older adults’ relative accuracy for global judgments was not different from 0, $t(45) = 0.09$. A main effect for standard, $F(1,99) = 5.29$, partial $\eta^2 = .051$, showed that relative accuracy was better with standards ($M = 0.35$, $SE = 0.08$) than without them ($M = 0.08$, $SE = 0.08$). When standards were not available, relative accuracy for global judgments was not different than 0, $t(51) = 0.33$. The Age X Standard interaction was not reliable, $F(1,99) = 1.51$, partial $\eta^2 = .015$. The main effect of providing a standard may seem somewhat counterintuitive given that it was

provided after global judgments for a text had been made. Providing standards for the first few texts may have helped individuals recalibrate confidence on later texts, thereby improving global metacomprehension accuracy.

Given the aforementioned age differences in vocabulary, we also conducted an analysis controlling for the effects of vocabulary performance. The relative accuracy for global judgments was still reliably lower for older than younger adults, $F(1,92) = 4.12$, partial $\eta^2 = .05$. Neither the main effect for the standard, $F(1,92) = 2.97$, $p = 0.09$, partial $\eta^2 = .043$, nor the interaction were reliable, $F(1,92) = 2.70$, partial $\eta^2 = .028$.

Absolute Accuracy of Term-Specific Judgments as a Function of Recall Status

We also assessed whether absolute accuracy of judgments increased as a function of providing standards. Term-specific judgments that participants made for commission errors were of the most interest. As described above, for consistency with our previous research (e.g., Dunlosky et al., 2005; Rawson & Dunlosky, 2007), we converted the categorical term-specific judgments to the following scale: “no credit” = 0, “partial credit” = 50, and “full credit” = 100. A mean of these ratings was computed for each participant, and then group means were computed from the individual’s means.

The mean term-specific judgment for commission errors are presented in Figure 2. Note that any value above 0 indicates overconfidence; in such cases, participants were assigning credit to recall responses that were totally incorrect. As shown in Figure 2, providing a standard reduced participants’ overconfidence, such that the mean judgment was lower for commission errors when standards were provided than when they were not. A 2 (Age) X 2 (Standard) ANOVA revealed only a main effect of the standard, $F(1,109) = 17.65$, $p < .001$, partial $\eta^2 = .14$, with lower ratings from participants who received the standards. When vocabulary was used as a covariate in the ANOVA (as in previous analyses described above), the pattern of effects remained the same. Finally, we also conducted non-parametric tests to compare ratings made by the standard groups (standard versus no standard) within each age group, which also revealed that providing standards led to the use of lower ratings for commission errors for older and younger adults, $ps < .05$.

Ratings given to the other classes of item recall outcomes (i.e., partially correct and entirely correct responses) are reported in Table 2 (the median judgment for omission errors was 0, so we omitted them from the table). Providing standards influenced only ratings for correct responses, and age had no influence on rating behavior. These observations were confirmed by a 2 (Age: younger vs. older adults) X 2 (Standard vs. no standard) X 2 (Item Class: partially correct response vs. correct response) ANOVA. Nine participants were dropped from the analysis because they did not have any entirely correct answers. A reliable main effect occurred for Item Class, $F(1,101) = 105.7$, $p < .001$, partial $\eta^2 = .51$, qualified by a reliable Item Class X Standard interaction, $F(1,101) = 9.6$, $p < .01$, partial $\eta^2 = .09$. The latter interaction indicated that ratings were higher for correct responses when standards were presented than when they were not, $t(93) = 2.19$, whereas providing standards did not influence ratings for partially correct answers, $t(103) = 0.02$. Thus, in the present case, not only did providing standards help to reduce overconfidence (Figure 2) but doing so also improved participants’ evaluations of their correct responses. As important, the main effect and interactions involving age were not reliable, indicating that providing standards influenced both groups’ judgments similarly.

Discussion

A main conclusion from this research is that both older and younger adults can obtain relatively high levels of judgment accuracy if they make term-specific postdictions of their text learning. Moreover, providing the correct definitions as standards for evaluation reliably reduced the

overconfidence of these judgments for both age groups, and doing so also helped them to evaluate the quality of their correct responses (albeit the benefit of standards for evaluating correct responses was not found by Rawson & Dunlosky, 2007). The implications are clear: Both younger and older adults can accurately evaluate the quality of their recall for key-term definitions when they (a) attempt to recall the specific definitions prior to making their judgments and (b) compare their response to the correct answer.

These outcomes and conclusions should be considered in the context of three issues. First, although age equivalence in judgment accuracy is the norm in studies of judgment accuracy for text materials (Table 1), one might have expected that the accuracy of term-specific judgments would be *greater* for older than younger adults, because it seems reasonable that accurately evaluating recall of definitions from texts is aided by one's verbal ability (e.g., Veenman & Beishuizen, 2004). In the present case, older adults' vocabulary was better than younger adults' vocabulary, yet older adults' judgment accuracy was not better. Second, for both age groups, although providing standards reduced overconfidence, doing so did not eliminate overconfidence completely (Figure 2), which itself indicates that even with the correct definition immediately available people still have difficulties evaluating their responses! Discovering techniques to entirely eliminate overconfidence is an important agenda for applied research.

Third, the outcomes are relevant to improving people's self-regulated learning that is needed to learn various kinds of content, such as new concepts relevant to job training, key information about new hobbies, and core concepts in classes. Given that older adults are staying active late in life and many are taking college classes, we find the applications to learning from standard textbooks particularly notable. Many chapters in textbooks end with a list of vocabulary, e.g., what are the definitions of "devitalized marriage," "passive-congenial marriage," and so forth. Anecdotally, after performing poorly on a test, some students claim that they were sure they knew most of the class materials. Ironically, this overconfidence may in part reflect the students' misuse of the end-of-chapter tests in textbooks, especially if they do not check their answers against the correct ones (cf. the current no standard group). But even if they do check their answers, their evaluations may still be overconfident (Figure 2). At the least, students of all ages should be instructed how to evaluate their learning of key-term definitions by recalling their definitions and explicitly comparing them to the correct definitions.

Concerning overconfidence, one intriguing issue is the degree to which overconfidence remains when people are asked to evaluate the quality of others' recall. Put differently, will overconfidence be as extreme when someone is grading another person's performance than when grading his or her own performance? Even with standards, people may be more overconfident when evaluating their performance, because their evaluations could be inappropriately influenced by idiosyncratic knowledge, such as how easily they retrieved a particular response and beliefs about whether they actually knew part of the correct answer but merely failed to report it during the test. Such idiosyncratic knowledge could not influence a person's evaluations of another's recall performance. Accordingly, an intriguing direction for future research involves having people self-evaluate their performance (as in the present study) and having them evaluate other's performance, which could reveal other sources of poor calibration in people's self assessments.

Although our main focus was on the accuracy of people's term-specific judgments, the present age-related difference in the accuracy of global predictions were considerable and deserve further scrutiny. In particular, the relative accuracy of older adults' global judgments was not reliably above the chance level of a zero gamma correlation, whereas younger adults' accuracy was near +.40, collapsed across groups. We did not predict this deficit a priori. When compared to other studies conducted on aging and metacomprehension accuracy (Table 1), it is notable

that the one other investigation reporting age-related deficits in global judgment accuracy involved predicting memory for individual sentences (Miles & Stine-Morrow, 2004). Given that age-related differences are likewise not always obtained for predicting sentence memory (e.g., Stine-Morrow et al., 2006), the age-related deficits found in the current study are inconsistent with a generic age deficit in metacomprehension. Instead, age differences may arise from several factors specific to text comprehension materials, the types of persons recruited into the study, or testing contexts. Thus, understanding whether and when older adults show differences in metacomprehension will apparently require further systematic research. What is particularly intriguing is that the age deficits in metacomprehension accuracy of global predictions for text recall were accompanied, in the same participants, by no age differences in term-specific postdictive judgments. Such outcomes might indicate that older adults could, in principle, spontaneously access specific information about the text content that would boost metacomprehension accuracy when making global predictions, but do not do so. Instead, they may construct global predictions using cues that dilute their accuracy, such as memory self-efficacy or prior familiarity with the text domain (Cavanaugh, Feldman, & Hertzog, 1998).

Regardless of whether age-related deficits occur in metacomprehension predictions, we find it notable that global accuracy for both older and younger adults' predictions tended to be low (Table 1). Thus, we argue that an alternative – and perhaps more promising – direction for research in this area is to discover techniques that will allow adults of all ages to obtain high levels of accuracy at evaluating their learning of text materials. To this end, we investigated term-specific judgments. Our main conclusion is that if older and younger adults want to accurately evaluate how well they have learned important key-term definitions, they can do so successfully by making term-specific judgments that are supported by presentation of the correct answers.

Acknowledgments

This research was supported by a grant from the National Institute on Aging, one of the National Institutes of Health (R37 AG13148).

References

- Cavanaugh JC, Feldman J, Hertzog C. Memory beliefs as social cognition: A reconceptualization of what memory questionnaires assess. *Review of General Psychology* 1998;2:48–65.
- Dunlosky J, Baker JMC, Rawson KA, Hertzog C. Does aging influence people's metacomprehension? Effects of processing ease on judgments of text learning. *Psychology & Aging* 2006;21:390–400. [PubMed: 16768583]
- Dunlosky J, Lipko AR. Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science* 2007;16:228–232.
- Dunlosky, J.; Metcalfe, J. *Metacognition*. Beverly Hills, CA: SAGE; 2009.
- Dunlosky J, Rawson KA, Middleton EL. What constrains the accuracy of metacomprehension accuracy? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language* 2005;52:551–565.
- Ekstrom, RB.; French, JW.; Harman, HH.; Dermen, D. *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service; 1976.
- Hultsch DF, Hertzog C, Dixon RA. Text recall in adulthood: The role of intellectual abilities. *Developmental Psychology* 1984;20:1193–1209.
- Koriat A, Bjork RA, Sheffer L, Bar SK. Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General* 2004;133:643–656. [PubMed: 15584811]
- Maki, RH. Test predictions over text material. In: Hacker, DJ.; Dunlosky, J.; Graesser, AC., editors. *Metacognition in Educational Theory and Practice*. Mahwah, NJ: LEA; 1998. p. 117-144.

- May CP, Hasher L, Kane MJ. The role of interference in memory span. *Memory and Cognition* 1999;27:759–767.
- Miles JR, Stine-Morrow EAL. Adult age differences in self-regulated learning from reading sentences. *Psychology & Aging* 2004;19:626–636. [PubMed: 15584788]
- Nelson TO. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin* 1984;95:109–133. [PubMed: 6544431]
- Olin JT, Zelinski EM. Age differences in calibration of comprehension. *Educational Gerontology* 1997;23:67–77.
- Pierce BH, Smith SM. The postdiction superiority effect in metacomprehension text. *Memory & Cognition* 2001;29:62–67.
- Rawson K, Dunlosky J, McDonald S. Influences of metamemory on performance predictions for text. *Quarterly Journal of Experimental Psychology* 2002;55A:505–524. [PubMed: 12047056]
- Rawson KA, Dunlosky J. Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology* 2007;19:559–579.
- Salthouse TA. The aging of working memory. *Neuropsychology* 1994;8:535–543.
- Stine-Morrow EAL, Shake MC, Miles JR, Noh SR. Adult age differences in the effects of goals on self-regulated sentence processing. *Psychology & Aging* 2006;21:790–803. [PubMed: 17201498]
- Thiede KW, Anderson MCM, Theriault D. Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology* 2003;95:66–73.
- Thiede KW, Dunlosky J, Griffin T, Wiley J. Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2005;31:1267–1280.
- Veenman MVJ, Beishuizen JJ. Intellectual and metacognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learning and Instruction* 2004;14:621–640.
- Weaver, CA., III; Bryant, DS.; Burns, KD. Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In: Weaver, CA., III; Mannes, S.; Fletcher, CR., editors. *Discourse comprehension: Essays in honor of Walter Kintsch*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1995. p. 177-193.

Appendix A

Text Information

Text Title	Text Characteristics	Terms
Types of Evaluation (sample text – not used in the experiment)	110 words	formative evaluation
	7 sentences	summative evaluation
	15.7 words/sentence	
	Flesch-Kincaid Grade Level = 13.0	
Marriage	271 words	devitalized marriage
	13 sentences	passive-congenial marriage
	20.8 words/sentence	vital marriage
	Flesch-Kincaid Grade Level = 16.0	total marriage
Energy Use	276 words	basal metabolism
	15 sentences	thermic effect of food
	18.4 words/sentence	adaptive thermogenesis
	Flesch-Kincaid Grade Level = 13.2	direct calorimetry
Gestures	271 words	emblems
	15 sentences	illustrators
	18.0 words/sentence	self-adaptors

Text Title	Text Characteristics	Terms
Measurement	Flesch-Kincaid Grade Level = 10.5	alter-adaptors
	277 words	nominal measurement
	14 sentences	ordinal measurement
	19.7 words/sentence	interval measurement
U.S. Government	Flesch-Kincaid Grade Level = 11.3	ratio measurement
	284 words	“layer cake” federalism
	13 sentences	9 th and 10 th amendments
	21.6 words/sentence	article 6
Investment	Flesch-Kincaid Grade Level = 15.4	“picket fence” federalism
	279 words	return on investment
	16 sentences	risk
	17.4 words/sentence	effective business
	Flesch-Kincaid Grade Level = 10.0	efficient business

Appendix B

Example of an Experimental Text

TITLE: ENERGY USE

The body uses energy for three major purposes, basal metabolism, physical activity, and the thermic effects of food. To a lesser extent, it also uses energy for adaptive thermogenesis. **BASAL METABOLISM** is the minimal energy required for a resting, awake body to stay alive. This requires about 60% to 70% of total energy use by the body, and includes maintaining a heartbeat, respiration, and temperature. The amount of energy used for basal metabolism depends primarily on lean body mass. Physical activity increases energy expenditure beyond our basal energy needs by as much as 25% to 40%. Unlike basal metabolism, energy expenditure from physical activity varies widely among people. The **THERMIC EFFECT OF FOOD** refers to the fact that the body uses energy to digest, absorb, and further process food nutrients. The energy cost of this thermic effect is analogous to a sales tax, in which you are taxed about 5% to 10% for the total energy you eat. Finally, **ADAPTIVE THERMOGENESIS** is when the body expends energy to produce heat in response to a cold environment or as a result of overfeeding. Though adaptive thermogenesis probably does not play a major role in weight regulation, it appears to represent a small portion of energy use. The amount of energy a body uses can be measured by direct calorimetry. **DIRECT CALORIMETRY** measures the amount of heat that emanates from a body. Usually, a person is put into an insulated chamber, and body heat released raises the temperature of a layer of water surrounding the chamber. The difference in the temperature of the water before and after the person entered the chamber indicates the number of calories expended.

Figure 1.

Relative accuracy of term-specific judgments. Participants in the standard group received the correct answer while evaluating their response; participants in the no standard group were not provided with the correct answer. Bars represent standard errors of the mean.

Figure 2.

Mean term-specific judgments (self-score magnitude) given to commission errors. Participants in the standard group received the correct answer while evaluating their response; those in the no standard group were not provided with the correct answer. Bars represent standard errors of the mean.

Table 1

Relative accuracy for global predictions of test performance for text learning

	Younger adults	Older adults
Olin & Zelinski (1997)	.10	.07
Miles & Stine-Morrow (2004)	.31	.18
Dunlosky, Baker, et al. (2006)		
Experiment 1	.16	.23
Experiment 2	.26	.34
Stine-Morrow et al. (2006)		
Experiment 1	.28	.29

Note: Entries are means across intra-individual correlations between global metacomprehension predictions and test performance for text materials.

Table 2

Mean Self-Score Ratings for Other Item Classes

	Partially Correct		Correct	
	<i>M</i>	(<i>SE</i>)	<i>M</i>	(<i>SE</i>)
Older Adults				
No Standard	72.6	(3.8)	81.2	(3.8)
Standard	70.2	(3.4)	87.1	(3.4)
Younger Adults				
No Standard	69.5	(3.6)	80.3	(3.5)
Standard	69.3	(3.2)	88.6	(3.2)