Metacognitive Judgments of Improvement are Uncorrelated with Learning Rate

Corinne L. Townsend (ctownsend@ucmerced.edu)
Evan Heit (eheit@ucmerced.edu)
Department of Social and Cognitive Sciences, 5200 North Lake Road
Merced, CA 95343 USA

Abstract

Being able to assess one's own learning rate is essential for optimal learning. Can students accurately assess their learning rate, and is the timing of judgments of improvement important? In this experiment, students were to estimate their learning rate on each trial, either before the trial, or immediately after. If students typically make these judgments before embarking on further study, accuracy might be greater in the predictive judgment condition. No evidence was found that students could accurately judge improvement, in either condition. Implications for models of self regulated learning are discussed in light of these findings.

Keywords: metacognition; self regulated learning; metamemory.

Introduction

Judgments of improvement are metacognitive judgments regarding one's speed of learning. These can be thought of as a student's estimation of how quickly he or she is acquiring more knowledge, or put into practical terms, how useful a given amount of study time is likely to be. These judgments are crucial, as the ability to estimate one's learning rate will affect how well students are able to allocate their time optimally during self regulated learning, which then in turn will influence academic achievement. One such example of how these judgments might inform study time allocation is in the proximal learning model (Kornell & Metcalfe, 2006; Metcalfe, 2002; Metcalfe & Kornell, 2005). In this model, it is proposed that decisions seek to maximize the rate of return per time studied, and those regarding when to switch topics or stop studying may rely on judgments of improvement. This way, students can avoid working in vain while not making progress, and instead move on to more fruitful pursuits. The proximal learning model contrasts with an earlier account, the discrepancy reduction model (Thiede & Dunlosky, 1999), which assumed that students focused on the most difficult items first, and stopped when material reached a satisfactorily high level of learning, thus depending on JOL level to determine stopping times. Additionally, Son and Sethi (2006, in press) have derived mathematically that the most optimal behavior is usually to focus on the items with the highest current rate of return, consistent with the proximal learning model. There is some evidence to support this account, which is sometimes referred to as the shift-to-easier-materials effect; this is the finding that under time pressure, students prioritize by studying the easiest (high rate of return) items first, before moving on to more difficult material (Metcalfe & Kornell, 2003; Dunlosky & Thiede, 2004; Kornell & Metcalfe, 2006).

However, there is not yet evidence to support the idea of using improvement rates to inform decisions, and current research has not shown that students have the ability to make judgments of improvement accurately in any sense. In our previous work (Townsend & Heit, 2010), participants estimated their amount of improvement after completing each study trial, in a repeated series of study trials for a set of verbal materials. Students' judgments of improvement (or JOIs) were not significantly correlated with actual improvement rates, and in some cases were even negatively correlated. The negative correlation occurred when judgments of learning and judgments of improvements were made using different rating scales, which prevented participants from attempting to infer their JOIs from their judgments of learning. Work by Kornell and Bjork (2009) has also shown that students have difficulties estimating how much they will learn during one or more study trials, dramatically underestimating the usefulness of study. They referred to this type of judgment as a prediction of learning, but the concept is the same. Thus, there is reason to be concerned that students are not able to make the metacognitive judgments that would lead to optimal learning.

Students' post-study JOIs showed an interesting shift from underconfidence to overconfidence over the course of learning (Townsend & Heit, 2010), but predictive JOIs that estimate the fruitfulness of further study may or may not show the same pattern. It is important to assess predictions of future learning (predictive, pre-study trial JOI) rather than just a postdictive assessment of learning during a study trial, as decisions regarding study time allocation may depend more on how much is expected to gain from further study, rather than how much was gained from recent study. This experiment was designed to compare the two conditions to evaluate how (or whether) timing affects JOIs. For comparison, we also collected judgments of learning (JOLs, which are predictions of recall test performance) from an additional group of participants, to compare the relative accuracy of JOIs and JOLs. For example, whereas it may be too difficult for students to judge their level of improvement, they still may be able to judge their level of learning in absolute terms.

Experiment

In this experiment, we compared two different rating scales (percentage vs. absolute number of words), as well as different types of improvement judgments. One might expect that judgments in terms of number of words learned would be easier and more successful, due to their simplicity as well as their close nature to other judgments of optimal foraging (Gigerenzer & Hoffrage, 1995). Judgment types were either postdictive (made after a study trial) or predictive, occurring before the next study trial, i.e. "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 informative than postdictive JOIs for study decisions, and if students do make predictive JOIs (and not postdictive) they should have better accuracy for this kind of judgment. It may be more likely that students would make predictive JOIs, especially if they are determining whether or not further study would be worthwhile. Type of judgment (Predictive JOI, Postdictive JOI, or JOL) and type of scale (percent or number of words) were both manipulated between subjects.

Method

Participants. 171 students from the subject 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, percent scale JOIs, 31 making prospective, numerical JOIs, 34 making postdictive percent JOIs, 30 making postdictive numerical JOIs, 23 making percent 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 a study phase, judgment phase, and test phase. All manipulations were between subjects. The design was 3 judgment types (predictive JOI, postdictive JOI, or JOL) by 2 scales (absolute number or percentage), so each subject only experienced one judgment type and one scale type for a total of 6 different 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 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 either not entering any judgments, responding with the same judgment on each trial, not learning more than 5 words after all 6 trials, or technical errors. There were a total 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 and 25 in the postdictive numerical rating condition. Finally, 15 participants gave percentage JOL judgments, and 20 gave numerical JOL judgments.

Judgments of Learning. Judgments of learning were compared to 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 conditions (ρ = .42, *min* = -.88, *max* = 1.0, *SD* =.68, t (18) = 2.67, $p < .015$). There was no significant difference between the two conditions, t (33) =.36, p = .72.

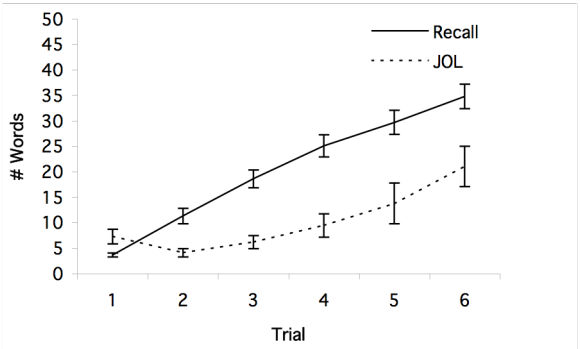

Figure 1: Mean JOL values and recall per trial, percent scale converted to number of words.

Confidence Bias. Relative accuracy of JOLs is not particularly informative, since it is reasonable to assume that participants understand that performance generally increases with each trial. For this reason, we examined absolute accuracy of these judgments as well. Absolute accuracy (in terms of bias) was assessed for JOLs 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 if 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, $\eta 2$ = .108. There was a significant effect of trial, F (5, 160) = 61.33, MSE = 44.76, p < .001, $\eta 2$ = .657. Similarly to previous work that included both JOLs and JOIs (Townsend & Heit, 2010), there appeared to be increasing underconfidence with practice, but with a small upturn on the last trials, as seen in Figures 1 (percentage scale judgments) and 2 (numerical scale judgments).

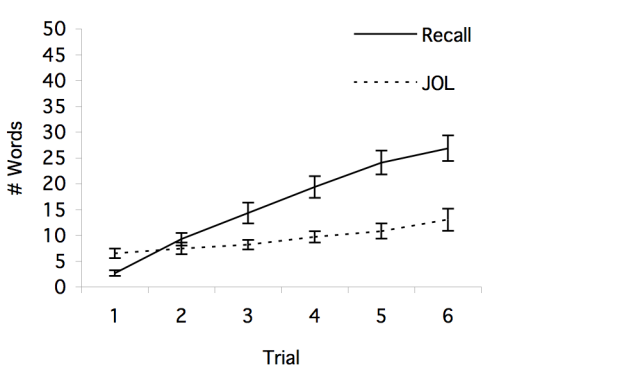

Figure 2. Mean JOL values and recall per trial, numerical scale

Judgments of Improvement. Judgments of improvement were compared with actual improvement, with no significant correlation found for either judgment type or either scale type. For predictive JOIs, neither percentage (average ρ = .11, min = -.89, max = .95, SD = .50) nor numerical judgments (average ρ = .06, min = -.98, max = 1.0, SD = .52) 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 non-significant.

Changes in JOLs are a possible basis of judgments of improvement. In this experiment, JOIs and JOLs were made between subjects to avoid influencing participants towards inferring JOIs this way. A between subjects repeated measures analysis of vaiance 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) = 13.302, p < .001, $\eta 2$ = .096.

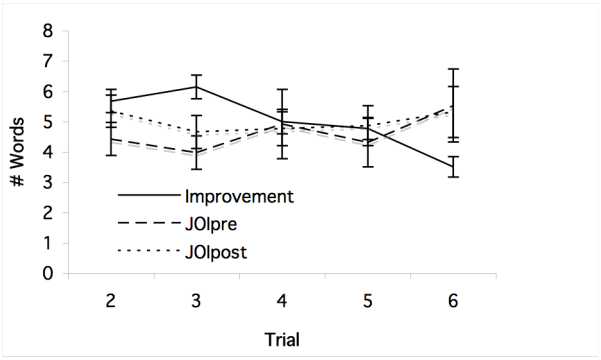

Figure 3. Average JOIs and improvement values per trial, by judgment time.

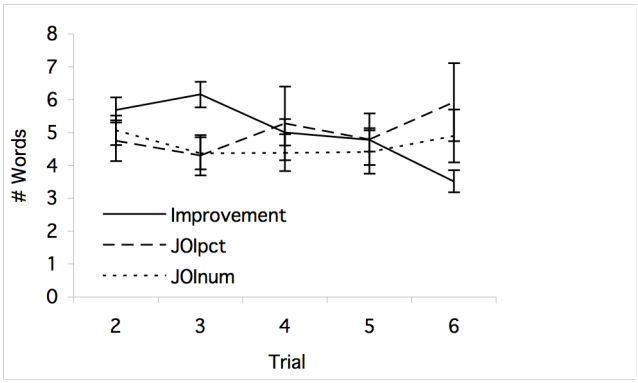

Figure 4. Average JOIs and improvement values per trial, by scale type.

JOI Bias. Absolute accuracy for JOIs was examined, and no significant differences in bias were found for judgment type or scale type, though there was a significant effect of trial, F (3.05, 259.51) = 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 Figures 3 and 4 which corroborates with the results from previous work (Townsend & Heit, 2010) which found that JOIs increased with trial, and in that case, were correlated with JOLs. Average total bias across participants was -.2872, min = -7.10, max = 20.0, SD = 4.67.

The low values for JOI biases may lead one to conclude that judged improvement was very close to actual improvement, despite the low correlations. This would be an erroneous conclusion, however, because an examination of the absolute accuracy (Schraw, 2009) of JOIs (average squared deviations between JOIs and improvement) 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) = -.262, p = .091, or for judgment type, t (96) = -.45, p = .66.

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, nor did the time of judgment; predictive JOIs were no more accurate than postdictive JOIs. In comparison, JOLs made before and after a test have been found to differ (Hacker, Bol, Horgan, & Rakow, 2000). One possible reason for why JOL values differ between pre and post test is that students may routinely make JOLs, assessing how well they are likely to do on exams, and then make post test judgments of performance, e.g. "I think I aced the exam!", and there are more cues with which to base posttest JOLs on, as compared to pretest JOLs (e.g. once they've taken the exam, they know what the actual questions were, how quickly the answers came to mind, etc). In contrast, JOIs may be a judgment that is not made very often, without as many informative cues, and is a judgment on which students don't 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 compared to pre-study. This is a cumbersome and unlikely task for a student to perform; more likely, students will rely upon subjective feelings, like how much more fluent the information seems, how answers may seem to come to mind faster, and perhaps even reduced feelings of anxiety about exams—and without feedback, students cannot learn whether or not these feelings are actually informative.

Other research that has looked at JOI predictions also found judgments to be uncorrelated with actual learning. In Kornell and Bjork (2009), they found a large degree of underconfidence in predictions of learning. Participants in their experiment made their predictions on the first study trial, so their results might not predict how learners will feel about the fruitfulness of study if asked beyond that point. For example, if asked initially about how much they will learn in four study trials, they may be incredibly unoptimistic, but if the students were to be asked after two study trials, they may have different predictions, perhaps based on their subjective experience of the task becoming easier. Kornell and Bjork (2009) showed that JOIs were inaccurate, observing that students were incredibly underconfident when it came to predicting future learning beyond one study trial, but they only experienced one trial at the time of judgment, and did not yet have the experience of repeating study (which is the very thing they are asked about). In our experiment, students made their judgments on each study trial, and a different pattern emerged: a shift from underconfidence in early trials, which is consistent with their results, to overconfidence in later trials. Unfortunately it would seem that experience with the task does not improve JOI accuracy at all, but rather shows a more interesting pattern of inaccuracy.

The inaccuracy of these judgments of improvements has significant implications for models of study time allocation that rely on them; specifically, it is highly unlikely that student behavior would approximate optimality by the use of JOIs. The inability to accurately assess the speed at which one is learning means that learners could not accurately make JOIs to reliably know if further study would be made in vain, and when time would be better spent on a different item or task, leading to much wasted time. Even worse, if students do make JOIs and base decisions on them, they may make bad decisions. Students may give up early in the process of learning (as JOIs are underconfident in the beginning of study), and instead work on better-learned material, on which they persist longer than they should due to overconfidence in the later periods of study. This would lead to very inefficient studying, and could have disastrous results- yet many students do manage to achieve reasonable performance in their courses, so this cannot be the whole story. It may be the case that stopping and switching is based not on explicit JOIs but is done implicitly; Reder and Schunn (1996) suggest that much of metacognitive monitoring and control may actually be implicit. Supporting this somewhat, Payne, Duggan, & Neth (2007), found that in a task switching situation where people performed two different tasks (scrabble and word search), they were sensitive to rate of rewards, and able to spend more time in the easier task. This possibility of implicit control will be investigated in future research that more closely resembles a learning situation, rather than tasks in which participants have such obvious successes and failures.

Whether they are informed by explicit JOIs or implicit control, decisions may also be based on other factors: subjective feelings such as frustration and fatigue, idiosyncratic rules (e.g. study for X amount of time, or until I fall asleep, or all day before the exam), JOLs, or on the results of self-testing. It would also be adaptive if students do not simply stop studying low JOI material, because no learning would take place, and that is not always a viable option. In the cases where the item has a low JOI and a low JOL (meaning that the item is not well learned, and is not being learned very quickly), the ideal behavior would be to change strategies, seek other sources of learning, or the guidance of the instructor.

We also leave open the possibility that students could be taught better study habits, and to make more accurate JOIs. In the framework of Stanovich (2009), otherwise intelligent students may act suboptimally because they lack the "mindware" that allows them to reflect on their own level of learning, simulate the possible consequences of further studying, and override their default study strategies. We are hopeful that at least some of these abilities are teachable. Future research will examine these possibilities.

References

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory and Cognition, 32*(5), 779-788.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

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.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449-468.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609-622.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363.

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*(4), 530-542.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463-477.

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

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General, 136*(3), 370-388.

Reder, L., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is goverened by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition (pp. 45-78)*. Mahwah, NJ: Erlbaum.

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

Son, L. K., & Sethi, R. (2006). Metacognitive control and optimal learning. *Cognitive Science, 30*(4), 759-774.

Son, L. K., & Sethi, R. (in press). Adaptive learning and the allocation of time. *Adaptive Behavior.*

Stanovich, K.E. (2009). *What intelligence tests miss: The psychology of rational thought.* New Haven: Yale.

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. L. & Heit, E. (2010). *Judgments of Learning and Improvement.* Manuscript submitted for publication.