1-1-2005

Understanding the Delayed-Keyword Effect on Metacomprehension Accuracy

Keith Thiede
Boise State University

John Dunlosky
Kent State University

Thomas D. Griffin
University of Illinois at Chicago

Jennifer Wiley
University of Illinois at Chicago

DOI: 10.1037/0278-7393.31.6.1267

Understanding the Delayed-Keyword Effect on Metacomprehension Accuracy

Keith W. Thiede
Boise State University

John Dunlosky
Kent State University

Thomas D. Griffin & Jennifer Wiley
University of Illinois at Chicago

Abstract

The typical finding from research on metacomprehension is that accuracy is quite low. However, recent studies have shown robust accuracy improvements when judgments follow certain generation tasks (summarizing or keyword listing), but only when these tasks are performed at a delay rather than immediately after reading (Thiede & Anderson, 2003; Thiede, Anderson & Therriault, 2003). The delayed and immediate conditions in these past studies confounded the delay between reading and generation tasks with other task lags, such as the lag between multiple generation tasks and the lag between generation tasks and judgments. The first two experiments disentangle these confounded manipulations and provide clear evidence that the delay between reading and keyword generation is the only lag critical to improving metacomprehension accuracy. The third and fourth experiments show that not all delayed tasks will produce improvements and suggest that delayed generative tasks provide diagnostic cues about comprehension that are necessary for improving metacomprehension accuracy.

Models of self-regulated learning describe learning as a dynamic process in which a learner monitors progress toward a learning goal and uses this information to regulate study (e.g., Metcalfe & Kornell, 2003; Nelson & Narens, 1990; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998). To make effective decisions about what to study or how long to study, the learner must accurately monitor his or her learning so as to identify materials that will benefit most from restudy. This link between accurate monitoring and learning has been empirically supported by recent research across a variety of domains (for a review, see Dunlosky, Hertzog, Kennedy, & Thiede, in press). In the present research, we focus in general on the accuracy of people's judgments of text learning—or metacomprehension accuracy—and in particular on why the delayed generation of keywords improves accuracy. To motivate our current approach to investigating these issues, we first briefly review the standard method and modal outcome from the metacomprehension literature and then describe the *delayed-keyword effect*.

To estimate the accuracy of people's judgments of text learning, participants in a typical experiment would read multiple texts, which range from about 200 to 400 words each. Sometime after studying a given text, a participant predicts how well he or she will perform on a test over the content of the text. After all texts have been read and judged, the participant takes a test of comprehension for each text. Metacomprehension accuracy is then estimated by correlating each participant's judgments with his or her own test performance. Higher correlations are indicative of better accuracy. In general, metacomprehension accuracy is notoriously low, with mean accuracy around .25 (for reviews, see Lin & Zabrucky, 1998; Maki, 1998a). Recent studies have demonstrated dramatic increases in metacomprehension accuracy when readers are asked to generate keywords (Thiede, Anderson, & Therriault, 2003) or summaries (Thiede & Anderson, 2003) following a delay after reading the text. The source of the improvements due to delayed keyword generation is explored in detail in the current article.

The Delayed-Keyword Effect

Thiede et al. (2003) demonstrated that metacomprehension accuracy could be dramatically increased by having participants read texts and then generate keywords that captured the essence of each text prior to judging comprehension for it. Importantly, the effect of generation on metacomprehension accuracy was moderated by the timing of keyword generation. That is, participants who read a text and immediately generated keywords for that text were no more accurate (mean accuracy = .23) than participants who did not generate keywords (mean accuracy = .36). By contrast, participants who read all the texts and then one by one generated keywords for each text after a delay had superior accuracy (mean accuracy = .70). The primary purpose of Thiede et al. was to empirically establish the link between metacomprehension accuracy and reading comprehension. They showed that improving metacomprehension accuracy led to more effective regulation of study and enhanced comprehension. However, they did not evaluate alternative explanations for the delayed-keyword effect. Thus, the question remains, Why does delayed generation of keywords improve the accuracy of people's metacomprehension judgments?

To answer this question, one must first scrutinize the differences in procedures for the immediate-keyword group and the delayed-keyword group. These groups differed on three factors (see Table 1). First, they differed on the lag between when a text was read and when keywords were generated (the *Reading-Keyword* or *RK* lag). For the immediate-keyword group, keywords were generated immediately after reading a text. By contrast, for the delayed-keyword group, keywords for a text were generated well after it had been read, because all texts were read prior to keyword generation. That is, this longer RK interval for each text of the delayed group was filled with reading the remaining texts and/or generating keywords for the other texts. The RK lag was used to derive the names of the groups in Thiede et al. (2003), but importantly, this lag was not the only factor that differed between groups. Second, the immediate-keyword and delayed-keyword groups also differed in the lag between generating keywords for one text and another (the *Keyword-Keyword* or *KK* lag). In particular, for the immediate-keyword group, keyword generation was spaced, because a text was read between the generation of keywords for each of the texts. By contrast, for the delayed-keyword group, keyword generation for each text was successive. That is, after participants had read all the texts, they generated keywords for one text and then immediately generated them for the next, and so forth, until keywords had been generated for all texts. Finally, the groups also differed in the lag between generating keywords and judging comprehension (the *Keyword-Judgment* or *KJ* lag). For the immediate-keyword group, the KJ lag was relatively long, because all texts were read and all keywords were generated before any of the judgments were made. In this case, the time between keyword generation and judging comprehension was filled by reading and generating keywords for the remaining texts. Of course, this lag was substantial for the early texts in the list. By contrast, for the delayed-keyword group, the KJ lag was relatively short. For this group, participants generated keywords for each text (which took relatively little time) and then made judgments for each text. Thus, the immediate group and delayed group differed on all three factors in Thiede et al., and hence it is not clear which factor(s) were responsible for the delayed-keyword effect.

In the present research, we systematically manipulate the factors above to estimate their relative contribution to the delayed-keyword effect. Most important, identifying which factor(s) produce the effect is critical for understanding why delayed generation of keywords boosts metacomprehension accuracy because each factor is related to a specific theoretical mechanism. In the remainder of the Introduction, we discuss these mechanisms and how each is related to a given factor. Afterwards, we provide a brief overview of all four experiments presented here and how they achieve our primary goal of understanding the delayed-keyword effect.

Alternative Mechanisms of the Delayed-Keyword Effect

Forgetting keyword-relevant information. This explanation points to the KJ lag as critical for improving metacomprehension accuracy. Generating keywords may improve metacomprehension accuracy by producing cues that can be used to judge comprehension. If the cues produced during keyword generation are not accessible at the time judgments of comprehension are made, then they could not influence metacomprehension accuracy. Moreover, the greater the time between generating keywords and making judgments, the more likely it is that a person will forget the cues produced during keyword generation. In Thiede et al. (2003), the KJ lag was much longer for the immediate-keyword group than for the delayed-keyword group, and it seems plausible that this difference could be responsible for the differential metacomprehension accuracy between these two groups. According to this explanation, if the lag between generating keywords and judging comprehension is minimized, metacomprehension accuracy should improve.

Relative comparison of keyword-relevant information. In Thiede et al. (2003), the KK lag may have affected one's ability to make relative judgments. For the delayed-keyword group, keyword generation for each text was successive. Generating keywords for texts one after another may have prompted participants in the delayed-keyword group to evaluate the learning of each text relative to the others, which could subsequently support better discrimination between texts and in turn increase metacomprehension accuracy. By contrast, for the immediate-keyword group, keyword generation for any two texts was spaced by the reading of one text, which may have increased the likelihood that participants judged each text individually. Spaced keyword generation would make it difficult to judge one text relative to another, thus failing to provide information on relative understanding which is the benchmark for relative accuracy that has been the focus of metacomprehension research (for a similar argument concerning feeling-of-knowing judgments, see Nelson & Narens, 1980). According to this explanation when the lag between generating keywords for one text to another is minimized (as when keyword generation is successive), metacomprehension accuracy should improve.

Accessing the situation-model or long-term memory for text. This explanation is that the delay between reading a text and generating keywords, the RK lag, is critical for improving metacomprehension accuracy. A delayed generation task will produce cues that are more predictive because the processing involved in the delayed generation task and comprehension test are similar in two ways. First, as both tasks are delayed, they both require accessing the long-term memory (LTM) text representation, whereas the immediate generation task makes use of representations that are still active in short-term memory (STM) or working memory (WM). Second, both tasks will tend to access the situation-model level of representation to a greater extent than immediate generation. Several findings have suggested that delays after reading decrease the accessibility of the exact words or discrete ideas that are read, but increase accessibility to the situation model (the conceptual gist of the text, the relations among its ideas, and what it implies). Kintsch, Welsch, Schmalhofer, and Zimny (1990) have shown that memory for the lexical and textbase representation of text decays more rapidly than that of the situation model. They found a rapid decay in recognition accuracy for verbatim sentences, an intermediate decay for paraphrased sentences, and a very slow decay for plausible inferences based on the situation model of the text (see also Fletcher & Chrysler, 1990; Schmalhofer & Glavanov, 1986). Thus, when a delay is inserted between reading a text and generating a list of keywords, it may decrease the likelihood that the reader uses their memory for the surface features or the textbase and increase the chances that they access their situation model of a text. Likewise, performance on delayed comprehension tests, especially assessments involving inference verification, will be influenced by the same situation models that readers accessed during delayed generation.

According to both of these explanations for the delayed-keyword effect (access to the situation model or to LTM), the delay after reading will produce diagnostic cues during keyword generation by affecting what representation is accessed, which in turn is responsible for improvements in accuracy of judgments of a delayed comprehension test.

Experimental Overview

In the Experiments 1 and 2, we competitively evaluated the three explanations above (forgetting of relevant information, relative comparison, and access to situation model or LTM) by estimating the relative contribution of the three factors to the delayed-keyword effect. Note, however, that the proposed mechanisms are not exhaustive for a given factor. For instance, generating keywords successively (versus spaced) may not only facilitate making relative judgments, but it may also reduce forgetting of the keywords. Accordingly, the abovementioned discussion of mechanisms was meant to identify potential factors, which is necessary to narrow the field of possible mechanisms. To foreshadow, Experiments 1 and 2 suggests that the RK lag is the most critical factor for improving metacomprehension accuracy.

It is not possible to create a simple 3-way factorial design where each lag varies independently of the others without confounding a whole new set of manipulations between groups. There are inherent constraints that stem from the fact that the three lags all overlap in one of their components (i.e., RK, KK, KJ), so any manipulation could at best vary one lag independently of the other two, while the other two remain confounded. Isolating different lags across different studies allows for testing the importance of each lag in producing improvements. For both Experiments 1 and 2, the original immediate-keyword and delayed-keyword conditions were included, plus each was modified to create two additional groups. In Experiment 1, the modifications were designed to isolate the KJ lag, and in Experiment 2, to isolate the RK lag. A third experiment isolating the KK lag was not necessary, because the obtained results of these experiments provided clear evidence that only the RK lag affects metacomprehension accuracy.

Given the observed importance of delaying keyword generation to improving metacomprehension accuracy, another question arises: What features of the generating keywords task are necessary for obtaining the effect? To provide a more comprehensive analysis of the delayed-keyword effect, we explore this issue in Experiments 3 and 4. In particular, we evaluate whether *generating* keywords (versus thinking about the text or merely reading keywords) is necessary for demonstrating the effect of delay on the metacomprehension accuracy.

Experiment 1

In this experiment, we manipulated the lag between generating keywords and judgments (KJ lag) independently of the RK and KK lags. The original (Thiede et al., 2003) immediate-keyword and delayed-keyword conditions were retained, but a modified version of each was added. In both cases, the modification completely eliminated any KJ lag by always collecting judgments immediately following the keyword generation for the respective text (see Table 1). This modification did not change the RK and KK lags, thus they remained confounded as with the original conditions. For the sake of simplicity, conditions will be referred to only by their RK and KJ lags. The RK-KK confound will be revisited in the discussion of the findings. If the KJ lag is the critical factor, there should be a main effect of modification and metacomprehension accuracy should be greater for the two modified KJ-no lag groups than for the KJ-lag groups. If the KJ lag does not affect metacomprehension accuracy, then these new conditions will parallel the original conditions and metacomprehension accuracy will be greater for both RK-delayed conditions, with or without a KJ lag.

Method

Design and Subjects. The time between reading and generating keywords (RK lag: immediate versus delayed), and the time between generating keywords and making judgments (KJ lag: lag versus no lag) were manipulated between subjects. The kind of test question (inference vs. detail) was manipulated within participants.

One hundred-thirty two students enrolled in a psychology course at the University of Illinois at Chicago participated as part of a subject pool and were randomly assigned to four groups by order of appearance. Participants were treated in accordance with the ethical standards of the American Psychological Association.

Materials. The texts and tests were the same as those used in Thiede et al. (2003). Texts were seven expository texts adapted from encyclopedia articles on different topics (i.e., communication styles of men versus women, the effects of alcohol on sleep, experimental design, stress, intelligence and IQ tests, Norse

settlements, and WWII naval warfare). The texts ranged in length from 1118 words to 1595 words, and ranged in Flesch-Kincaid readability scores from 9.5 to 12.0. The 12 test questions per text consisted of six questions that required an inference to answer correctly and six questions that required only memory of details about the text content (for an example of a text and test items see Appendix A).

Procedure and lag manipulation. An overview of the experimental procedure is presented in Table 1. All participants were instructed that they would read texts, rate their comprehension for each, and then answer test questions for each text. Participants were also instructed that they would be asked to type a list of keywords that captured the essence of a text. These instructions included an example of keywords (i.e., for a text on the Titanic one might write: iceberg, shipwreck, tragedy…), but there was no formal training on how to generate keywords. To be consistent with Thiede et al. (2003), following the instructions, participants were asked to make an ease-of-learning (EOL) judgment for each text topic[1]. After making an EOLs for each text, participants read the sample text, rated their comprehension of the text, and answered the sample questions. Participants were encouraged to ask questions about the procedure during the practice trial.

For the critical trials, the order of text presentation was randomized anew for each participant. For all six texts, all participants performed the same tasks: reading, generating keywords, judging comprehension, completing comprehension tests. Texts were presented paragraph by paragraph, presentation time was controlled by the participant[2]. For the keyword generation task, participants were presented with the title of a text and were instructed to type five keywords that captured the essence of the text. The comprehension judgments were prompted with the title of the text at the top of the screen and the query, "How well do you think you understood the passage whose title is listed above? 1 (very poorly) to 7 (very well)." After judging their comprehension of the last text, participants answered 12 4-alternative multiple-choice questions for each text; six test questions were inference questions designed to assess knowledge of a person's situation model (for a detailed description of various representations of texts and how to assess knowledge of these representations see Kintsch, 1988; Graesser, Millis, & Zwaan, 1997) and six questions were designed to assess memory of a text. The inference questions were written so that information from non-adjacent paragraphs were required to correctly answer questions. The texts were rated for comprehension and tested in the same order as they were presented for reading.

For each group, after answering the last test question, participants were presented the number of questions they correctly answered over all six tests. That is, they received feedback regarding overall performance, but not text specific feedback.

The difference between groups was the procedural order of the tasks creating a different pattern across the three types of lags (see Table 1). The RK-delayed groups generated keywords only after reading all six texts, whereas the RK-immediate groups generated keywords immediately after reading the corresponding text. The original groups with a KJ lag made all their judgments only after generating keywords for all six texts, whereas the modified groups without a KJ lag made each judgment immediately following keyword generation for that text.

The length of the delays or lags are defined by the number of intervening reading and/or keyword tasks. The RK-delayed groups had five tasks between reading a text and generating keywords for that text, and zero tasks between generating keywords for one text and the next. By contrast, the RK-immediate conditions had zero tasks between reading and generating keywords and one task between generating keywords for one text and the next.

Orthogonal to these differences was the KJ lag. The modified groups (KJ-no lag) had no tasks between generating keywords and judging comprehension for a text. The remaining two groups (KJ-lag, the immediate-keyword and delayed-keyword groups in Thiede et al.) had either 10 or 5 intervening tasks between generating keywords and judging comprehension for the first text. That is, the KJ lag for the original immediate-keyword group was twice as long (10 tasks) as the lag for the original delayed-keyword group (5 tasks). This differing size of the KJ lag is confounded with both other lags, so direct comparison between these two groups is only interesting in terms of replicating the original finding. The test of the KJ

lag effect is the comparison between these original groups and the new KJ-no lag groups that are the same in their RK and KK lags.

Results and Discussion

For each dependent variable, analyses consisted of a 2 (RK-delayed versus RK-immediate) X 2 (KJ-no lag versus KJ lag) analysis of variance.

Test performance and metacomprehension judgments. Because metacomprehension accuracy describes the relation between comprehension ratings and performance on a test of reading comprehension, descriptive analyses of these variables are reported first. For each participant, we computed the median proportion of correct test response and comprehension rating across the six texts. We used the median because it is the recommended measure of central tendency for small sets of scores where extreme scores may have an undue influence on the mean (Gravetter & Wallnau, 1999). The mean of the medians was then computed across participants for each group. Test performance and comprehension ratings (Table 2) did not differ across groups. Neither the main effects nor the interaction were significant for test performance [for inference questions, all $F(1, 128)$s < 1, MSE = .02, p > .10; for detail questions, all $F(1, 128)$s < 1, MSE = .02, p > .10], or for comprehension judgments [all $F(1, 128)$s < 1, MSE = 1.08, p > .10].

Metacomprehension accuracy. As in previous studies (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987; Maki & Serra, 1992; Thiede et al. 2003), metacomprehension accuracy was operationalized as a Goodman-Kruskal gamma correlation between a participant's comprehension judgments and test performance across texts. For each participant, we computed two gamma correlations. We computed metacomprehension accuracy separately for the six test questions that required inference and for the six test questions that required memory of details. The mean of these intra-individual correlations was then computed across participants for each group separately for each kind of test. No participants had indeterminate gamma correlations.

When test performance was measured by inference questions, metacomprehension accuracy (Figure 1) was affected by the RK lag, $F(1, 128) = 12.86$, MSE = .16, p < .001, effect size = .09, but neither the effect of the KJ lag, $F(1, 128)$ < 1, p > .10, nor the interaction were significant, $F(1, 128)$ < 1, p > .10. As evident from inspecting Figure 1, accuracy was substantially higher for both RK-delayed groups than for the RK-immediate groups. The RK delay effect occurred even when comparing the two modified groups that had no KJ lags. Further, the KJ-no lag groups showed the same levels of metacomprehension accuracy as their respective KJ-lag groups.

When test performance was measured by detail questions, neither the RK lag nor the KJ lag affected metacomprehension accuracy, and the interaction was not significant, all $F(1, 128)$s < 1, MSE = .23, p > .10. Mean metacomprehension accuracy (and Standard Error of the Mean) for the RK-immediate-KJ lag, RK-immediate-KJ-no lag, RK-delayed-KJ lag, RK-delayed-KJ-no lag were respectively, .29 (.09), .37 (.10), .40 (.06), and .38 (.08). Thus, accuracy for predicting memory of details was not affected by either manipulation.

These results replicate the delayed-keyword effect reported by Thiede et al. (2003). More important, they implicate the RK lag manipulation as critical to producing the difference in metacomprehension accuracy. By contrast, given that there were no differences in metacomprehension accuracy when the KJ lag was longer versus shorter and that the RK-delayed groups were more accurate regardless of KJ lag, the KJ lag seems to be neither necessary nor sufficient in affecting metacomprehension accuracy. However, it is important to recall that the RK lag and the KK lag were confounded in this design. Thus, although we can rule out the influence of the KJ lag, we cannot conclude whether the observed improvements in metacomprehension accuracy resulted directly from the RK delay or from the successive keyword generations. Thus, Experiment 2 was designed to test for independent effects of the KK and RK lags.

Experiment 2

In this experiment, we manipulated the delay between reading and generating keywords independently of the lag between each keyword generation task. Again, we retained the original immediate-keyword and delayed-keyword groups from Thiede et al. (2003). Two modified versions of these conditions were created by adding filler texts between each keyword generation (see Table 1). The modified groups each had longer KK lags than their respective original groups, with the modified immediate group having the longest KK lag and the modified delayed group having an intermediate KK lag equal to the original immediate group. Thus, if the KK lag is a critical factor, then the modification will *reduce* metacomprehension accuracy, and the modified delayed group will have accuracy roughly equal to the original immediate group. In addition to affecting the KK lag, the modification also increased the length of the RK delay. As a result, the modified delayed group had the longest RK delay and the modified immediate group had an intermediate RK delay. In contrast to the predicted effects of the KK lag, if the RK delay is the critical factor then metacomprehension accuracy will improve even in the modified delayed condition and may be higher for the modified immediate group compared to the original immediate-keyword group.

Method

Design and Subjects. The time between reading and generating keywords (RK lag: immediate versus delayed), and the time between generating keywords for one text and another (KK lag: no filler versus added filler) were manipulated between subjects. Type of test was not manipulated in this experiment. That is, only inference questions were used in this experiment because the delayed-keyword effect occurred only with inference tests and the purpose of this investigation is to better understand this effect.

One hundred students enrolled in a psychology course at the University of Illinois at Chicago participated as part of a subject pool and were randomly assigned to four groups by order of appearance. Participants were treated in accordance with the ethical standards of the American Psychological Association.

Materials. The critical texts were the same as in Experiment 1. In the KK-spaced groups, spacing between keyword generation for one text and another was created by inserting a filler text. The filler texts were approximately 1,000 word expository texts on a variety of unrelated topics.

Procedure and lag manipulation. An overview of the experimental procedure is presented on the bottom of Table 1. As in Experiment 1, all participants received the instructions and completed the procedure for a sample text. For the critical trials, the order of text presentation was randomized anew for each participant. Again, for all six texts, all participants performed the same tasks: reading, generating keywords, judging comprehension, completing inference comprehension tests. All aspects of the experimental procedure were the same as in Experiment 1. The difference between groups was the procedural order of the tasks and presence or absence of filler texts between keyword generations.

Participants in the RK-delayed-KK-successive group completed the same procedure as the RK-delayed-KJ-lag group in Experiment 1, which also replicated the original delayed-keyword group in Thiede et al. (2003). They first read the six texts. After reading, they were presented the title of a text and were instructed to type five keywords for the text. They typed keywords for each of the six texts, and then judged their comprehension for each text. After judging their comprehension of the last text, they answered six 4-alternative multiple-choice inference questions for each text. Participants in the modified delayed group (RK-delayed-KK-spaced) followed almost this same procedure, except that just prior to generating keywords for each text they read a filler text on an unrelated topic.

Participants in the RK-immediate-KK-spaced group completed the same procedure as the RK-delayed-KJ-lag group in Experiment 1, which also replicated the original immediate-keyword group in Thiede et al. (2003). They read a text. They were then shown the title of a text and instructed to type keywords for that text. They read and immediately wrote keywords for each text. After writing keywords for the last text, participants judged their comprehension of each text. Following the last comprehension judgment, they answered six 4-alternative multiple-choice inference test questions for each text. Participants in the modified immediate (RK-immediate-KK-double-spaced) group followed almost this same procedure, except that following each reading and just prior to each keyword generation task they read a filler text on an unrelated topic.

As with Experiment 1, the length of the delays and lags were defined by the number of intervening reading and keyword generation tasks. The RK-delayed-KK-successive group had five tasks between the first reading and first keyword generation task, but zero tasks between keyword generations. The RK-delayed-KK-spaced group had six tasks between the first reading and the first keyword generation task, and one task between each keyword generation. The RK-immediate-KK-spaced group had zero tasks between each reading and keyword generation task, but one task between each keyword generation. Finally, the RK-immediate-KK-double-spaced group had one task between each reading and keyword generation task, and two tasks between each keyword generation. It is important to recognize that this modified immediate group is not truly RK-immediate, but rather has a short one-task delay between reading and keyword generation. If the RK delay is critical, this short delay may show modest improvements in metacomprehension accuracy, somewhere between the true immediate and delayed groups.

Results and Discussion

For all dependent variables, analyses consisted of a 2(RK-delayed versus RK-immediate) X 2(KK filler texts versus no-KK-filler) analysis of variance. However, since the addition of filler texts created multiple levels of RK delays and KK spacing, it was especially important to also compare individual groups more directly to evaluate the potentially conflicting influences of KK and RK lags on metacomprehension accuracy.

Test performance and metacomprehension judgments. Descriptives on test performance and comprehension judgment are presented in Table 2. Test performance and comprehension ratings did not differ across groups. Neither the main effects nor the interaction were significant for test performance, all $F(1, 96)$s < 1, MSE = .02, p > .10, or for comprehension judgments, all $F(1, 96)$s < 1, MSE = 1.40, p > .10.

Metacomprehension accuracy. Metacomprehension accuracy was operationalized as a Goodman-Kruskal gamma correlation between a participant's comprehension judgments and test performance across texts. The mean of the intra-individual correlations (between judgments and test performance) was then computed across participants for each group. No participants had indeterminate gamma correlations.

Metacomprehension accuracy was affected by the RK lag, as indicated by a main effect of delay versus immediate conditions, $F(1, 96) = 12.25$, MSE = .15, p = .001, effect size = .11. As shown in Figure 2, accuracy was substantially higher for the groups that generated keywords after a delay than for the groups that generated keywords immediately after reading. Contrary to the expected effect of the KK lag, the presence of filler texts slightly improved metacomprehension accuracy overall, but this main effect was not significant, $F(1, 96) = 1.2$, MSE = .15, p > .10. The interaction approached significance, $F(1, 96) = 3.43$, MSE = .15, p = .07, which was due to the marginally improved accuracy for adding filler texts to the RK-immediate condition, $F(1, 48) = 3.38$, p = .07. The RK-delayed-KK-successive and RK-delayed-KK-spaced groups did not differ, $F(1, 48) = 0.39$, p > .10. In addition, the two groups with identical KK lags but different RK lags (RK-delayed-KK-spaced and RK-immediate-KK-spaced) were significantly different, $F(1, 48) = 13.75$, p < .001, effect size = .52.

As in Experiment 1, these results revealed the delayed-keyword effect and implicate the longer RK interval as critical to producing the difference in metacomprehension accuracy. The strongest evidence for this conclusion is the superior metacomprehension accuracy of the two RK-delayed groups versus the RK-immediate groups. Even when comparing the two groups with identical KK lags, metacomprehension accuracy was greater for the RK-delayed group than for the RK-immediate group. In addition, the modest increase in RK-immediate metacomprehension accuracy for the KK double-spaced group over the original KK spaced group provides further evidence that the RK lag is critical and direct evidence against any negative influence due to longer a KK lag. More specifically, in the present experiment, Experiment 1, and Thiede et al. (2003), the original delayed groups with higher accuracy always had a shorter KK lag and a longer RK delay than the original immediate groups. Thus, if the KK lag was responsible for this replicated effect, then shorter KK lags must somehow improve metacomprehension accuracy. The RK-immediate group with filler texts had double the KK spacing, so the observed greater accuracy for this group was just the opposite of any expected KK lag influence. However, this KK-double spaced group also had a short,

one task RK delay. Thus, the greater metacomprehension accuracy for this group over the true RK-immediate group is consistent with theoretically predicted and previously observed effects of an RK delay.

Across Experiments 1 and 2, only the delay between reading and keyword generation showed a systematic relationship to the observed differences in metacomprehension accuracy. Differences in the KJ lag were neither necessary nor sufficient to produce differences in metacomprehension accuracy in the first experiment, and differences in the KK lag were neither necessary nor sufficient to produce differences in metacomprehension accuracy in the second experiment. By contrast, the RK delay was systematically associated with improvements in metacomprehension accuracy. Even a short RK delay appeared to produce a modest improvement, and the RK delay effect was observed regardless of whether the other two lags were shorter, longer, or the same. Experiments 3 and 4 attempted to further investigate what it is about generating keywords that leads to more accurate monitoring.

Experiment 3

Results from Experiments 1 and 2 demonstrate that *delaying* keyword generation and judgments after reading (a longer RK interval) is critical for obtaining the effect on metacomprehension accuracy. This particular conclusion gains particular importance when contrasted with results from Maki (1998b), who compared accuracy for metacomprehension judgments made immediately after study versus for those made after a delay. That is, some participants judged comprehension immediately after reading each text (i.e., *immediate* judgment), whereas others first read all the paragraphs and then were provided with each text title to make the judgments (i.e., *delayed* judgments). As delaying judgments of learning has consistently improved monitoring accuracy in associative learning tasks (e.g., Nelson & Dunlosky, 1991), the prediction was that delaying comprehension judgments would also improve metacomprehension accuracy. However, in Maki's study, mean metacomprehension accuracy was not different from zero for participants who made delayed judgments (mean gamma correlation = .10). Thus, it seems that improved metacomprehension accuracy does not result from delaying judgments alone (see also Dunlosky, Rawson, & Middleton, 2005). What may be critical then in the present experiments is that improved metacomprehension accuracy was seen as the result of delayed *generation tasks*, in particular keyword generation. The presence of such a task may enhance the likelihood that the participants access a text representation from memory prior to making a metacomprehension judgment, and delaying judgments alone may not be enough to provide the reader with diagnostic monitoring cues.

However, an alternative possibility is that the delayed keyword generation tasks simply prompt more extensive consideration of memory for texts. In the case of delayed judgments alone, individuals typically take little time (e.g., 6 to 8 seconds) to generate their metacomprehension judgments, suggesting they are driven by a shallow analysis of information about the text that can be readily accessed in the moments prior to making the judgment (Morris, 1990). By contrast, participants in Experiments 1 and 2 usually took 30 seconds or longer to generate keywords (Experiment 1 overall mean = 65.1, Experiment 2 overall mean = 74.0 seconds) prior to making the metacomprehension judgments, which presumably would afford a more thorough search of memory.

In the present experiment, we further explore the conditions that support high levels of metacomprehension accuracy by comparing two groups who made metacomprehension judgments either immediately after reading or after a delay. Participants were not asked to generate keywords. Instead, they were instructed to use the time prior to making the judgment to think about the essence of the text, which included reflecting on the main points and details of the text. The idea here is that if keywords solely encourage a more thorough consideration of memory for the text, then instructing participants to do such an analysis alone will yield greater metacomprehension accuracy for delayed than immediate judgments.

Method

Subjects. Fifty-four students enrolled in a psychology course at the University of Illinois at Chicago participated as part of a subject pool and were randomly assigned to two groups by order of appearance.

Participants were treated in accordance with the ethical standards of the American Psychological Association.

Materials and design. The texts were the same as in the previous experiments. The timing of when participants were asked to think about each text was manipulated between subjects. That is, participants were asked to think about a text either immediately after reading it or after a delay. As in Experiment 2, only inference questions were used in this experiment.

Procedure. Participants were in one of two groups (think about a text immediately after reading versus after a delay). An overview of the procedure is presented in Table 3. Participants in the delayed-think-about-the-text group, read all the texts. They were then presented the title of a text and were instructed to "think about the essence of the text, including main ideas and details." They could think about the text as long as they liked and signaled that they were done thinking about the text by typing the return key. After thinking about the last text, they judge their comprehension of each text. Following the judgment for the last text, they answered six 4-alternative multiple-choice inference test questions for each text. Participants in the immediate-think-about-the-text group, read a text and were immediately instructed to think about the text, the instructions were the same as for the delayed group. After reading and thinking about each text, they judged their comprehension of each text. They then answered inference test questions for the texts. Results and Discussion

Test performance and comprehension judgments. As before, descriptives on test performance and comprehension judgments are provided first because metacomprehension accuracy describes the relation between these variables. As reported on the bottom of Table 2, test performance and comprehension ratings did not differ across groups, both $t(52)$s < 1.0, $p > .10$.

Time to think about texts. For each participant, we computed the median time spent generating keywords across the six texts. The mean of the medians was then computed across participants for each group (Table 4). The mean time spent thinking about the texts did not differ for the delayed group and the immediate group, $t(52) < 1.0$, $p > .10$. Note, that as expected, participants instructed to reflect upon the main ideas and details of the text used quite a bit of time doing so, and much more so than participants typically use (< 10 sec) to make either immediate or delayed metacomprehension judgments (Baker & Dunlosky, in press; Morris, 1990).

Metacomprehension accuracy. As before, metacomprehension accuracy was operationalized as a Goodman-Kruskal gamma correlation between a participant's comprehension judgments and test performance across texts. No participants had indeterminate gamma correlations. The mean gamma correlation was not different for the delayed group and the immediate group, $t(52) = .20$, $p > .10$, see the leftmost panel of Figure 3. Apparently, the extra time spent thinking about the texts after a delay did not improve metacomprehension accuracy in an analogous manner to the delayed generation of keywords.

Experiment 4

Delaying metacomprehension judgments—even with specific instructions to consider the to-be-judged text in depth—is evidently not sufficient to boost metacomprehension accuracy, which implicates the keywords as an important factor for boosting the accuracy of delayed judgments. However, the instruction to think about the text is a vague direction for participants. Keywords may act as more specific cues, and may constrain the way readers reflect on their understanding of the text. To test the possibility that providing keywords guides participants' metacomprehension judgments, we employed a yoked design in which half the participants generated keywords, while the other half were asked to read the keywords that had been generated by others. If keywords act as specific cues to help readers better reflect on their understanding, then metacomprehension accuracy should improve when keywords are provided to readers prior to delayed judgments as much as when keywords are generated.

Method

Subjects. Sixty students enrolled in a psychology course at the University of Illinois at Chicago participated as part of a subject pool and were randomly assigned to two groups by order of appearance. Participants were treated in accordance with the ethical standards of the American Psychological Association.

Materials and design. The texts were the same as in the previous experiments. Participants were in one of two groups. The *generate group* generated a list of keywords as in Experiment 1. The *read group* read a list of keywords that had been generated by a yoked participant from the generate group. As in the previous experiment, only inference questions were used in this experiment.

Procedure. The *generate group* generated a list of keywords as in the delayed-keyword group in Thiede et al. (2003). That is, participants read all the texts. They then generated keywords for each text. They then judged their comprehension of texts and finally answered six 4-alternative multiple-choice inference questions for each text. The *read group* completed the same procedure as the generate group, except instead of generating keywords, they read a list of keywords that had been generated by a yoked participant from the generate group. Participants were instructed to study the keywords that were presented, as the keywords had been generated to capture the essence of the text. The participants controlled the amount of time allocated to studying the keywords.

Results and Discussion

Test performance and comprehension judgments. As before, descriptives on test performance and comprehension judgments are provided first because metacomprehension accuracy describes the relation between these variables. As reported on the bottom of Table 2, test performance did not differ across groups, $t(58) < 1.0$, $p > .10$. Comprehension judgments were significantly higher for the read group than for the generate group, $t(58) = 2.49$, $p = 02$, effect size = .10.

Time processing keywords. For each participant, we computed the median time spent generating or reading keywords across the six texts. The mean of the medians was then computed across participants for each group, as reported in the bottom of Table 4. It took longer to generate keywords than it did read the keywords, $t(58) = 7.3$, $p < .001$, effect size = .48.

Metacomprehension accuracy. As before, metacomprehension accuracy was operationalized as in the previous experiments. One participant in the read group was not included in the analysis because he had an indeterminate gamma due to invariance in comprehension ratings. As shown in the rightmost panel of Figure 3, the mean gamma correlation was significantly greater for the generate group than for the read group, $t(57) = 2.64$, $p = .01$, effect size = .11. These results suggest that the act of generating keywords is critical to improving metacomprehension accuracy.

The difference in the magnitude of comprehension judgments and the shorter keyword processing times in the read keyword condition suggest that keyword generation is altering the processes that readers use to reflect on and evaluate their comprehension of the texts. Only the generate condition may be prompting readers to move beyond a superficial reflection on the texts.

General Discussion

Until recently, the general conclusion to be drawn from metacomprehension research was that people demonstrate only meager levels of accuracy at judging their comprehension of recently read texts, which is particularly disheartening given that monitoring accuracy is positively related to learning (Thiede, 1999). Moreover, attempts to improve accuracy have typically produced less than impressive results (for a review see Maki, 1998b). However, recent research has revealed that monitoring accuracy can be improved using delayed generation tasks (Thiede & Anderson, 2003; Thiede et al., 2003). Even so, the delayed and immediate conditions in these past studies confounded the delay between reading and generation tasks with other task lags, such as the lag between multiple generation tasks and the lag between generation tasks and judgments. The first two experiments disentangle these confounded manipulations.

Metacomprehension accuracy describes a person's ability to judge his or her understanding of one text relative another. Therefore, contexts that facilitate relative comparison of texts should improve accuracy

(Nelson & Narens, 1980). According to this account, generating keywords for texts successively (rather than spaced) should have facilitated a relative comparison of texts because the cues produced by generating keywords can be easily compared from one text to the next. Yet, we demonstrated here that metacomprehension accuracy was not influenced by spacing keyword generations. Once successive keyword generation was unconfounded from RK delay, successive generations were no longer associated with improved metacomprehension accuracy. A second alternative hypothesis was that shorter lags between keyword generation and metacomprehension judgments might also be the source of improved comprehension monitoring. The rationale behind this prediction is that when a person makes a metacomprehension judgment, one may be most accurate when one has easy access to relevant cues that serve as bases of judgments. Perhaps surprisingly then, metacomprehension accuracy was not affected by delaying judgments until well after keyword generation, which presumably would have provided the cues that boosts metacomprehension accuracy. Thus, one important contribution of the present research was in establishing that these factors (KK lag and KJ lag) are not responsible for improving metacomprehension accuracy, which disconfirms two plausible theoretical accounts for the delayed-keyword effect.

The present research isolated the most critical factors for improving metacomprehension accuracy, which are the *generation* of keywords and the *delay* between reading a text and generating keywords (i.e., the RK lag). The need for this combination of factors can be explained in terms of accessing a text representation that will be critical for comprehension test performance. The generation task may provide the necessary framework to direct participants to retrieve a representation of the text from memory, while the RK lag allows for changes in the relative accessibility of text representations. Delayed generation tasks may elicit relatively diagnostic cues for predicting test performance, either simply because they require retrieval from LTM, or alternatively, because the delay allows for the surface and text-base levels of representation to become less accessible, forcing readers to access their situation model which is more central to comprehension. Thus, judgments will be more accurate because the reader has already tried to access the text representation that constrains their performance on a test. The process of delayed retrieval from LTM may provide better feedback cues for predicting future access to a LTM representation. In addition, judgments may also be more accurate because the cues generated during the keyword task are based more on the situation model of the text, which would be most diagnostic of inference test performance or other tests requiring conceptual understanding (e.g., Dunlosky & Rawson, 2005; Rawson et al., 2000; Thiede & Anderson, 2003; Wiley, Griffin & Thiede, in press). On the other hand, keyword generation immediately after reading does not require retrieval from LTM and can be performed using a highly accessible textbase representation that is not diagnostic of performance on a delayed test of comprehension. The general idea here is that the combination of *delay* and *keyword generation* increases the likelihood that a text representation that is diagnostic of future test performance will be accessed, and that cues from accessing this representation allow the reader to make more accurate metacomprehension judgments.

In Experiments 3 and 4, we found that the metacomprehension accuracy of delayed judgments was relatively low compared to keyword generation, which suggests that the improvements result from a particular interaction between generating one's own keywords and performing this task at a delay. Delaying comprehension judgments alone does not improve metacomprehension accuracy. This is supported both by our data and previous finding of Maki and Dunlosky (Maki, 1998b; Dunlosky et al., 2005). Further, extending the amount of time in which a reader reflects on the text before judgment also does not lead to improved metacomprehension accuracy. Moreover, delayed activities like vaguely thinking about the text or reading other people's keywords does not require the same process and representation access as delayed generation, thus does not provide diagnostic feedback about comprehension. Instead, what seems critical for improving metacomprehension accuracy is the presence of certain cognitive tasks, such as delayed keyword generation, that involve the kind of processing and access to representations relevant to future comprehension performance.

At present it is impossible to distinguish the relative contribution of these two possible factors. More work is needed to discriminate between them and determine whether it is LTM access or, more specifically, situation model access that produces the cues for predicting delayed comprehension performance. Such

future work will face methodological obstacles given the construct overlap between LTM text representations and situation model representations, and the possibility that features of the situation model are a major factor in determining LTM accessibility.

Our goal in future investigations is to discover other interventions that may also improve metacomprehension accuracy. In particular, we are interested in finding tasks or reading strategies that may require situation model use either without a delay or even during the act of reading. Not only would such tasks aid in testing use of the situation model independent of LTM access, but these kinds of on-line tasks may be most useful in real classroom instruction. Future research focusing on the nature of such tasks and what they require or afford is likely to either provide further support for the mechanism of situation model use or provide insights into alternative mechanisms that might account for the delayed-keyword effect.

In summary, discovering ways to improve metacomprehension accuracy is an important step to understanding the processes involved in monitoring comprehension. The present research illuminated such processes by systematically ruling out factors that could have contributed to enhancing metacomprehension accuracy, and isolating the key factors (a generation task, and a delay between reading and generation) that are responsible for one of the largest and most robust effects on metacomprehension accuracy—the delayed-keyword effect. Although our preferred account of this effect involves the role of monitoring at the level of the situation model, further progress is likely to be made when this account is competitively evaluated against other contenders. Perhaps most important, by demonstrating the key factors as well as the replicability of the delayed-keyword effect, we suspect that such contenders will be forthcoming, which in turn will promote further progress toward understanding the delayed-keyword effect in particular and metacomprehension in general.

References

Baker, J., & Dunlosky, J. (in press). Does Momentary Accessibility Influence Metacomprehension Judgments? The Influence of Study-Judgment Lags on Accessibility Effects**.** *Psychonomic Bulletin & Review*..

Dunlosky, J., Hertzog, C., Kennedy, M., & Thiede, K. (in press). The self-monitoring approach for effective learning. *Cognitive Technology*.

Dunlosky, J., & Rawson, R. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37-55.

Dunlosky, J., Rawson, K., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551-565

Fletcher, C. R., & Chrysler, S. T. (1990). Surface forms, textbases, and situation models: Recognition memory for three types of textual information. *Discourse Processes*, 13, 175-190.

Glenberg, A.M., Sanocki, T., Epstein, W., and Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*, 119-136.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163-189.

Gravetter, F.J. & Wallnau, L.B. (1999). *Essentials of statistics for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*, 133-159.

Lin, L. & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345-391.

Maki, R. H. (1998a). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds). *Metacognition in Educational Theory and Practice.* (pp. 117-145). Hillsdale, NJ: LEA.

Maki, R.H. (1998b). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26*, 959-964.

Maki, R.H. & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 116-126.

Metcalfe, J. & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530-542.

Morris, C.C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 223-232.

Nelson, T.O. and Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267-270.

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

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. In G. H. Bower (Ed.), The psychology of learning and motivation, vol. 26, (pp. 125-173). New York: Academic Press.

Rawson, K.A., Dunlosky, J., & Thiede, K.W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory and Cognition, 28*, 1004-1010.

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25*, 279-294.

Thiede, K.W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667.

Thiede, K.W. & Anderson, M.C.M. (2003). Summarizing can improve meta-comprehension accuracy. *Contemporary Educational Psychology, 28*, 129-160.

Thiede, K.W., Anderson, M.C.M. & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66-73.

Thiede, K.W. & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024-1037.

Wiley, J., Griffin, T., & Thiede, K. W. (in press). Putting the Comprehension in Metacomprehension. *Journal of General Psychology*.

Winne, P. H. & Hadwin, A. F. (1998). Studying as self-regulated learning. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds). *Metacognition in Educational Theory and Practice.* (pp. 277-304). Hillsdale, NJ: LEA.

Appendix A

Sample Text and Test Questions

The testing of intelligence has a history going back to the turn of the century when Binet in Paris attempted to select children who might profit from public education. Since that time the notion of intelligence has been the subject of considerable scrutiny, especially by Spearman in England in the 1930s, and of much and often bitter controversy. Intelligence is defined as a general reasoning ability which can be used to solve a wide variety of problems. It is called general because it has been shown empirically that such an ability enters into a variety of tasks. In job selection, for example, the average correlation with occupational success and intelligence test scores is 0.3. This is a good indication of how general intelligence is, as an ability.

This general intelligence must be distinguished from other abilities, such as verbal ability, numerical ability, and perceptual speed. These are more specific abilities which, when combined with intelligence, can produce very different results. A journalist and engineer may have similar general intelligence but would differ on verbal and spatial ability. The illiterate scientist and innumerate arts student are well-known stereotypes illustrating this point.

Most of our knowledge of intelligence has come about through the development and use of intelligence tests. In fact, intelligence is sometimes defined as that which intelligence tests measure. This is not as circular as it might appear: what intelligence tests measure is known from studies of those who score highly and those who do not, and from studies of what can be predicted from intelligence test scores. Indeed, the very notion of intelligence as a general ability comes about from investigations of intelligence tests and other scores. Well-known tests of intelligence are the Wechsler scales (for adults and children), the Stanford-Binet test and the British Intelligence Scale. These are tests to be used with individuals. Well-known group tests are Raven's Matrices and Cattell's Culture Fair test.

The IQ (intelligence quotient) is a figure which makes any two scores immediately comparable. Scores at each age group are scaled such that the mean is 100 and the standard deviation is 15 in a normal distribution. Thus a a score of 130 always means that the individual is two standard deviations above the norm, that is, in the top 2 1/2 percent of the age group.

Modern intelligence tests have been developed through the use of factor analysis, a statistical method that can separate out dimensions underlying the observed differences of scores on different tests. When this is applied to a large collection of measures, an intelligence factor emerges which can be shown to run through almost all tests. Factor loadings show to what extent a test is related to a factor. Thus a test of vocabulary loads about 0.6, that is, it is correlated 0.6 with intelligence. Such loadings, of course, give a clear indication of the nature of intelligence.

The results of one of the most technically adequate factor analyses, by Cattell, can be summarized as follows. Intelligence breaks down into two components. Fluid ability is the basic reasoning ability--which in Cattell's view is largely innate and depends upon the neurological constitution of the brain. It is largely

independent of learning and can be tested best by items which do not need knowledge for their solution. Crystallized ability is fluid ability as it is used in a culture. In Cattell's view, crystallized ability results from the investment of fluid ability in the skills valued by a culture. This involves the traditional academic disciplines, for example, physics, mathematics, classics, or languages.

Many social class differences in intelligence test scores and educational attainment are easily explicable in terms of these factors especially if we remember that many old-fashioned intelligence tests measure a mixture of these two factors. Thus in middle-class homes, where family values and cultural values are consonant, a child's fluid intelligence becomes invested in activities which the culture as a whole values (verbal ability, for example). Performance in education is thus close to the full ability, as measured by the fluid ability of the child. In children from homes where educational skills are not similarly encouraged there may be a considerable disparity between ability and achievement. On intelligence tests where crystallized ability is measured, social class differences are greater than on tests where fluid ability is assessed.

Thus a summary view of intelligence based on the factor analysis of abilities is that it is made up of two components: one a general reasoning ability, largely innate, and the other a set of skills resulting from investing this ability in a particular way. These are the two most important abilities. Others are perceptual speed, visualization ability and speed of retrieval from memory--a factor which affects how fluent we are in our ideas and words.

We are now in a position to examine some crucial issues in the are of intelligence and intelligence testing, issues which have often aroused considerable emotion but have been dealt with from bases of ignorance and prejudice rather than knowledge. Positions on the controversial question of the heritability of intelligence polarize unfortunately around political positions. Opponents of the hereditary hypothesis were heartened by the evidence (now generally accepted) that Sir Cyril Burt had manufactured his twin data which supported this hypothesis. However, there are other more persuasive data confirming this position—data coming from biometric analyses.

First, what is the hereditary hypothesis? It claims that the variance in measured intelligence in Britain and the USA is attributable about 70 percent to genetic factors and 30 percent to environmental factors. It is very important to note that this work refers to variance within a particular population. If the environment were identical for individuals, variation due to the environment would be zero. This means that figures cannot be transported from culture to culture or even from historical period to period. This variance refers to population variance; it does not state that 70 percent of the intelligence in any particular individual is attributable to genetic factors. Finally, a crucial point is that interaction takes place with the environment; there is no claim that all variation is genetically determined.

These figures have been obtained from biometric analysis, which involve examining the relationship of intelligence test scores of individuals of differing degrees of relatedness. This allows variance to be attributed to within-family and between-family effects, as well as enabling the investigator to decide whether, given the data, assortative mating, or other genetic mechanisms can be implicated. Work deriving from this approach is difficult to discount.

The issue of racial differences in intelligence is even more controversial, with potentially devastating political implications. Some social scientists feel that this is a case where research should be stopped, as for example with certain branches of nuclear physics and genetic engineering. Whether suppression of the truth or the search for it is ever justifiable is, of course, itself a moral dilemma.

The root of the problem lies in the inescapable fact that, in the USA, Blacks score lower on intelligence tests than do any other group. Fascists and members of ultra right-wing movements have immediately interpreted this result as evidence of Black inferiority. Opponents of this view have sought the cause in a variety of factors: that the tests are biased against Blacks, because of the nature of their items. Other arguments are that Blacks are not motivated to do tests set by Whites; that the whole notion of testing is foreign to Black American culture; that the depressed conditions and poverty of Black families contribute to their low scores; that the prejudice against Blacks creates a low level of self-esteem so that they do not perform as well as they might; and that verbal stimulation in the Black home is less than in that of Whites. Jensen investigated the whole issue in great detail. Many of these arguments were refuted by experimental evidence, especially the final point--for Blacks do comparatively better on verbal than non-verbal tests. But to argue that this is innate or biologically determined goes far beyond the evidence. Motivational

factors and attitudes are difficult to measure and may well play a part in depressing Black scores. What is clear, however, is that on intelligence tests, as a group, American Blacks perform less well than other racial or cultural groups, while these tests still predict individual success in professional, high-status occupations.

Intelligence as measured by tests is important because in complex technologically advanced societies it is a good predictor of academic and occupational success. That is why people attach great value to being intelligent. For example, cross-cultural studies of abilities in Africa have shown that the notion of intelligence is different from that in the west and is not there so highly regarded. Many skills in African societies may require quite different abilities. Thus as long as, in a society, it is evident that a variable contributes to success, that variable will be valued; and even though intelligence is but one of a plethora of personal attributes, there is, in the west, little hope that more reasoned attitudes to intelligence will prevail.

Two further points remain to be made. First, the fact that there is a considerable genetic component does not mean that the environment (family and education) do not affect intelligence test scores. It has clearly been shown that even with 80 percent genetic determination, environmental causes can produce variations of up to 30 points. Finally, the rather abstract statistically defined concept of intelligence is now being intensively studied in cognitive experimental psychology in an attempt to describe precisely the nature of this ability.

Detail question

Which of the following tests is suitable for group administration?

A. British Intelligence Scale
B. Raven's Matrices*
C. Stanford-Binet test
D. Wechsler scales

Inference question

What would be more helpful in scoring well on an achievement test?

A. Crystalized ability*
B. Fluid ability
C. neither affects performance on achievement tests
D. they would affect performance on an achievement test equally

Author Note

The current experiments were supported by a grant from the U. S. Department of Education's Institute of Education Sciences (Grant # R305H030170) to Keith Thiede and Jennifer Wiley. All opinions expressed herein are those of the authors and do not reflect those of the funding organization. We thank Qiong Fu for her research assistance on this project.

Footnotes

1. These data are not needed to evaluate hypotheses related to the delayed-keyword effect. Nonetheless, we analyzed EOL data. Across all four experiments, we showed that the mean magnitude of EOLs and the correlation between EOL and test performance did not differ across groups, all $Fs < 1.7$, $p > .10$.

2. These data are not needed to evaluate hypotheses related to the delayed-keyword effect. However, we analyzed reading time data to establish equivalence for time on task or engagement with the texts. Across all four experiments, we showed that the mean reading time did not differ across groups, all $Fs < 1.2$, $p > .10$.

Figure Captions

Figure 1. Metacomprehension Accuracy for Experiment 1. For the data presented comprehension was assessed with inference test questions. On the horizontal axis, Immediate Keyword Generation is RK-immediate and Delayed Keyword Generation is RK-delayed. In the legend, No Keyword-Judgment Lag is KJ-no lag and Keyword-Judgment Lag is KJ-lag. The error bars are the standard errors of the means.

Figure 2. Metacomprehension Accuracy for Experiment 2. Comprehension was assessed with inference test questions. On the horizontal axis, the KK-spaced groups had generation of keywords for one text and the next filled with reading of one text. The KK-double spaced group had generation of keywords for one text and the next filled with reading of two texts. The KK-successive group had no delay between generation of keywords for one text and the next. In the legend, the RK-immediate group generated keywords immediately after reading a text. The RK-short delay group had a filler text between generation of keywords for one text and the next. The RK-delayed groups had the delay between reading and generating keywords filled with the reading of the remaining texts. The KK-spaced group, the rightmost bar, had a filler text between generation of keywords for one text and the next. The error bars are the standard errors of the means.
Note: The 2nd and 4th bars represent groups with filler texts.

Figure 3. Metacomprehension Accuracy for Experiments 3 and 4. For both experiments, comprehension was assessed with inference test questions. The error bars are the standard errors of the means.

Table 1. *Overview of Procedures for Thiede et al. 2003, Experiments 1 and 2.*

Thiede, et al. (2003): Immediate and Delayed Keyword Conditions				
Delayed-Keyword	R1, R2, R3…R6	K1, K2, K3…K6	J1, J2, J3… J6	T1, T2, T3…T6
Immediate-Keyword	R1, K1; R2, K2; R3, K3…R6, K6		J1, J2, J3… J6	T1, T2, T3…T6
Experiment 1: Lag versus No-lag between Keywords and Judgments				
*RK-delayed-KJ-lag	R1, R2, R3…R6	K1, K2, K3…K6	J1, J2, J3… J6	T1, T2, T3…T6
RK-delayed-KJ-no lag	R1, R2, R3…R6	K1, J1; K2, J2; K3, J3… K6, J6		T1, T2, T3…T6
**RK-immediate-KJ-lag	R1, K1; R2, K2; R3, K3…R6, K6		J1, J2, J3… J6	T1, T2, T3…T6
RK-immediate-KJ-no lag	R1, K1, J1; R2, K2, J2; R3, K3, J3… R6, K6, J6			T1, T2, T3…T6
Experiment 2: Successive versus Spaced and Double-spaced Keyword Generations				
*RK-delayed-KK-successive	R1, R2, R3…R6	K1, K2, K3…K6	J1, J2, J3… J6	T1, T2, T3…T6
RK-delayed-KK-spaced	R1, R2, R3…R6	F1, K1, F2, K2, F3…K6	J1, J2, J3… J6	T1, T2, T3…T6
**RK-immediate-KK-spaced	R1, K1; R2, K2; R3, K3…R6, K6		J1, J2, J3… J6	T1, T2, T3…T6
RK-immediate-KK-dbl spaced	R1, F1, K1; R2, F2, K2…R6, F6, K6		J1, J2, J3… J6	T1, T2, T3…T6

Note. R1 = Participants read text 1… ; K1 = Participants generated a list of keywords for text 1…; J1 = Participants judged their comprehension of text 1…; T1 = Participants took test on text 1…; F1 = Participants read filler text 1…
* Indicates the condition replicates the original Delayed-Keyword condition in Thiede et al. (2003).
** Indicates the condition replicates the original Immediate-Keyword condition in Thiede et al. (2003).

Table 2. *Mean Test Performance and Comprehension Ratings*

Group	Test Performance		Comprehension Rating
		Experiment 1	
	Inference	Detail	
RK-Delayed-KJ-lag	.50 (.03)	.52 (.02)	4.30 (.21)
RK-Delayed-KJ-no lag	.46 (.03)	.49 (.02)	4.21 (.16)
RK-Immediate-KJ-lag	.46 (.02)	.48 (.02)	4.30 (.19)
RK-Immediate-KJ-no lag	.47 (.03)	.52 (.03)	4.52 (.17)
		Experiment 2	
RK-Delayed-KK successive	.46 (.03)		3.74 (.26)
RK-Delayed-KK spaced*	.43 (.03)		4.00 (.24)
RK-Immediate-KK spaced	.42 (.03)		4.02 (.24)
RK-Immediate-KK double spaced*	.42 (.03)		4.24 (.21)
*These conditions contained filler texts			
		Experiment 3	
Immediate-think about text	.51 (.05)		3.74 (.31)
Delayed-think about text	.49 (.05)		4.04 (.31)
		Experiment 4	
RK-Delayed-generate keywords	.40 (.03)		3.58 (.19)
RK-Delayed-read keywords	.42 (.02)		4.25 (.19)

Note. Entries are mean across individual's median test performance (proportion correct) and median judgments. Values in parentheses are standard errors of the means.

Table 3. *Overview of Procedures for Experiments 3 and 4.*

Experiment 3: Delayed versus Immediate Thinking-about-text Conditions				
Delayed-Thinking	R1, R2, R3…R6	Th1, Th2, Th3…Th6	J1, J2, J3… J6	T1, T2, T3…T6
Immediate-Thinking	R1, Th1; R2, Th2; R3, Th3…R6, Th6		J1, J2, J3… J6	T1, T2, T3…T6
Experiment 4: Delayed Read versus Delayed Generate Keyword Conditions				
Delayed-Reading	R1, R2, R3…R6	RK1, RK2, RK3…RK6	J1, J2, J3… J6	T1, T2, T3…T6
*Delayed-Generate	R1, R2, R3…R6	GK1, GK2, GK3…GK6	J1, J2, J3… J6	T1, T2, T3…T6

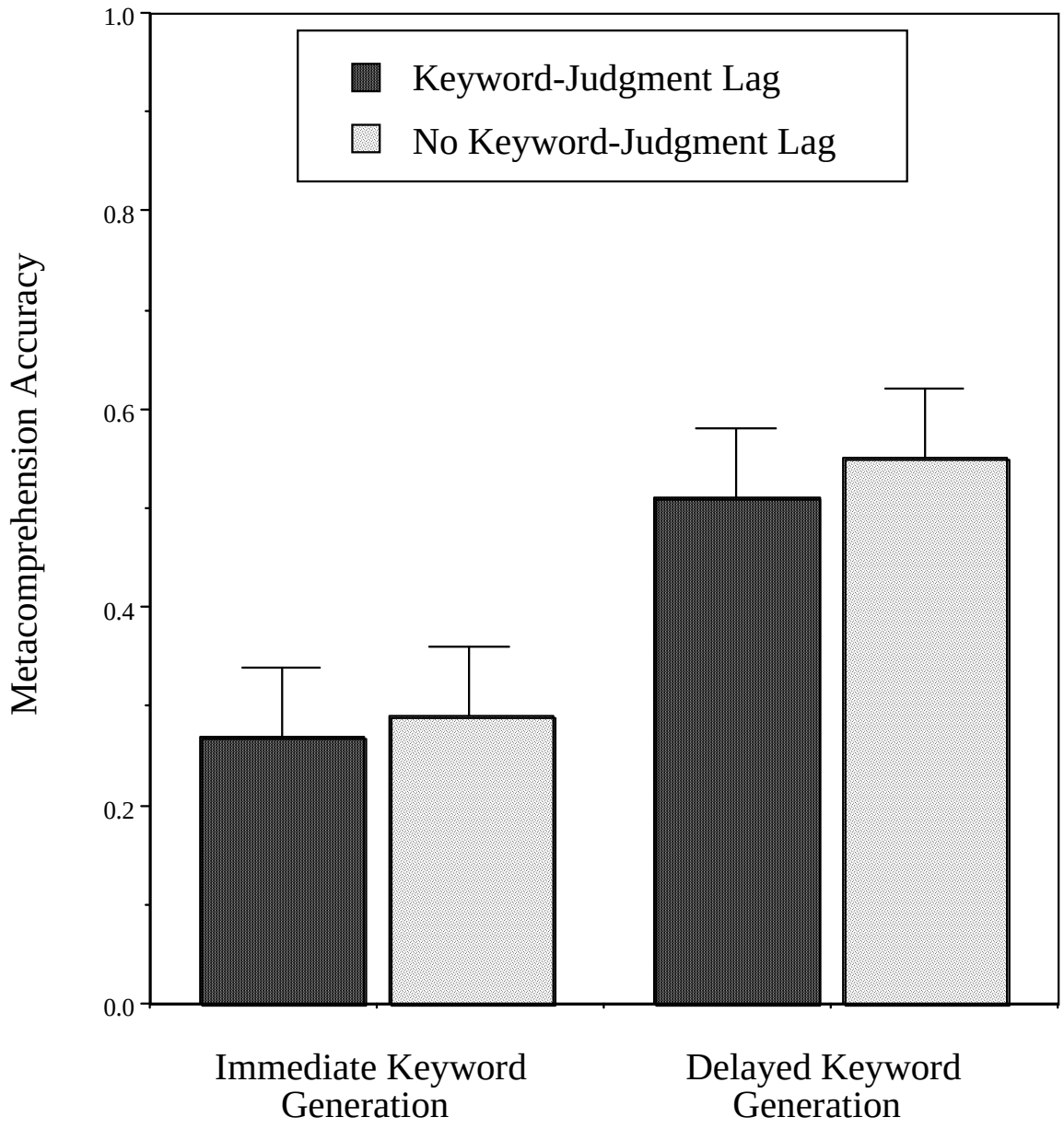
Note. R1 = Participants read text 1… ; Th1= Participants was instructed to think about text 1… ; J1 = Participants judged their comprehension of text 1…; T1 = Participants took test on text 1…; RK1 = Participants read keywords for text 1... ; GK1 = Participants generated keywords for text 1…

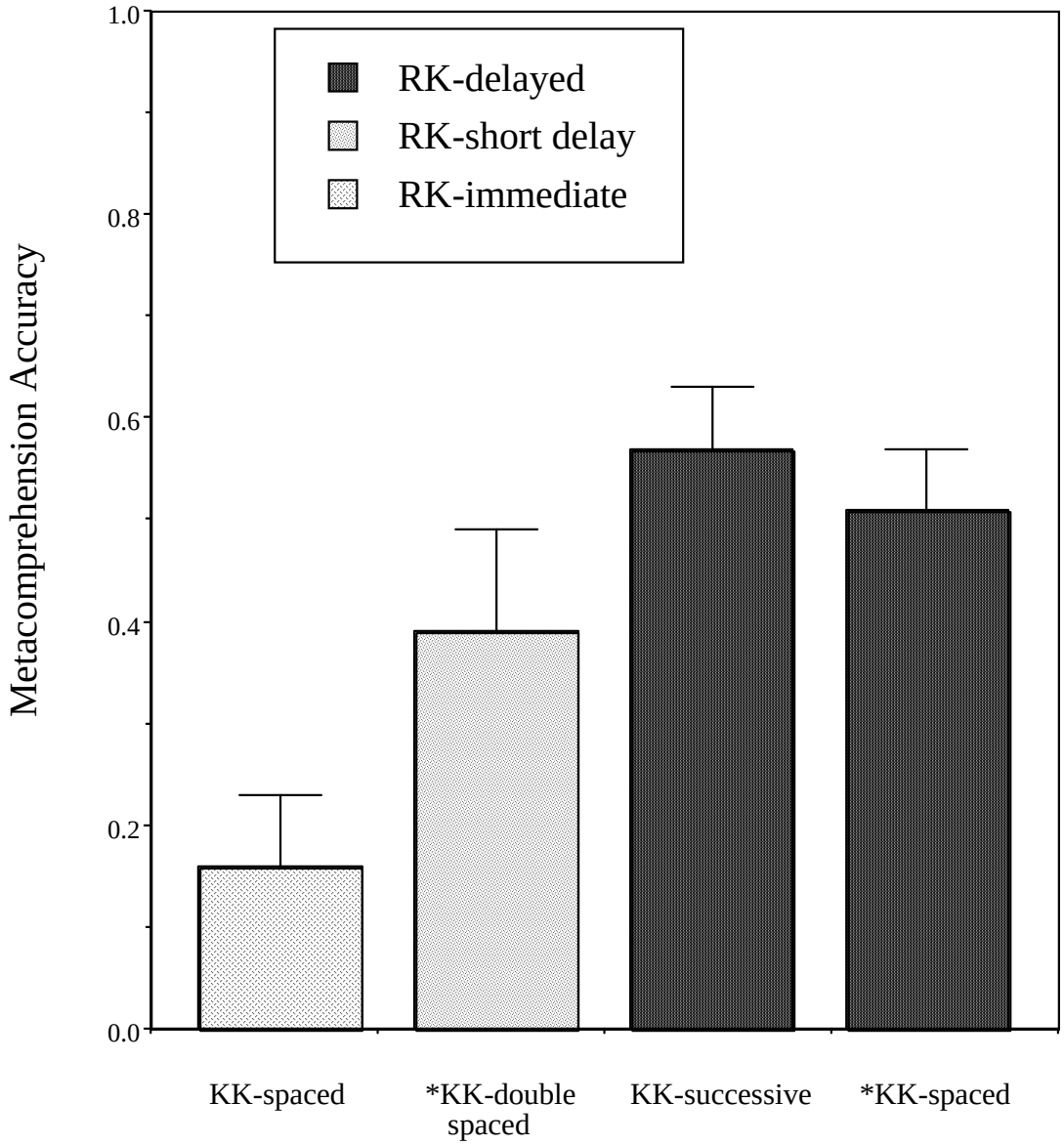
* Indicates the condition replicates the original Delayed-Keyword condition in Thiede et al. (2003).

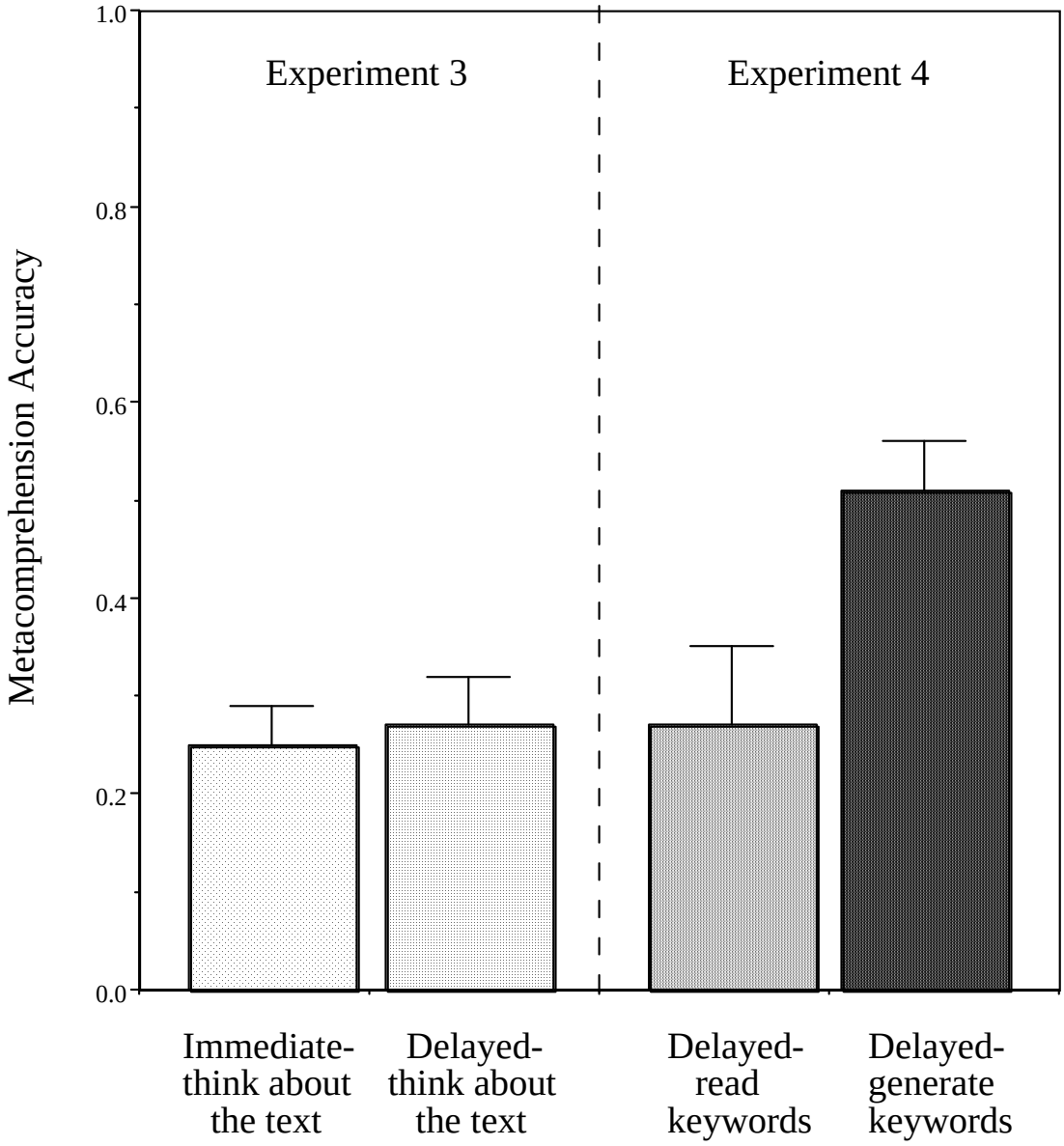
Table 4. *Mean Latency to Think about Texts, Generate and Examine Keywords*

Group	Latencies
Experiment 3	
Immediate-think about text	42.4 (5.2)
Delayed-think about text	50.2 (8.9)
Experiment 4	
RK-Delayed-generate keywords	45.8 (3.7)
RK-Delayed-read keywords	14.6 (2.3)

Note. Entries are mean across individual's median latencies to generate five keywords. Values in parentheses are standard errors of the means.


1.0
0.8
0.6
0.4
0.2
0.0
Metacomprehension Accuracy
Keyword-Judgment Lag
No Keyword-Judgment Lag
Immediate Keyword Generation
Delayed Keyword Generation


RK-delayed
RK-short delay
RK-immediate
Metacomprehension Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
KK-spaced
*KK-double spaced
KK-successive
*KK-spaced


Experiment 3
Experiment 4
Metacomprehension Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Immediate-think about the text
Delayed-think about the text
Delayed-read keywords
Delayed-generate keywords