Mem Cogn (2011) 39:204–216
DOI 10.3758/s13421-010-0019-2

Judgments of learning and improvement

Corinne L. Townsend · Evan Heit

Published online: 18 November 2010

Abstract Can learners accurately judge the rate of their learning? Rates of learning may be informative when study time is allocated across materials, and students' judgments of their learning rate have been proposed as a possible metacognitive tool. Participants estimated how much they improved between presentations in multitrial learning situations in which *n*-gram paragraphs (in Experiments 1 and 2) or word pairs (Experiments 3 and 4) were learned . In the first experiment, participants rated improvement on a percentage scale, whereas on the second and third, judgments were given on a 0–6 scale. Experiment 4 used both a percentage scale and an absolute number scale. The main result was that judgments of improvement were poorly correlated with actual improvement and, in one case, were negatively correlated. Although judgments of improvement were correlated with changes in judgments of learning, they were not reliable indicators of actual improvement. Implications are discussed for theoretical work on metacognition.

Keywords Metacognition · Metamemory

Successful learning depends not only on the ability to learn, but also on the ability to monitor learning. Such metacognitive judgments are important for the allocation of study time, which, in turn, impacts performance (Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2005; Thiede, Anderson, & Therriault, 2003; Thiede & Dunlosky, 1999). If learners cannot accurately monitor what they have learned, decisions about what material needs to be studied, and for how long, will not be optimal, resulting in lower test performance. Improving metacognitive accuracy is also associated with better test performance. Thiede et al. (2003) showed that improved metacomprehension (comprehension judgments for text material) enhanced students' abilities to effectively focus their efforts when given time to restudy items. They more often chose to restudy texts that were less well learned, leading to superior test performance, as compared with students who did not generate key words. These results suggest that metacognitive accuracy guides learning and affects student performance.

Research on the relation between metacognitive accuracy and performance has focused on two related kinds of judgments. One is a judgment of learning (JOL; Kornell & Metcalfe, 2006; Metcalfe & Finn, 2008), which assesses how much information a person feels is known, usually solicited on a percentage scale (such as the judgment that 80% of the items have been learned). The second is a metacomprehension judgment (Dunlosky & Lipko, 2007), which is a judgment of how well a piece of text is understood, also generally scaled in terms of performance on a test concerning the material. Still, there may be other metacognitive judgments besides JOLs and metacomprehension judgments that are important for study time allocation and subsequent performance—specifically, judgments concerning learning rate (Metcalfe & Kornell, 2003, 2005), or in other words, a judgment of improvement (JOI). JOIs would benefit study because learners could choose to quit studying an item when the rate of learning drops, with little to no progress. JOIs would be especially useful when there is limited time, because no time would be wasted on unlearnable items. Efforts could be focused on more quickly learned material in order to maximize overall performance.

Rather than directly evaluating JOIs, researchers have used repeated JOLs to represent the subjective learning curve (Koriat, Sheffer, & Ma'ayan, 2002; Metcalfe & Kornell, 2005). Metcalfe and Kornell (2005) inferred participants' judged rates of improvement by subtracting the stopping JOLs from the starting JOLs; yet this JOL difference score,

C. L. Townsend (✉) · E. Heit
School of Social Sciences, Humanities, and Arts,
University of California, Merced,
5200 North Lake Rd,
Merced, CA 95343, USA
e-mail: ctownsend@ucmerced.edu

although likely somewhat related to actual rate of learning, may not reflect the subjective sense of improvement. To infer that these JOL difference scores represent the subjective sense of improvement may be inaccurate, because it presumes that people remember their previous JOL states. It is not certain that all previous JOL states (or even just the one preceding the JOL) would be remembered—especially in a complex learning scenario—so it is important to measure JOIs directly. If the previous assumptions are correct (Koriat et al., 2002; Metcalfe & Kornell, 2005), JOIs should simply reflect JOL difference scores. JOIs may indeed be based, in part, on JOLs, but they may also be influenced by subjective cues, such as a sense of fluency, interest, frustration, and so on (Metcalfe & Kornell, 2005).

Metcalfe and Kornell (2005) explained how a JOL rate would be a useful judgment during study time. In the region-of-proximal-learning model, learners use a judgment of rate of learning as a stopping rule while studying. According to this model, learners first choose items in the order of how easy they are to learn. They should stop studying a particular item when the judged learning rate drops to an unacceptably low value and should move on to the next item, to maximize the rate of return per unit time spent studying. Similarly, Son and Sethi's (2006) model of optimal learning describes how, in most learning scenarios, focusing on the item with the highest current rate of improvement will result in the maximum amount of learning per unit time spent studying.

These analyses contrast with an earlier model of study time allocation, discrepancy reduction (Carver & Scheier, 1990; Koriat & Goldsmith, 1996; Thiede & Dunlosky, 1999), in which learners choose the most difficult (low-JOL) items for restudy. Some evidence had suggested that people choose to study the most difficult items, which have the largest discrepancies between the current state of learning and the goal, and spend more time on them. This model is referred to as *discrepancy reduction* because it is assumed that people work to reduce discrepancies between the current state and the goal state (Carver & Scheier, 1990). Recently, however, conditions have been found in which people did not choose to study the most difficult items and, instead, started with easier (yet unlearned) items (Dunlosky & Thiede, 2004; Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2003). These findings lend some support to the proximal learning model, but it still remains to be seen whether or not JOIs are important for stopping decisions.

JOLs and study behavior

Students can explicitly produce and use JOLs when studying, in order to appropriately decide what to work on and for how long (Kornell & Bjork, 2007; Kornell & Metcalfe, 2006; Nelson, Dunlosky, Graf, & Narens, 1994). In a situation in which learners select items for restudy, better performance results when learners' choices are honored, as opposed to dishonored (Kornell & Metcalfe, 2006), showing that people generally choose items that could benefit from restudy. Metcalfe and Finn (2008) also demonstrated that JOLs have a direct relationship with which material is selected for restudy; manipulations that influence JOLs also influence study choices.

Much research has been conducted concerning the accuracy of these judgments, with the general finding that they are somewhat accurate. Nelson and Dunlosky (1991) found a gamma correlation of .93 between JOLs and actual recall when JOLs were delayed for a short period of time after study. Immediate JOLs also show above-chance accuracy, although the correlations are quite significantly lower than those for delayed JOLs. People generally show biases in such judgments (Finn & Metcalfe, 2008; Koriat, Sheffer, & Ma'ayan, 2002; Meeter & Nelson, 2003). These biases are best illustrated in multitrial experiments, which exhibit the underconfidence-with-practice effect (Koriat, 1997; Koriat et al., 2002; Meeter & Nelson, 2003). On the first trial, people generally show a bias toward overconfidence, giving JOLs that are higher than performance, but on succeeding trials, they give JOLs that are lower than actual performance (Koriat et al., 2002). This effect may translate into inaccurate JOIs; if JOIs are related to JOLs, these biases will impair the ability to accurately judge improvement, and improvement could be increasingly underestimated as well.

Overview of experiments

The present experiments were designed to assess the accuracy of JOIs. The most optimistic prediction would be, consistent with the region-of-proximal-learning model, that learners are sensitive to rate of improvement and can make accurate statements about rate of improvement. On the other hand, if improvement is inferred indirectly, from changes in JOLs, JOIs should be affected by the trend of increasing underconfidence, so that improvement is increasingly underestimated over time. Finally, if other variables, such as fluency, interest, or other subjective cues, are used to inform JOIs, a very different pattern may arise, including an increasing sense of JOI confidence with practice. In the extreme, if JOIs are responses to other cues, rather than true improvement, there may be a zero or even negative correlation between JOIs and improvement.

In Experiments 1 and 2, we examined learning of paragraphs of words in the form of *n*-grams (Shannon, 1948). In Experiments 3 and 4, we used the somewhat more conventional paired-associate learning paradigm. In the first three experiments, judgments made immediately after each study trial were used, whereas in Experiment 4, judgments preceding study were also examined. We assumed that for JOIs to be useful, they would be needed during study, so we solicited JOIs either just before or just after study and test, rather than sandwiched within periods of delay. The JOIs in these experiments were, as such, judgments of current memory improvements (which may not persist after a delay). A second assumption was that a JOI would typically be an aggregate, general judgment of how much material one is learning in a given amount of time. In each experiment, participants were asked to make explicit judgments. In Experiment 1, they were asked to make JOIs on the same scale as JOLs. In Experiments 2 and 3, to minimize the possibility that the participants were not judging improvement from a subjective feeling of improvement but were simply subtracting the second most recent JOL from the most recent JOL, we solicited JOIs on a different rating scale. Experiment 4 had JOLs and JOIs made between participants (to entirely avoid participants' being compelled to base JOIs on JOLs) and investigated the use of percentage and absolute number scales.

Experiment 1

This experiment was designed to investigate the accuracy of JOIs for simple text materials, using multiple study–test trials. The experiment consisted of six repeated trials, in order to approximate a natural study situation. Difficulty was manipulated by presenting 50-word paragraphs of random words that differed in *n*-gram size. An *n*-gram is defined as a sequence of *n* words from a body of text (Shannon, 1948); for example, the words in italici would be a 4-gram: The dog *barked at the cat* (see Appendix A for materials). We used these materials with the intention of approximating a common learning scenario in which students attempt to learn material contained in a text. We chose *n*-grams, rather than actual prose passages, however, because they are simple text, yet do not require interpretation or overall comprehension to learn, and it is reasonable to score them in terms of number of individual words remembered.

Method

Participants Forty-six students from the participant pool at the University of California, Merced, volunteered to participate for class credit.

Materials Three passages were constructed using *n*-grams from the "Phrases in English" database from the British national corpus (Fletcher, http://pie.usna.edu/index.html). Passages were constructed of randomly chosen 4-grams, 6-grams, or 8-grams, and all the passages were 50 words long. Any British spellings were changed to American spellings. Differences in *n*-gram size were expected to result in differing levels of learning: 4-gram paragraphs were expected to be the most difficult to learn because they were the most random, and 8 grams were expected to be the easiest because they were the most coherent. However, this was a secondary prediction, since the intent was mainly to have a variety of texts.

Procedure The experiment was conducted using a computer program, which randomly selected the order in which the paragraph types were seen. Each paragraph was presented for six trials, and each participant saw all three paragraphs. For each trial, participants viewed the paragraph for 60 s, made JOLs and JOIs for the paragraph, and were asked to recall the paragraph. JOLs were prompted with the question, "What percent of the paragraph (0–100) do you think you will be able to recall?" JOIs (on trials 2–6 only) were prompted with the question, "Compared to the previous trial, what percent more of the paragraph will you be able to recall?" Actual knowledge of the paragraph was prompted with asking, "Please type in as much as you can recall." Participants then typed in as much of the material as they could.

Scoring Responses were scored according to how many words recalled matched words from the paragraph. No points were taken off for misspellings or being out of order, since we were concerned strictly with recall of the words. (The instructions did not inform participants about the exact details of scoring.) For analysis, judgments were converted from percentages to number of words; for example, 10% was converted to 10% of 50, or 5 words. Improvement values were defined as the increase in recall from one trial to the next.

Results

Most important, JOIs were compared with actual improvement in recall, to investigate the relative accuracy of the judgments. JOIs were also compared with changes in JOLs across trials, to investigate whether JOIs could be based on differences in JOLs. The correlations between JOLs and recall were calculated, as well as biases in JOLs across trials, to check for the underconfidence-with-practice effect. All the correlations were calculated within participants. Some analyses excluded a few participants because correlations could not be calculated, due to the same response having been given for each judgment or to missing data.

Judgments of improvement The most important analysis was accuracy of JOIs. There were 15 observations per participant (improvement scores on trials 2–6 for each paragraph), and we assessed the correlation between judged improvement and actual improvement in recall over these 15 points (see Table 1 for JOI correlations for all three experiments). This correlation was significant but small; mean ρ was .19 (min = −.66, max = .67, *SD* = .27), *t*(45) = 4.64, *p* < .01. JOIs were also compared with JOL difference scores (difference scores are the difference in JOL values from one trial to the next, such as trial 2 JOL minus trial 1 JOL) in order to assess the possibility that participants used differences in JOLs across trials to make JOIs. The mean ρ for JOI and change in JOL was .31, significantly different from zero, *t*(45) = 6.05, *p* < .01. This was also significantly greater than the JOI–improvement correlation, *t*(45) = 2.34, *p* < .05. Hence, participants may have been relying on changes in reported JOLs, rather than on other cues more related to actual improvement. However, changes in JOLs are not indicative of actual improvement; the correlation between JOL difference scores and true improvement was only .11 (min = −.58, max = .62, *SD* = .25), *t*(45) = 3.00, *p* < .01.

It is possible that using the same percentage rating scale for JOLs and JOIs encouraged participants to answer consistently (e.g., always giving a JOI of 10%, and raising the JOLs 10%). Through this mechanism, the correlation between JOIs and JOL difference scores could arise. This is illustrated in Fig. 1; improvement ratings and changes in JOLs are flat across trials, especially for trials 3–6. This suggests that the low correlation between JOIs and actual improvement could be due to participants' answering in a consistent manner between JOIs and JOLs.

Absolute accuracy Bias (average difference score) and absolute accuracy (mean of squared deviations) (Schraw, 2009) were calculated for JOIs. JOI bias showed overconfidence, relative to actual improvements in recall, with mean bias = 3.44 (min = −4.56, max = 23.57*, SD* = 6.70). Mean absolute accuracy was 109.99 (min = 4.37, max = 804.8*, SD* = 194.38). A comparison with JOL change values showed very similar values, on average: a mean bias of 4.63 (min = −1.93, max = 25.03, *SD* = 6.47), and a mean absolute accuracy value of 128.01 (min = 1.35, max = 895.12, *SD* = 201.07). However, paired samples *t* tests showed that the biases were significantly different (*t*(45) = 5.16, *p* < .01) but that the absolute accuracy values were not significantly different, *t*(45) = −1.19, *p* = .242. A comparison with JOLs showed a mean bias of −12.86 (min = −33, max = 0.37, *SD* = 8.78), and an absolute accuracy value of 333.45 (min = 0, max = 1,267.87, *SD* = 341.30). These numbers were significantly different from the bias, *t*(45) = −12.96, *p* < .01, and accuracy, *t*(45) = 3.61, *p* < .01, statistics between JOIs and improvement.

Table 1 Correlations between judgments of improvement (JOIs), improvement, and judgment of learning (JOL) increase

Experiment		Improvement	JOL increase
1	JOI	.19*	.31*
	Improvement		.11*
2	JOI	.04	.34*
	Improvement		.18*
3	JOI	−.37*	.36*
	Improvement		.07

Note. Spearman correlations between JOIs and actual improvement, JOIs and increases in JOLs, and actual improvement and increases in JOLs are listed for Experiments 1–3. Values are not listed for Experiment 4, because JOIs and JOLs were made between subjects

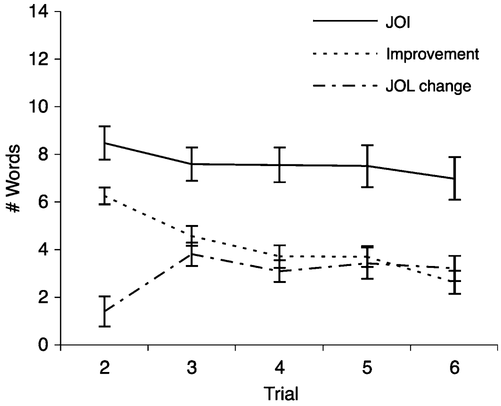

Fig. 1 Average judgments of improvement (JOIs), improvement values, and changes in judgment of learning (JOL) per trial (Experiment 1)

Judgments of learning We compared JOLs with actual recall for each of the paragraph types for each participant to see whether judgment accuracy depended on the difficulty of the material. Mean JOL correlations were all significantly different from zero (see Table 2). No significant differences were found among the JOL correlations for the different paragraphs, so further analyses were collapsed across paragraph type (mean correlations for individual paragraphs are listed in Table 2). Correlations were calculated using data from all six trials for each of three paragraphs, leading to 18 data points in total for each participant. The mean JOL Spearman correlation was ρ = .65, (min = −.29, max = .98, *SD* = .30), significantly different from zero, *t*(45) = 14.78, *p* < .01. Correlations were also calculated using Goodman–Kruskal gamma and Pearson's *r* for all four experiments; since they reached the same conclusions, only Spearman values are listed in the text (refer to Table 2 for gamma values).

Table 2 Goodman–Kruskal gamma correlations

Experiment					
1		Overall	4 gram	6 gram	8 gram
	JOL	.59*	.70*	.69*	.75*
	JOI	.19*	.20*	.21*	.11
2		Overall	4 gram	6 gram	8 gram
	JOL	.57*	.68*	.71*	.61*
	JOI	.04	−.09	−.05	.12
3		Overall	Related	Unrelated	
	JOL	.57*	.80*	.75*	
	JOI	−.40*	−.17	−.64*	
4		Overall	Percent	Number	
	JOL	.52*	.64*	.43*	
	JOIpre	.078	.13	.024	
	JOIpost	.056	.10	.006	

Note. Values listed in this table are the gamma correlations for judgments of learning (JOLs) and judgments of improvement (JOIs), both overall and for individual conditions in Experiments 1–4

Confidence bias Confidence bias is a measure of the absolute accuracy of judgments, in terms of whether learning is over- or underestimated. Confidence bias was calculated by subtracting recall from JOL (see Fig. 2 for recall vs. JOL values). There was evidence for the underconfidence-with-practice effect, with bias scores becoming increasingly negative over time, $F(5, 685) = 17.42$, $MSE = 48.04$, $p < .001$, $\eta^2 = .113$.

Discussion

The key finding was a low correlation between JOIs and actual improvement, suggesting a relatively poor ability to make JOIs that would be useful for guiding learning. Although the results do not rule out that possibility that people are (weakly) sensitive to a subjective sense of improvement and make judgments on this basis, a possible alternate explanation is that people simply remember their previous JOL and report the difference between the two most recent judgments. Indeed, JOIs were more strongly correlated with the difference in JOLs than with true improvement; however, absolute accuracy measures showed that there were substantial differences between JOI values and JOL difference scores.

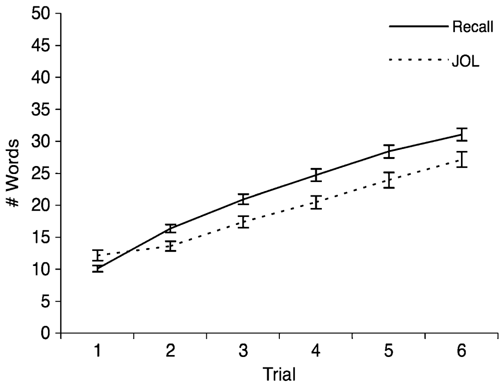

Fig. 2 Mean judgment of learning (JOL) values and recall, per trial (Experiment 1)

There are other possible explanations for the near-zero correlation between JOIs and improvement. For example, it is also possible that participants misunderstood the question, or that it was difficult for them to give percentage ratings for improvement. The finding that JOIs are somewhat related to JOL difference scores suggests three possible scenarios: People may use their JOLs to infer improvement, the percentage scale may encourage a consistent relationship between JOLs and JOIs (e.g., answering 60, 70, 80, or 90 for JOLs and always 10 for JOI), or subjective cues, such as fluency, may underlie both JOIs and JOLs, leading to a correlation without an actual relationship. This may be more likely, since absolute accuracy measures show the relation between JOIs and JOL differences to be no better than the relation between JOIs and actual improvement values.

Experiment 2

This experiment was conducted to examine whether the same findings would appear when the percentage rating scale for improvement was not used. In Experiment 2, a 0 to 6 rating scale was used for JOIs. The value of 0 meant *no improvement occurred*, whereas a value of 6 meant *a lot of improvement*. The 0 to 6 rating scale should discourage participants from attempting to calculate JOIs from differences in JOLs, as they may have done in Experiment 1. Whereas, in Experiment 1, it was possible for participants to always give a JOI of, say, 10, and always give JOLs an increment of10% higher, that would be more difficult when the two scales are different. If people are truly sensitive to improvement, without mediating JOIs by remembering previous JOLs, JOIs and actual improvement should still be correlated. Other than the changed rating scale, the method for Experiment 2 was similar to that in Experiment 1.

Method

Fifty-six students participated. Three new *n*-gram passages were constructed, again as 4-, 6-, and 8-grams. New passages were created in case there had been idiosyncratic effects of the specific passages used in Experiment 1.

The experiment was conducted as in Experiment 1, except that judgments of improvement were prompted with the question, “On a scale of 0 to 6, with 0 being not at all, and 6 being a lot, how much do you think you improved relative to the last trial?" Again, responses were scored according to how many words recalled matched words present in the paragraph. JOIs were converted to *Z* score values for each participant, to minimize noninformative individual differences in what was an arbitrary rating scale.

Results

As in Experiment 1, we assessed accuracy of JOIs. JOIs were also compared with JOL differences, to investigate whether JOIs were based on changes in JOLs. Spearman correlations were calculated for JOLs, as well as biases in JOLs.

Judgments of improvement JOIs were very poor. The average correlation between JOI and actual improvement was ρ = .04 (min = −.67, max = .86, *SD* = .34), which was not significantly different from zero, *t*(57) = 0.97, *p* = .34. The correlation between JOIs and the increase in JOLs was ρ = .32 (min = −.27, max = .84, *SD* = .31), *t*(57) = 7.82, *p* < .01. As in Experiment 1, changes in JOLs were only a weak predictor of actual improvement, ρ = .21, (min = −.46, max = .83, *SD* = 31), significantly different from zero, *t*(57) = 5.11, *p* < .01. Similar to Experiment 1, the correlation between JOIs and changes in JOLs was significantly greater than the correlation between JOIs and actual improvement, *t*(57) = 6.15, *p* < .001.

Average JOIs were also found to increase with each trial, which is interesting because actual improvement decreased over time. JOIs were found to be moderately correlated with JOLs: The average ρ = .51, (min = −.37, max = .98, *SD* = .34), *t*(57) = 11.39, *p* < .01, and this relationship was stronger than the relationship between JOIs and increases in JOL, *t*(57) = 3.30, *p* < .01. Fig. 3 illustrates the relationship between JOIs and actual improvement:Judgments increased over trials, whereas improvement decreased over trials. Just as in Experiment 1, JOL difference scores seem flat across trials (Figs. 1, 3). A stark contrast between the two figures is the relationship between JOI and improvement; whereas in Fig. 2 the two lines seem somewhat parallel, in Fig. 3 they appear to go in opposite directions, with JOIs increasing while actual gains diminish.

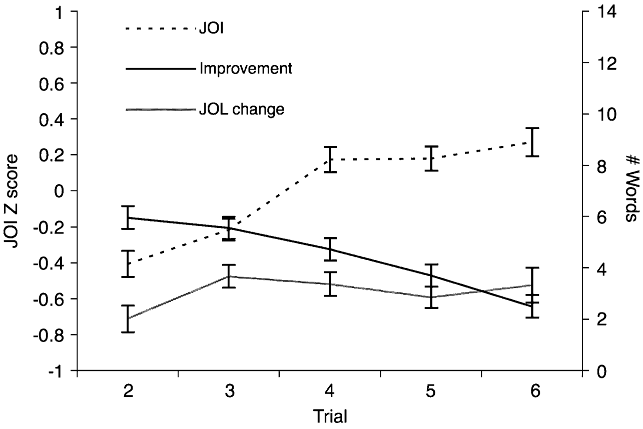

Fig. 3 Mean judgments of improvement (JOIs), improvement values, and changes in judgment of learning (JOL) per trial (Experiment 2)

Judgments of learning JOLs were compared with actual recall for each paragraph, but the average correlations were not significantly different, so the analyses were collapsed across paragraph type. The mean JOL Spearman correlation, across 18 observations per participant, was .63 (min = −.22, max = .96, *SD* = .29), significantly different from zero, *t*(57) = 16.70, *p* < .01.

Confidence bias Average bias exhibited the underconfidence-with-practice effect, replicating the results of Experiment 1, with mean bias becoming increasingly negative across trials. This relationship between JOLs and recall performance can be seen in Fig. 4. A repeated measures ANOVA showed a significant effect of trial on average bias, *F*(5, 865) = 30.64, *MSE* = 45.13, *p* < .001, η^2 = .15.

Discussion

This experiment did not yield evidence that students are sensitive to actual rates of improvement; the mean correlation was close to zero. Putting together the results from Experiments 1 and 2, people's ability to judge their own rate of improvement in learning seems very poor. The lack of a significant correlation in Experiment 2 gives support to the hypothesis that the small correlation found in Experiment 1 was an effect of the percentage rating scale for both improvement and learning. The percentage scale used in Experiment 1 enabled participants to respond consistently, whereas the 1 to 6 scale used in Experiment 2 did not present this easy opportunity.

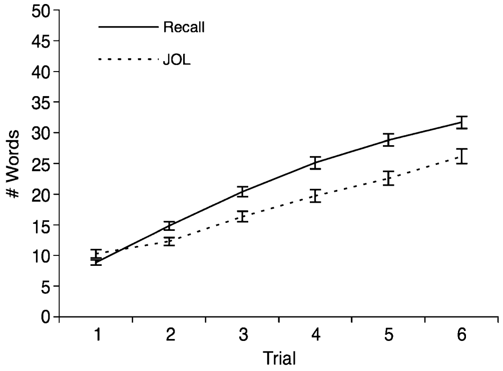

Fig. 4 Mean judgment of learning (JOL) values and recall per trial (Experiment 2)

As in Experiment 1, there was some consistency between JOIs and JOLs, as if people used the difference between the two most recent JOLs to formulate a JOI. This explanation does not suffice, however, because JOIs increased over trials and were more highly correlated with JOLs than with *changes* in JOLs. An alternate explanation of this effect is that JOIs may be influenced by other cues that are more salient over time, such as feelings of familiarity or fluency.

Experiment 3

Experiment 3 was conducted to assess JOIs for learning of somewhat more conventional stimuli: word pairs. Word pairs were chosen on the assumption that they might provide a wider range of improvement values and because other research has shown that the relatedness of word pairs influences JOLs (Dunlosky & Matvey, 2001; Matvey, Dunlosky, & Schwartz, 2006), with more related pairs having larger JOL values. We expected that more related pairs might also have higher JOIs if memory performance or fluency is a misleading, yet influential, cue for JOIs. There would also be increasing JOI values over time if this were the case. If the JOIs are influenced by changes in JOLs (as opposed to memory performance), JOI values should not increase over time, and there should be no differences between JOI values for related versus unrelated pairs (unless there are also substantial differences in the learning curves).

Method

Participants Forty-one students participated.

Materials Fifty related word pairs and 50 unrelated word pairs (nouns) were constructed from the Birkbeck word association norms (Moss & Older, 1996). Related pairs were defined as words with a high association (M = 34.9%, min = 19.6, max = 63.0), and unrelated pairs were defined as words with low association (M = 2.2%, min = 1.1, max = 2.4).

Procedure Each participant studied either unrelated or related pairs. The experiment was conducted using a computer program that presented the word pairs in random order. There were six trials in total. For each trial, participants viewed each word pair for 1 s. After all 50 pairs had been seen, participants were prompted to make made JOLs and JOIs. JOLs were prompted with the following: “You just studied many word pairs. Next, you will be presented with the first word of each pair, and will be asked to recall the second word. What percent do you think you can recall correctly?” JOIs were prompted with the following statement: “Compared to the last time you studied the list, how much do you feel that you improved? Please answer on a scale of 0 to 6, with 0 meaning 'I didn't improve at all', and 6 meaning 'I improved very much'.” Actual knowledge of the words was tested by showing the first word of each pair, and the participant typed in the second word in response.

Scoring Responses were scored according to how many correct target words were recalled. No points were taken off for misspellings. JOLs were converted from percentage to number of words. JOIs were standardized to Z score values for each participant.

Results

Judgments of improvement Most important, JOIs were compared with actual recall improvements in order to examine their accuracy. We found that JOIs were not just lacking in accuracy but that they were completely inaccurate, to the degree that they were negatively correlated with improvement. When people were improving the most (in the first few presentations), they gave their lowest JOI ratings, and when they improved the least (in the last few presentations), they gave their highest JOI ratings (Fig. 5). The average JOI ρ was −.37 (min = −1.0, max = 1.0, SD = .65), which was significantly different from zero, $t(40) = -3.61$, $p < .01$. As in the first two experiments, we also compared JOIs with changes in JOL values. The correlation between JOIs and the increase in JOLs was again significant, with an average ρ = .36 (min = −.5, max = 1.0, SD = .41), $t(40) = 5.56$, $p < .01$. However, JOL changes were found to be poor indicators of improvement, just as in Experiments 1 and 2, with an average correlation of .07 (min = −1, max = .97, SD = .59), $t(40) = 0.79$, $p = .43$. Average JOIs increased with each trial, so they

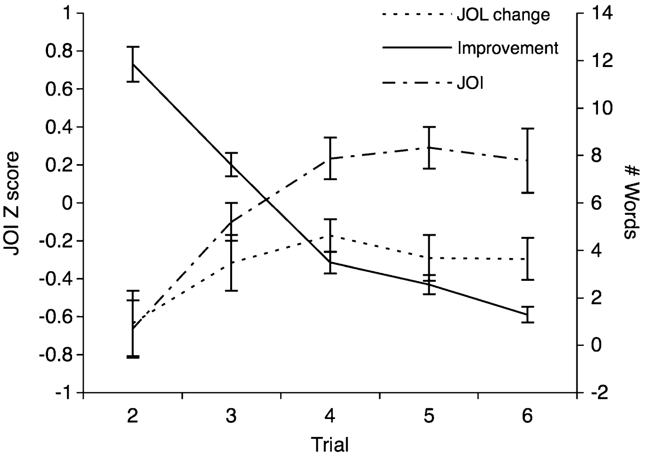

Fig. 5 Mean judgments of improvement (JOIs), improvement values, and changes in judgment of learning (JOL) per trial (Experiment 3)

were compared with JOL values (average ρ = .42, min = −1.0, max = 1, *SD* = .60), *t*(40) = 4.48, *p* < .01. The correlation between JOIs and increase in JOLs was not significantly different from the correlation between JOIs and JOLs, *t*(38) = 0.27, *p* = .79. Examining Fig. 5 reveals that JOIs no longer show a steadily increasing trend over time; average JOIs are highest on trial 5 and decrease on trial 6. This may be a result of high levels of recall on both trials 5 and 6; some participants may have been aware of this and, as such, may have given very low JOIs on the last study trial.

Average JOI–true-improvement correlations were compared for the two conditions, to see whether the relatedness of the words affected JOI accuracy. There was a significant difference, with the unrelated word pairs having a much larger, negative correlation than the related condition. The related condition mean ρ was −.15, whereas the unrelated pair mean ρ was −.58, *t*(39) = 2.19, *p* < .05. Participants gave the highest JOIs on later trials, whereas this relationship was much less severe for related words.

Judgments of learning Relative accuracy of JOLs was measured by comparing JOLs with recall performance and assessing whether or not the JOLs were affected by the relatedness of the word pairs. Significant correlations were found for both conditions, and the two conditions were not significantly different, *t*(39) = 0.143, *p* = .89. Correlations were calculated for each participant, using data from all six trials. The overall mean JOL correlation (across conditions) was .74, which was significantly different from zero, *t*(40) = 11.66, *p* < .01.

Confidence bias Biases were analyzed to see whether they differed for unrelated and related word pairs. Average biases were not different, but there was a significant effect of trial, *F*(5, 125) = 8.95, *MSE* = 52.40, *p* < .01, η^2 = .263. The average biases also showed the underconfidence-with-practice effect, as can be seen in Fig. 6. Similar to Experiment 1, there appeared to be increasing underconfidence with practice.

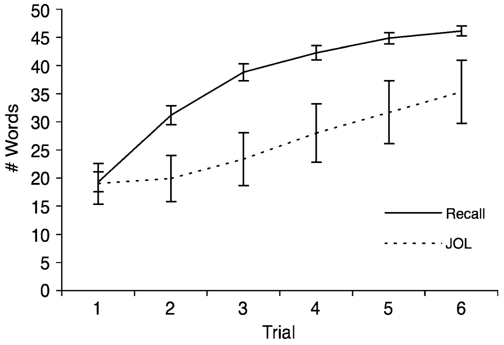

Fig. 6 Mean judgment of learning (JOL) values and recall per trial (Experiment 3)

Discussion

This experiment again showed that JOIs were poor; in this case, the correlations were significant but negative. We found that, as in the previous experiment, JOIs were significantly correlated with change in JOLs, giving further support to the hypothesis that change in JOLs has at least some influence on perceptions of improvement. Change in JOL was not significantly correlated with improvement, however, so this is not an ideal source of information for these judgments. JOIs were found to increase over trials and correlated significantly with JOLs (the JOI–JOL correlation was not significantly different from the JOI–JOL-change correlation, however). It is possible that both cues might be used to make JOIs: changes in JOL values and/or cues such as fluency that are related to overall performance. Using these problematic sources of information to make JOIs may be responsible for the negative correlation between JOIs and improvement. The results of this experiment are similar to those of Experiment 2, although more dramatic; it may be that fluency cues were stronger with these word pair stimuli, as compared with the *n*-gram paragraphs in Experiment 2 (since the stimuli were the only change from Experiment 2 to Experiment 3).

Experiment 4

In this experiment, we compared two different rating scales (percentage vs. absolute number of words), as well as different types of JOIs. One might expect that judgments in terms of number of words learned would be easier and more successful, due to their simplicity. Judgment types were either postdictive (after a study trial) or predictive, occurring before the next study trial—that is, “If you were to study this list for another minute, how much do you think you would improve? Answer: I think I would learn another ___% of the material.” Predictive JOIs may be more helpful to learners than postdictive JOIs for study decisions, and if students do make predictive JOIs (and not postdictive ones). they should have better accuracy for this kind of judgment. Predictive JOIs may be more likely when it is determined whether further study would be worthwhile. Both type of judgment (predictive JOI, postdictive JOI, or JOL) and type of scale (percentage or number of words) were manipulated between subjects.

Method

Participants One hundred seventy-one students from the participant pool at the University of California, Merced, volunteered to participate for class credit. The number of participants in each condition was as follows: 32 making prospective, percentage scale JOIs, 31 making prospective, numerical JOIs, 34 making postdictive percentage JOIs, 30 making postdictive numerical JOIs, 23 making percentage scale JOLs, and 21 making numerical JOLs.

Materials A list of 50 Swahili–English word pairs was constructed from the Nelson and Dunlosky (1994) norms. These stimuli have been used in much previous metacognitive research. The list of word pairs was constructed to include a range of difficulty.

Design and procedure The experiment consisted of six trials, with each trial consisting of study, judgment, and test phases. All manipulations were between subjects. The design was 3 judgment types (predictive JOI, postdictive JOI, or JOL)×2 scales (absolute number or percentage), so each participant experienced only one judgment type and one scale type, for a total of six conditions. For the prospective JOI conditions, each trial consisted of judgment–study–test (with the exception of the first trial, which did not include a judgment). Judgments were solicited with the question, "If you were to study this list for another minute, how much do you think you would improve? Answer: I think I would learn another ___[% or words] of the material."

For the postdictive JOI conditions, each trial consisted of study–judgment–test (with the first trial not including a judgment). These judgments were made after the question, "Compared to the previous trial, what percent more of the list will you be able to recall? Answer: I will recall another ___ % of the list" or "Compared to the previous trial, how many more words of the list will you be able to recall? Answer: I will recall another ___ words of the list."The JOL conditions consisted of study–judgment–test. Participants were asked, "What percent of the list will you be able to recall? Answer: I will recall __ % of the list" or "How many words of the list will you be able to recall? Answer: I will recall ___ words of the list."

Scoring Responses on the test trial were marked as correct if they matched the target word. No points were deducted for misspellings. Percentage judgments were converted to number of words for the purpose of analysis.

Results

Preliminary analyses revealed that some participants were not successful in learning Swahili–English word pairs. On this basis, 37 participants were removed from analyses due to not entering any judgments, responding with the same judgment on each trial, not learning more than five words after all six trials, or technical errors. There were totals of 25 participants in the predictive-JOI–percent-judgment condition, 23 in the predictive-JOI–numerical-judgment condition, 25 in the postdictive-JOI–percent-rating condition, and 25 in the postdictive–numerical-rating condition. Finally, 15 participants gave percentage JOL judgments, and 20 gave numerical JOL judgments.

Judgments of improvement JOIs were compared with actual recall improvement, with no significant correlation found for either judgment type or for either scale type. For predictive JOIs, neither percentage (average ρ = .11, min = −.89, max = .95, *SD* = .50) nor numerical (average ρ = .06, min = −.98, max = 1.0, *SD* = .52) judgments were significantly different from zero; for postdictive JOIs, percentage (average ρ = .05, min = −.89, max = .95, *SD* = .51) and numerical (average ρ = .04, min = −.89, max = .89, *SD* = .52) judgments were also nonsignificant.

Changes in JOLs are a possible basis of JOIs. In this experiment, JOIs and JOLs were made between subjects to avoid influencing participants toward inferring JOIs this way. A between-subjects repeated measures ANOVA comparing mean JOIs and mean JOL difference scores by trial suggests that participants may not have been covertly making JOLs and using them to infer JOIs, $F(1, 125) = 13.30$, $p < .001$, $\eta^2 = .096$.

JOI bias Absolute accuracy for JOIs was examined, and no significant differences in bias were found for judgment type or scale type, although there was a significant effect of trial, $F(4, 340) = 9.13$, $MSE = 25.34$, $p < .001$, $\eta^2 = .097$. Percentage judgments were converted into number of words for the purpose of comparison. There appeared to be increasing confidence with trial, as illustrated in Figs. 7 and 8, which corroborates the results from the previous experiments. Average total bias across participants was −.2872 (min = −7.10, max = 20.0, *SD* = 4.67.

The low JOI biases may suggest that JOIs were very close to actual improvement, despite the low correlations. However, absolute accuracy (Schraw, 2009) of JOIs (average squared deviations between JOIs and improve-

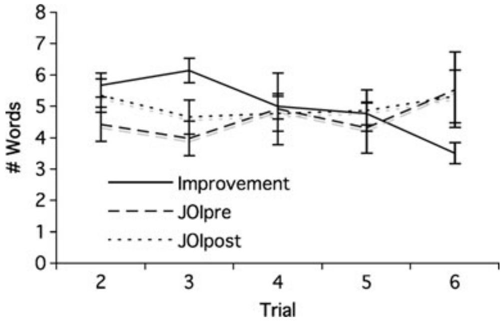

Fig. 7 Average judgments of improvement (JOIs) and improvement values per trial, by judgment time (Experiment 4)

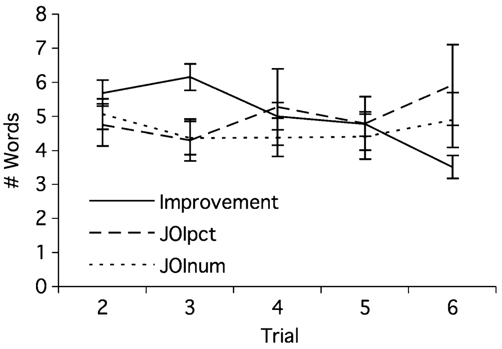

Fig. 8 Average judgments of improvement (JOIs) and improvement values per trial, by scale type (Experiment 4)

ment) shows a large discrepancy. The average value of absolute accuracy across participants was 45.68 (min = −78.8, max = 897.0, *SD* = 115.21. No significant differences in absolute accuracy were found for judgment time, *t*(96) = −0.262, *p* = .091, or for judgment type, *t*(96) = −0.45, *p* = .66.

If we compare these measures with those in Experiment 1, in an exploratory analysis, we find that the bias was significantly more overconfident in Experiment 1 (*MD* = −3.72), *t*(141) = −3.85, *p* < .01. Also, the absolute accuracy measures illustrate significantly better accuracy in Experiment 4 as well (*MD* = −64.31), *t*(142) = −2.48, *p* = .014.

Judgments of learning JOLs were compared with recall performance, and significant correlations were found for both percentage (ρ = .61, min = −.58, max = 1.0, *SD* = .56), *t*(15) = 4.38, *p* < .001, and number rating (ρ = .42, min = −.88, max = 1.0, *SD* = .68), *t*(18) = 2.67, *p* < .015 conditions. There was no significant difference between the two conditions, *t*(33) =*0*.36*, p* = .72.

Confidence bias JOL bias was assessed by computing the difference between JOLs and actual recall. For percentage judgments, the percentage was converted to number of words. Biases were also analyzed to see whether they differed for judgment type. There was a trend toward more underconfidence for percentage judgments, *F*(1, 32) = 3.86, *MSE* = 111.22, *p* = .058, η^2 = .108. There was a significant effect of trial, *F*(5, 160) = 61.33, *MSE* = 44.76, *p* < .001, η^2 = .657. Similar to the preceding experiments, there was increasing underconfidence with practice, but with a small upturn on the last trials, as can be seen in Figs. 9 and 10 (percentage scale judgments) and 10 (numerical scale judgments).

Discussion

In this experiment, we failed to find a significant correlation between JOIs and actual improvement. The type of scale (percentage or number of words) did not make a difference for judgment accuracy (either relative or absolute), nor did the time of judgment. The predictive JOIs were no more accurate than the postdictive JOIs, whereas JOLs made before and after a test have been found to differ (Hacker, Bol, Horgan, & Rakow, 2000). One possible reason JOL values differ between pre- and posttest is that there are more cues on which to base posttest JOLs, as compared with pretest JOLs (e.g., once the participants have taken the exam, they know what the actual questions were, how quickly the answers came to mind, etc.). In contrast, JOIs do not generally get feedback; JOLs do get feedback over time, as students are given grades on assignments and exams (and this feedback may also help savvy students to learn what cues are more informative). To get feedback on a JOI, it would be necessary to test oneself before and after a study session and then calculate how much more information was known, as compared with prestudy. Instead, students will probably rely on subjective feelings, like increased fluency of information, how answers come to mind faster, and reduced feelings of anxiety about exams, but without feedback, students cannot learn whether or not these feelings are actually informative.

We also found that, in this experiment (which used word pairs), JOIs were more accurate and less biased than the JOIs

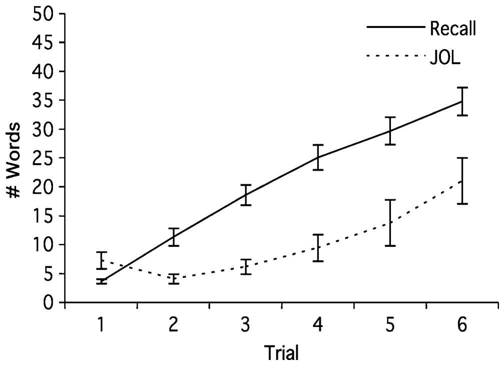

Fig. 9 Mean judgment of learning (JOL) values and recall per trial (Experiment 4), percentage scale converted to number of words

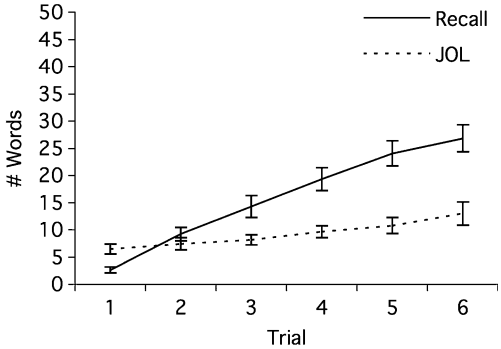

Fig. 10 Mean (judgment of learning (JOL) values and recall per trial (Experiment 4), numerical scale

made in Experiment 1 (which used *n*-gram paragraphs). There are two possible explanations for this difference, since there were two key differences between the experiments: In Experiment 1, participants made JOLs and JOIs, whereas in Experiment 4, they made only one type of judgment, and the stimuli were different. Making both JOLs and JOIs may have biased learners (toward overconfidence) and contributed to reduced accuracy. Alternately, the difference may be due to the familiar format of learning foreign vocabulary a word at a time. Students might be more accurate in this scenario because of more experience with this type of study material. Unfortunately, absolute accuracy measures could not be calculated for the other two experiments, due to the rating scales employed.

General discussion

In these experiments, we found little or no evidence to suggest that people can accurately assess their rate of improvement. People generally thought that they were improving the most when they were learning the least (on later trials) and thought that they were improving the least when they were actually improving the most (on earlier trials). Experiment 1 showed a small positive correlation between JOIs and actual improvements in recall, but JOIs were not significantly different across trials. This means that it was likely that some people were giving consistent responses, perhaps due to the rating scale (by this we mean giving the same JOI on each trial and always raising their JOL by that amount). This interpretation is supported by the findings in Experiment 4: Participants in the percentage rating condition had nonsignificant correlations between JOIs and improvement, as opposed to the significant correlation in Experiment 1; unlike in Experiment 1, they made only JOIs, not JOLs and JOIs. If participants show a positive correlation only when giving *both* JOIs and JOLs, it can be inferred that those giving both JOIs and JOLs are responding differently. In Experiments 2 and 3 (in which JOIs and JOLs were given on different scales), we also found a nonsignificant correlation, and the average JOIs increased across trials, whereas improvement decreased.

From the point of view of a learner choosing what to study, the judgment of interest may be a predictive JOI, made *before* studying, rather than a postdictive JOI, made *after* studying. We found in Experiment 4 that it made no difference whether JOIs were predictive or postdictive; there was no difference between the two kinds of judgments in their accuracy. Perhaps other changes to the question wording could provide more informative cues, but whether the judgment is a prediction of future improvement or an assessment of recent improvement does not seem to be important.

This finding (that JOIs are not indicative of improvement) is surprising, since it suggest that learners may be unable to use JOIs as a cue to stop studying a particular item (when improvement rate drops), or, if learners were to use these judgments, it might lead to counterproductive study patterns. This finding suggests that models of study time allocation that assume access to the rate of improvement (Metcalfe & Kornell, 2005; Son & Sethi, 2006) are at risk. The ability to make accurate JOIs would certainly be valuable in many instances, especially under time pressure where students want to ensure that their efforts do not go to waste. At the very least, it remains a task for future research to document any situations in which learners can make accurate JOIs. If learners cannot explicitly give accurate JOIs, it is difficult to see how they would be used during study to make informed decisions, and learners may be better off sticking to a discrepancy reduction approach to their studies.

Some monitoring and control processes may be implicit (Reder & Schunn, 1996), and perhaps people do have the ability to modify behavior on the basis of rates of improvement. If this is so, it is possible that people just cannot verbally report accurate JOIs. Implicit JOIs would still have implications for models of study time allocation: It is assumed that monitoring is explicit, as well as decisions about what to study and when to stop. This explicitness allows people to make decisions about their behavior on the basis of their current goal, and it is not clear that implicit metacognitive processes could be used to *explicitly* guide behavior toward different goals. It would also be difficult to teach students study strategies if some processes and decisions are not explicit while others are explicit. Note that in meta-memory research, it is routine to ask explicitly for JOLs, and although such judgments do have some systematic biases, overall, they do reflect degree of learning (Finn & Metcalfe, 2008;Koriat et al., 2002; Meeter & Nelson, 2003; Nelson & Dunlosky, 1991).

The hypothesis that multiple JOLs are a basis for making JOIs appeared to have a small degree of merit, since JOIs were more correlated with changes in JOLs than with actual improvement (when both judgments were made), although in terms of absolute accuracy, there were large discrepancies between JOL change values and JOIs. Changes in JOL were not a reliable cue for improvement, because they did not accurately reflect changes in recall. In order for JOIs to be accurate, they would have to rely on cues that were genuinely informative of improvement.

If the findings in these experiments are truly representative of learners' abilities, students may be unable to allocate their study time optimally, persevering when they should move on, and moving on when they should persist (because of low JOIs given in the beginning, when the most learning occurred). Son and Sethi (2006) concluded that a simple rule of choosing the task with the steepest current learning curve would be optimal with diminishing returns scenarios, but even this is not always optimal if the learning curves are

S-shaped rather than concave; even if students could make JOIs, this would not guarantee optimal allocation. If students managed to make good allocation decisions on the basis of JOLs, without accurate JOIs, learning may be suboptimal because of ideas about whether or not further gains can be made on a particular item. Stopping decisions may indeed be suboptimal, as evidenced by a study by Kornell and Bjork (2008), who looked at flashcard dropping while studying English–Swahili translations. Flashcard dropping is a custom among many students, when they decide to exclude some cards from further study. Flashcards might be dropped because an item is already learned or because the item has no chance (presumably) of being learned. Students who dropped items had their performance suffer, in comparison with those who did not follow that strategy. If those items were truly unlearnable, overall performance would benefit from their exclusion. This suggests that students were moving on because their perceptions of improvement rate were incorrect. Clearly, some students are able to attain high levels of learning, so we may assume that they have more accurate JOIs, their threshold for giving up is different, they return to difficult-to-learn items once they have mastered other material, or they use strategies to learn those items differently.

It is important to investigate learners' abilities to assess their learning and improvement while studying, because such information will aid in the development of study strategies and skills training. It may be possible to improve the calibration of JOIs by directing attention to relevant cues, such as giving performance feedback or by having people switch between two lists of items so that there is a source of comparison. The ability to make accurate JOLs may be necessary for optimal study time behaviors, and students who possess this ability will be able to make more informed, better choices. If judgments are not highly correlated with actual learning, learners may have difficulty in knowing what areas need work and when to move on.

Acknowledgments We thank Lisa Son for comments on a previous version of the manuscript and Melissa Newland for research assistance. A preliminary report of Experiment 4 appears in Townsend and Heit (2010).

Appendix A

Sample 50 word 4-gram paragraph.

sole legal political party recently there has been release of growth hormone back into working order early and late promoters rein on public spending never seen anything like wanted to know how care in the community does not make sense next time you go academic or other problems people who

Sample 50 word 6-gram paragraph.

are determined to bring unemployment down ethipoian people's revolutionary dramatic front joint meetings with our sister society other parliamentary material as is necessary registered foreign lawyer or a recognized must be taken into account when scottish national gallery of modern art car audio specialist who will top left

Sample 50 word 8-gram paragraph.

optional qualifying subject for the first level certificate and without risks to health when properly used while any motor vehicle insured by this policy old man has appeared in court charged with is a clear expression of local preference supported already served five weeks on remand he was cancel this

Sample related word pairs.

Hand	Glove
Bread	Butter

Sample unrelated pairs.

Vow	Reason
Zoo	Bars

References

Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control process view. *Psychological Review, 97*, 19–35.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1180–1191.

Dunlosky, J., & Lipko, A. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition, 32*, 779–788.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34.

Fletcher, W. "Phrases in English" database. Retrieved October 2006, from http://pie.usna.edu/index.html.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*, 125–136.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32*, 609–622.

Matvey, G., Dunlosky, J., & Schwartz, B. (2006). The effects of categorical relatedness on judgments of learning (JOLs). *Memory, 14*(2), 253–261.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*, 123–132.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174–179.

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.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Moss, H., & Older, L. (1996). *Birkbeck word association norms*. Hove, U.K.: Psychology Press.

Nelson, T. O., & 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., & Dunlosky, J. (1994). Norms of paired associate recall during multitrial learning of Swahili–English translation equivalents. *Memory, 2*, 325–335.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213.

Reder, L., & Schunn, C. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. Reder (Ed.), *Implicit memory and metacognition*. Mahwah, NJ: Erlbaum.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.

Son, L. K., & Sethi, R. (2006). Metacognitive control and optimal learning. *Cognitive Science, 30*, 759–774.

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.

Townsend, C., & Heit, E. (2010). Metacognitive judgments of improvement are uncorrelated with learning rate. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.