Do repeated judgments of learning lead to improved memory?
DO REPEATED JUDGMENTS OF LEARNING LEAD TO IMPROVED MEMORY?

Max Larsson Sundqvist

Judgments of Learning (JOL) that are made after a delay, instead of immediately after study, are more accurate in terms of predicting later recall (the delayed JOL effect). The Self-Fulfilling Prophecy (SFP) theory explains the delayed JOL effect as the result of a testing effect. In the current study we tested the prediction that performing delayed JOLs leads to a memory improvement. During learning, 79 participants studied Swahili-Swedish word pairs, immediately followed by a cued recall test, and then made either one single or three repeated, spaced JOLs. A final cued recall test was given after either 5 minutes or 1 week. Making repeated JOLs did not increase memory performance compared to the single JOL condition, hence lending no support to the SFP theory. However, making repeated JOLs did improve their relative accuracy, which suggests that the delayed JOL effect mainly concerns memory monitoring and not performance.

In order to successfully regulate study efforts, people will need to engage in memory monitoring and control. For instance, being certain that you know something relies on memory monitoring, and knowing that you don’t know something will likely result in an effort to restudy the material that needs to be learned. This example shows how monitoring and control regulate one another, or how people’s metacognitive judgments and assessments affect their subsequent study behavior(e.g., Metcalfe & Finn, 2008). Judgments of learning (JOL; see Metcalfe & Dunlosky, 2008, for a review) are one example of such metacognitive assessments, and they serve as indicators of what a person believes she knows and what she does not. One key aspect of these judgments is, of course, their accuracy. In other words, only judgments of learning that actually reflect the current state of learning will serve to successfully regulate study efforts. For example, feeling like you know something that, it turns out, you don’t know, would be considered a very poor and inaccurate JOL. Therefore, knowing what makes a JOL accurate or not, is essential.

In the field of metacognitive research, one of the more robust findings is the delayed JOL effect, which states that judgments of learning have considerable predictive value in terms of subsequent recall on a memory task when made at a delay rather than immediately after study (Dunlosky & Nelson, 1992; Jang & Nelson, 2005; Nelson et al., 2004). While there are several different accounts of the delayed JOL effect, they all converge on one crucial point: the delay between study and JOL effectively rules out the possibility that the judgment itself is based on information retrieved from short-term memory. Making a delayed JOL will thus cause people to make a diagnostic retrieval attempt from long-term memory. And since the diagnostic aspect of the retrieval attempt is only present when accessing information from long-term and not short-term memory, this is believed to be the reason why delayed (but not immediate) JOLs are so very accurate. At face value, it appears rather intuitive; if the conditions, under which the judgments of learning are made, mirror the conditions under which the retrieval will take place; this congruity is likely to increase the validity of the judgments of learning.
The preferred means of evaluating the accuracy of delayed JOLs is by calculating an item-by-item Goodman-Kruskal gamma correlation (see Nelson, 1984) between the JOL and recall performance. Whereas absolute JOL accuracy is measured simply by subtracting JOL magnitude from actual recall performance (which provides an index of over/under-confidence), the gamma correlation provides a measure of relative JOL accuracy. The delayed JOL effect, then, describes an increase in the mean relative JOL accuracy for JOLs that are made at a delay compared to immediate JOLs.

Theories of the delayed JOL effect

Over the course of the last decades, different theories have emerged which address the question of why delayed JOLs are substantially more accurate than immediate JOLs. These accounts are explained below.

**The Monitoring Dual Memories (MDM) hypothesis.**

The MDM hypothesis (Nelson & Dunlosky, 1991) suggests that making JOLs involves attempted retrieval of the sought-after information, and that the information is gathered from two memory systems, namely long-term memory and short-term memory. Thus, the outcome of a JOL is governed by the monitoring of these dual memories. Moreover, the timing of the JOL affects what kind of information is retrieved from the dual memories. If the JOL is immediate, the judgment will be affected by information that is stored in short-term memory, and because that information is ephemeral, it will most likely have vanished before the final test. Thus, the immediate JOL will prove inaccurate. If the JOL is delayed, it will instead use information stored in long-term memory as its decision basis, and since that information is likely to still exist at the final test, the JOL will more accurately predict final recall.

**The transfer-appropriate monitoring hypothesis.**

The transfer-appropriate monitoring hypothesis (Begg et al., 1989; Dunlosky & Nelson, 1997) emphasizes the degree of equivalence between processes involved at the time of the judgment and the time of the final test. That is to say that the retrieval processes involved in making the JOL should match (or transfer onto) the retrieval processes associated with taking the final test. This account is similar to Tulving and Thomson’s (1973) encoding specificity principle, which points to the contextual equivalence or overlap between, for instance, judgment and testing. In other words, making a judgment of learning in a setting that is as homogeneous as possible to that of the final test will increase the accuracy of that judgment. Therefore, since making delayed JOLs is contextually and operationally more similar to a delayed test than is an immediate JOL, the delayed JOL will be more accurate, and this is especially the case when retrieval is effortful.

The transfer-appropriate monitoring and the MDM hypothesis are similar insofar as both underline the importance of congruency between the judgment and the final test, whether it be by information or context. However, the transfer-appropriate monitoring hypothesis hits a snag when its predictions are put into practice. For instance, recognition of stimulus-response pairs are more accurately predicted when JOLs are cued by only the stimulus, although a stimulus-response cue is more similar to the recognition test itself (Dunlosky & Nelson, 1997). Thus, the transfer-appropriate monitoring hypothesis appears to hold true only under certain conditions which are not yet fully determined.
The stochastic drift model.

Drawing upon the assumptions of the MDM hypothesis, the stochastic drift model (Sikström & Jönsson, 2005) explains the differences in accuracy as the result of a drift in memory strength from the time of judgment to the time of testing. In contrast to the MDM hypothesis, however, the stochastic drift model does not differentiate between short-term and long-term memory systems, but instead explains the decay of information in terms of fast or slow memory traces. The memory strength of any information is governed by different exponential decay processes that cause the information to decay at different speeds, and because immediate JOLs exhibit greater drift from time of judgment to time of testing (compared to delayed JOLs), their accuracy will inherently be lower.

The self-fulfilling prophecy (SFP) theory.

Another potentially viable explanation for the delayed JOL effect is the self-fulfilling prophecy theory (SFP) as presented by Spellman and Bjork (1992). According to this theory, the delayed JOL effect is present when information in long-term memory is successfully accessed and, consequently, tested. This process in turn effectuates the JOL outcome; an item, which can be retrieved from long-term memory, is given a high JOL, whereas an item, which cannot be retrieved, is given a low JOL.

Furthermore, the SFP theory explains the correlation between delayed JOLs and recall performance as moderated by a memory effect, which is brought about by the attempted retrieval of items. This side effect, which is also known as the testing effect (e.g., Roediger & Karpicke, 2006a), occurs simply because the act of retrieval itself constitutes a learning opportunity, and for this reason, the SFP theory suggests an equivalence between testing and additional study in terms of memory gains. Indeed, it has been shown that testing memory may be even more beneficial to memory than equal amounts of additional study (Carpenter & DeLosh, 2006; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006a), although depending on conditions such as retention interval (Roediger & Karpicke, 2006a). In sum, the key assumption is that an item which is successfully retrieved will not only be given a high JOL, but also have an increased likelihood of being successfully retrieved in the future, whereas low JOL items (that have not been retrieved) do not receive this enhancement in memory. This in turn is expected to lead to a boost to the JOL-recall correlation.

The testing effect

Another phenomenon, which is closely related to the main concepts of the SFP theory, is the testing effect. Applied mainly to verbal learning, the testing effect and its occurrence are well documented (e.g., Bartlett & Tulving, 1974; Donaldson, 1971; Izawa, 1967; McDaniel & Masson, 1985; Wheeler & Roediger, 1992; Whitten & Bjork, 1977), but outside the domain of cognitive psychology, however, this effect is not well known. Thus, it appears that most people are unaware of the potential memory boost stemming from the testing effect, and therefore do not guide their learning efforts accordingly, that is towards continuously testing themselves during study. If anything, the preferred means of learning appears to be repeated study, and often at the last minute (Kornell & Bjork, 2007).

The basic premise of the testing effect is that retrieval, when successful, has a reinforcing impact on memory. Testing whether or not we know something causes an attempted retrieval
of the appropriate information, and if this information is successfully retrieved, it will also, under some circumstances, increase likelihood of being successfully retrieved in the future, and sometimes more than an equal amount of study (see Roediger & Karpicke, 2006a). Although the underlying mechanisms of the testing effect are not yet fully understood, there are very promising theories that account for the memory boost (see Roediger & Karpicke, 2006a, for a review). One viable candidate is the retrieval hypothesis (Dempster, 1996), where the key predictor of recall performance is the encoding or processing depth. According to Bjork (1975), information that requires greater retrieval effort or greater depth of processing will consequently be encoded or re-encoded at a deeper level, which in turn strengthens memory. Therefore, information, which is repeatedly accessed and then restored in memory, will become increasingly available and resilient to decay.

Another explanation for the testing effect is the concept of transfer-appropriate processing (see Craik & Tulving, 1975), which, as mentioned earlier, is also a possible candidate for explaining the delayed JOL effect (but then referred to as the transfer-appropriate monitoring hypothesis). Transfer-appropriate processing points to the correspondence between learning or encoding processes and the tasks required in the memory task itself. In other words, if a memory task consists of visual stimuli, but recall performance is measured by their verbal representations (e.g. a picture of a boat vs. the word boat), the testing effect will only benefit memory to the extent these processes match (i.e. transfer onto) one another. However, according to findings by Glover (1989) and Carpenter and DeLosh (2006), the testing effect cannot be fully accounted for by the matching processes elicited by intervening and final tests.

In addition to transfer-appropriate processing, it seems that desirable difficulties improves recall performance, as they inherently force participants to process information in a more effortful manner (Bjork, 1994). For instance, having to memorize something, which at face value is incomprehensible, causes participants to make greater efforts towards comprehension, which in turn enhances the testing effect for that particular item. Thus, the difficulty posed by the task itself is, according to Bjork (1975; 1994; 1999), desirable for the purposes of learning. However, Roediger and Karpicke (2006b) have found that, in the case of studying versus testing memory, testing memory inherently poses a challenge which satisfies the need for desirable difficulties. In a sense, this finding supersedes Bjork’s (1994) account, simply because the desirable difficulties are posed naturally by the testing. In other words, testing memory renders the inclusion of desirable difficulties in the learning environment obsolete in terms of memory benefits, because its need has already been satisfied.

More recent findings by Karpicke and Blunt (2011) have shown that retrieval practice (testing memory) is more beneficial to learning than is elaborated study (concept mapping), while the accuracy of metacognitive predictions do not improve in a similar fashion – instead JOL accuracy benefited the most from repeated study. It should be pointed out that the study conducted by Karpicke and Blunt (2011) used global JOL measures and not item-by-item JOLs, and for this reason, there is no possibility of producing a testing effect in terms of JOL outcome. Notwithstanding this fact, their finding points to a distinction between what underlying processes in the delayed JOL effect and the testing effect influence either recall performance, relative JOL accuracy or both, all the while the SFP theory does not take this disparity into account. That is, the SFP theory predicts (i) increased memory performance
with repeated testing, and (ii) in turn, a higher relative JOL accuracy (gamma) as a result of this memory benefit.

The question, then, is whether or not making delayed JOLs leads to a memory boost equivalent to that of explicit memory testing. If it does, this would support the SFP theory by demonstrating the presence of a testing effect. In the same fashion, making delayed JOLs repeatedly would be expected to result in a larger memory boost than would making delayed JOLs only one time, since delayed JOLs (which generate a retrieval attempt and thus covertly test participants’ memory) serve as additional learning opportunities according to the SFP theory (Spellman & Bjork, 1992). Similarly, the SFP theory also predicts that JOL accuracy (i.e., the correspondence between JOL and actual recall) should increase with repeated JOLs.

Another important question is that of retention interval. One might expect JOLs that reflect a short retention interval to differ from JOLs that reflect a longer retention interval. That is to say, the answer to the question “How much will you remember in five minutes?” is probably not the same as the answer to the question “How much will you remember in one week?”. However, according to Koriat et al. (2004), participants tend to largely ignore the prospective aspect of a JOL instruction. That is, participants appear to make a JOL that does not take the specified retention interval into account.

By introducing different retention intervals, we will be able to compare delayed JOLs elicited by their respective durations. If participants do not adjust their JOLs according to the specified retention interval, this will support the findings of Koriat et al. (2004). In addition, we might also observe when the aforementioned memory boosts are present, and when they are not. When making repeated JOLs, we should expect a more sizeable memory boost for a longer retention interval, compared to a short retention interval, as more forgetting will have taken place by that time. For a shorter retention interval, much of the forgetting that needs to take place for the memory boost to reveal itself will not yet have occurred (see Roediger & Karpicke, 2006b).

Furthermore, if making delayed JOLs does not always produce a testing effect, collected data might indicate what boundary conditions need to be satisfied for this memory boost to occur. As previously mentioned, only items that have been encoded correctly will benefit from repeated testing, and this initial encoding can thus be considered a boundary condition. Therefore, it is not unlikely that delayed JOLs will, potentially, exhibit the very same kind of boundary conditions with respect to memory performance.

Testing the assumptions that can be deduced from the SFP theory will need manipulation of two variables above all, namely repetition and retention interval. If the act of making a delayed JOL is truly equivalent to explicit memory testing, then the act of repeated testing should increase recall performance, whether it be by making delayed JOLs or by actually testing memory – the SFP theory regards them as comparable in terms of memory benefits. Therefore, a design that contains either single or repeated “testing” sessions will positively test the validity of this assumption. Since initial encoding is needed for future analyses, a preferable experimental design would contain initial study (S) sessions, followed by a single testing (T) session, followed by either a single or three consecutive delayed JOL (J) sessions before final recall. This makes for either two (TJ) or four (TJJJ) sessions in which the testing effect may manifest itself, and according to the SFP theory, four sessions should lead to a greater testing effect than two (cf. Roediger & Karpicke, 2006b; Tulving, 1967).
In addition to the manipulation of single or repeated JOLs, the experiment will also apply two different retention intervals in order to assess its role as a possible boundary condition. As previously mentioned, we should expect a larger testing effect in cases where the retention interval is longer, simply because participants will have forgotten more by that time, and thus benefit more from the effects of testing memory.

By examining more closely the effects of single versus repeated JOLs and retention interval, this experiment may contribute to a better understanding of the SFP theory and its underpinnings. In a review by Sikström and Jönsson (2005), comparisons were made in 19 experiments between immediate and delayed JOLs, and no systematic support of the SFP theory was found. While there was a tendency toward better recall after delayed JOLs compared to immediate, a sign test showed that the advantage was not significant. A more recent review by Rhodes and Tauber (2011) compared effect sizes from 98 experiments, and found a modest but significant advantage for delayed JOLs compared to immediate JOLs with respect to memory performance.

Interestingly, while the review yielded very modest effect sizes for the delayed JOL effect in terms of memory performance, Rhodes and Tauber (2011) discovered very robust benefits to relative JOL accuracy. This finding goes some way towards delineating the delayed JOL effect, as it appears to affect mainly memory monitoring (JOL accuracy) rather than actual memory performance.

In sum, the present experiment aimed to investigate the validity of the SFP theory, using retention interval and number of JOL sessions as independent variables and recall performance and relative JOL accuracy as main dependent variables. More specifically, the retention interval serves as a possible boundary condition for the effects associated with single versus JOL sessions to appear.

Method

Participants

The participants were 79 undergraduate students within the Humanities and Social Sciences (53 women; age: M = 26.49; SD = 7.96; range = 18-56 years) at Stockholm University. In return for their participation, they received either movie vouchers or course credit. Participants were recruited using billboard notifications and online recruitment forms.

Materials

The to-be-remembered items consisted of a total of 40 Swahili-Swedish word pairs adapted and translated from the Swahili-English word standards by Nelson and Dunlosky (1994). The experiment was conducted within an E-Prime 2.0 Professional (Psychology Software Tools, Inc.) software environment running on PC computers, and responses were collected entirely via keyboard input.
Filler tasks and retention intervals

A 45 second distractor task was produced by displaying random selections from a large list of mathematical equations (one to two digit multiplications), which were either correctly or incorrectly displayed, where the task consisted in verifying their solutions (accepting or rejecting them as correct). For the longer 5-minute retention interval, a movie trailer (Wall-E featurette) was presented on the computer screen; the movie trailer was shown twice and the participants were asked if there was any difference between the materials that had been shown (there was none). For the longest 7-day retention interval, an appointment was made at the end of the initial learning session for the participants to finish the experiment one week later. See Figure 1 for an outline of the sequence of sessions and filler tasks.

In addition to the explicit memory performance data, scores of the Mental Effort Tolerance Questionnaire (METQ; Dornic, Ekehammar & Laaksonen, 1990) were also collected at the end of the experiment (after the final cued recall test). However, this measure was not included here and is therefore not further discussed. Because the questionnaire was administered at the very end of the experiment, there is no possibility that other measures may have been affected by completing the METQ questionnaire.

Design

Participants were divided into groups according to a 2 x 2 between subject design, using two different JOL conditions (single or repeated JOLs) and two different retention intervals (either 5 minutes or one week). The two JOL conditions contained different sequences of three learning procedures, namely Study, Testing or JOL sessions (referred to as S, T or J for the sake of brevity). Between any two sessions, a 45 sec distractor task (verifying solutions) was administered. Randomization in terms of group assignment was maintained insofar as participants were assigned to different groups, in order of appearance, with respect to JOL condition. Thus, no two consecutive participants would perform in the same JOL condition. Retention interval, however, was not randomized in the same fashion. Instead, all of the 5-minute retention intervals were performed during one phase of data collection, and the 1-week retention intervals during another.

Procedure

The participants were first welcomed and seated near the computer. At this point, an informed consent / confidentiality form was given to the participants, along with a brief description of the experiment. The experimenter entered age and other background variables, after which the experiment was commenced and the participants began to follow the instructions on the screen in front of them. Participants were either alone with the experimenter during the experiment, or with another participant running the experiment simultaneously. Participants were separated by a wooden screen and instructed not to interact in any way during the experiment.

After the initial instructions, all participants would perform three consecutive study sessions (see Figure 1). The first two study sessions consisted of self-paced exposures (referred to as $S_{sp}$ as in self-paced study) of the first half of the word list, and subjects could choose freely within a 20 sec interval when to move on to the next word pair. The third and last study
session consisted of 6 sec exposure of each of the 40 Swahili-Swedish word pairs (i.e. both halves of the word list) with a randomized presentation order for each participant. The division into two separate word lists and the self-paced approach ensured that participants would be very likely to correctly encode and recall some word pairs, but not others. This was done mainly as an attempt to minimize the risk of encountering ceiling and/or floor effects, as the second half of the word pairs will naturally be more difficult to remember because it was presented only once, compared to three times for the first half. In other words, we should expect high recall performance for the items that were exposed three times, and low recall performance for items that were exposed once. This approach also increases the likelihood of observing a testing effect as well as differences in JOL accuracy in terms of gamma correlations.

A testing session consisted of a cued recall test of the word pairs in randomized order, and with a 15 sec response time limit. A JOL session was practically identical to a testing session, but instead of prompting the target (Swedish) word, participants were asked "How certain are you that you will recall the Swedish word in one week (5 minutes), given that you are shown the Swahili word?" after which they responded on a percental scale from 0 to 100% with increments of 20%.

The two experimental conditions were composed of varying combinations of sessions outlined in Figure 1. Please note that the initial sequence of sessions is identical for all groups in the experiment ($S_{sp}$, $S_{cp}$, S, T), as it establishes a certain level of initial encoding or memorization which is thus equal for all participants.

![Figure 1. The sequence of sessions for the different JOL conditions and retention intervals. Brackets indicate points where the same procedure may be repeated. For instance, “x1 or x3” denotes sessions that differ between groups, and are completed either once or three times. Likewise, the arrow indicates a retention interval of 5 minutes or 1 week, again depending on group. As for the rest, the sequence is identical for all groups.](image-url)
The 5-minute retention interval was maintained using the Wall-E featurette as a nonsensical distinction task; the movie trailer was shown twice and the participants were asked if there was any difference between the materials that had been shown (there was none). For the longer, one week retention interval, an appointment was simply made for the participants to finish the experiment seven days later. Upon completion of the experiment, participants were instructed to fill out the METQ questionnaire, again using the keyboard as input device.

**Ethical considerations**

Given the non-intrusive nature of the experimental procedure, there were no ethical considerations to be made, except, of course, requirements of informed consent and confidentiality. To that end, the written consent form satisfied those requirements.

**Results**

For the analyses described below, data was entered into a two-way analysis of variance (ANOVA) with the between subjects variables JOL condition and retention interval as independent variables, and the between subject variable final recall (as well as forgetting) as dependent variable. An alpha level of .05 was used, and missing values were handled by using case-wise deletion. Effect-sizes are denoted by partial eta squared \( \eta^2 \) for the ANOVAs.

**Measures of JOL accuracy**

The SFP theory predicts increased measures of JOL accuracy with repeated testing. To this end, we compared relative measures of JOL accuracy with respect to JOL order (see Table 1) for the repeated JOL condition. For the sake of clarity, absolute measures of JOL accuracy were also included in the comparison, although the SFP theory makes no predictions regarding absolute JOL accuracy.

Gamma correlations from the first JOL session in both JOL conditions were entered into a two-way ANOVA as the dependent variable and retention interval as independent variable. There was a main effect of JOL condition, (single JOL: \( M = .90, SD = .02 \); repeated JOL: \( M = .83, SD = .02 \)), \( F(1, 78) = 4.98, p < .05, \eta^2 = .06 \), but no main effect of retention interval. There was also an interaction between JOL condition and retention interval (single JOL, 5 minutes: \( M = .95, SD = .03 \); single JOL, 1 week: \( M = .86, SD = .03 \); repeated JOL, 5 minutes: \( M = .81, SD = .03 \); repeated JOL, 1 week: \( M = .85, SD = .03 \)), \( F(1, 78) = 3.93, p = .05, \eta^2 = .05 \). This finding suggests, rather paradoxically, that relative JOL accuracy is significantly higher in the single JOL condition compared to the repeated JOL condition, but only when the retention interval is 5 minutes and not 1 week.

**Effects of JOL order.**

For the repeated JOL condition, a total of three 1 x 3 repeated measures ANOVAs were also conducted, using the within subject variables JOL accuracy, mean item-by-item JOL and over/underconfidence index as dependent variables. See Table 1 for a breakdown of the different dependent variables as a function of JOL order. Because sphericity could not be assumed, the degrees of freedom have been altered according to the Greenhouse-Geisser correction.
Table 1. The dependent variables, as a function of JOL order in the repeated JOL condition, displayed in means and (standard deviations).

<table>
<thead>
<tr>
<th>Measure</th>
<th>JOL session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>JOL</td>
<td></td>
</tr>
<tr>
<td>5 minutes</td>
<td>.38 (.14)</td>
</tr>
<tr>
<td>1 week</td>
<td>.29 (.18)</td>
</tr>
<tr>
<td>Gamma</td>
<td></td>
</tr>
<tr>
<td>5 minutes</td>
<td>.81 (.22)</td>
</tr>
<tr>
<td>1 week</td>
<td>.85 (.14)</td>
</tr>
<tr>
<td>O/U index</td>
<td></td>
</tr>
<tr>
<td>5 minutes</td>
<td>-.02 (.15)</td>
</tr>
<tr>
<td>1 week</td>
<td>.04 (.14)</td>
</tr>
</tbody>
</table>

First, we calculated three Goodman Kruskal gamma correlations for each participant in the repeated JOL condition, between the JOLs and final recall, one for each JOL session. These gamma correlations served as a measure of relative JOL accuracy, and were then entered into a 1x3 repeated measures ANOVA as the dependent variable, using retention interval as independent variable. There was no main effect of retention interval ($F < 1$). There was, however, a main effect of JOL order, ($1^{st}$ JOL: $M = .83$, $SD = .18$; $2^{nd}$ JOL: $M = .86$, $SD = .17$; $3^{rd}$ JOL: $M = .88$, $SD = .18$), $F(1.64, 60.69) = 7.49, p < .01$, $\eta^2_p = .17$, suggesting that making repeated JOLs increases their accuracy. There was no JOL order x retention interval interaction ($F < 1$).

Second, we calculated three mean item-by-item JOL ratings for each participant in the repeated JOL condition, one for each JOL session, which were then entered into the ANOVA as the dependent variable. There were no significant main effects of either JOL order or retention interval ($F < 1$), and there was no significant JOL order x retention interval interaction ($F < 1$), which simply shows that participants did not increase the average magnitude of their JOLs with each JOL session, nor did the retention interval affect the average magnitude of the JOLs.

Third and finally, we calculated three mean indices of over/underconfidence by subtracting the mean recall performance from the mean item-by-item JOL ratings for each participant in the repeated JOL condition, one for each JOL session. These indices served as a measure of absolute JOL accuracy, and were then entered into the ANOVA as the dependent variable. There were no main effects of either JOL order or retention interval ($F < 1$), and there was no JOL order x retention interval interaction ($F < 1$), indicating that participants’ calibration, in terms of JOL accuracy, did not change over the course of three JOL sessions, nor did the notion that the JOL reflected upon five minutes or one week’s retention interval. This finding is not in accordance with the underconfidence-with-practice effect (Koriat, Sheffer & Ma’ayan, 2002) which expects a lower over/underconfidence index for repeated JOLs. In other words, participants appear to have made the same JOLs, even repeatedly, regardless of retention interval, and these JOLs most likely reflect what they recall at the very moment of judgment. For that reason, we would expect JOLs to better map onto the five minute retention interval than the one week interval, simply because five minutes is temporally more similar to “right now” than is one week.
Measures of memory performance

Besides relative JOL accuracy, the SFP theory predicts that memory performance should be higher in the repeated JOL group than the single JOL group. Nevertheless, a two-way ANOVA showed no main effect of JOL condition \( (F = 1) \). There was, however, a main effect of retention interval, \((5 \text{ minutes}: M = .42, SD = .03; 1 \text{ week}: M = .27, SD = .03), F(1,78) = 11.97, p < .001, \eta^2 = .14\), pointing to the rather obvious fact that participants tend to remember more after five minutes than they do after a week has passed. There was no significant JOL condition x retention interval interaction.

Forgetting was calculated by subtracting the proportion of correctly recalled items in the final recall test from the proportion of correctly recalled items in the initial recall test. A two-way ANOVA, using forgetting as dependent variable, showed a main effect of retention interval, \((5 \text{ minutes}: M = .02, SD = .07; 1 \text{ week}: M = .16, SD = .11), F(1,78) = 43.16, p < .001, \eta^2 = .37\), but no main effect of JOL condition \((F < 1)\) and no JOL condition x retention interval interaction \((F < 1)\). This, again, underscores that not only did JOL condition not influence recall performance, it did not mitigate forgetting either.

Discussion

The main purpose of this study was to investigate the effects of single versus repeated delayed JOLs and retention interval with respect to the predictions of the SFP theory. More precisely, whether multiple spaced JOLs lead to better memory performance at a final test than a single JOL. To that end, the current data provided no conclusive evidence of a testing effect as posited by the SFP theory. Making repeated delayed JOLs did not differ significantly from making a single delayed JOL with respect to final recall performance, which contradicts the first prediction of the SFP theory. Relative JOL accuracy did, however, increase with repeated JOL sessions, which is somewhat akin to the second prediction of the SFP theory. Yet, these two predictions are contingent on one another in such a way that the increase in relative JOL accuracy is a direct consequence of the increase in memory performance. If there is no increase in memory performance to account for the increase in relative JOL accuracy, then the second prediction of the SFP theory was not corroborated either.

In addition to JOL condition, this study also assessed the role of retention interval as a possible boundary condition within the SFP theory. As noted previously, a common finding in the testing effect literature is that testing benefits long-term memory (e.g., when the final test is given after a week), but is not necessarily evident in the short term (e.g., after 5 minutes; Roediger & Karpicke, 2006b). The typical finding, then, is an interaction between retention interval and amount of testing. Results show that while participants, obviously, tend to remember more after 5 minutes has passed than after 1 week, there where no interactions between JOL condition and retention interval with respect to memory performance. If retention interval acted as a boundary condition for the testing effect, we should have seen such an interaction.

In terms of relative JOL accuracy, however, there was an interesting interaction between JOL condition and retention interval, indicating that relative accuracy was, surprisingly, higher in the single JOL condition, but only at a retention interval of 5 minutes. The explanation for this finding may be, as stated earlier, that participants’ judgments of learning probably reflect their
recall at the very same moment, and thus are more similar to the notion of 5 minutes than 1 week of retention. Nonetheless, if the predictions of the SFP theory were to be supported, we should have expected an interaction where relative accuracy was higher for repeated JOLs and a retention interval of 1 week.

What, then, caused the increase in relative JOL accuracy, and by what mechanism? In the Rhodes and Tauber (2011) meta-analysis, a general finding is that delayed JOLs primarily affect memory monitoring and, to a much lesser extent, performance. However, in the work by the main comparison is between immediate and delayed JOLs in terms of memory performance and relative JOL accuracy, not between single versus repeated delayed JOLs.

Nevertheless, the finding in the current experiment – that relative JOL accuracy increased with repeated JOLs – corroborates the findings of Rhodes and Tauber (2011). That is to say that delayed JOLs predominantly influence memory monitoring, because they (unlike immediate JOLs) are conducted in a setting which provides diagnostic information regarding the current memory state. As noted earlier, this increase in relative JOL accuracy (monitoring) did not significantly affect the memory performance in any way. This positively contradicts the SFP theory, which proposes an increase in relative JOL accuracy that is contingent on a concurrent increase in memory performance.

The fact that gamma correlations increased with repeated JOLs may be the mere result of a practice effect, such that participants become increasingly good at making metacognitive judgments as they keep making them repeatedly. Another possible explanation is that participants, in fact, remembered more at the time of the third JOL, compared to the first, which makes the JOLs based on the third retrieval attempt more relevant and accurate than the preceding two.

At present time, there are no similar studies that have specifically compared the effects of single versus repeated JOLs on either memory performance or relative JOL accuracy. For instance, experiments by Townsend and Heit (2010) have shown that JOLs do predict subsequent recall, but in a design where not only judgment, but also testing, was repeated. Therefore, there is no way of determining whether repeated testing or judgment (or some combination of the two) caused the testing effect that in turn boosted memory performance.

So, while there is ample evidence of the delayed JOL effect, there is still no single theory that comprehensively accounts for it. The SFP theory simply does not accommodate the fact that memory monitoring (e.g., relative JOL accuracy) and memory performance (e.g., cued recall) appear to be governed by different and separate processes. For this reason, a likely scenario is that a future, more developed theory of the delayed JOL effect is likely to be a combination of the most promising theories that exist today (see Rhodes & Tauber, 2011).

The present experiment compared differences in memory performance and JOL accuracy for participants who made either single or repeated JOLs, at different retention intervals before final recall. In order to fully compare and isolate the effects associated with studying, testing and judging, a more comprehensive design is eventually needed. For instance, single and repeated JOLs need to be compared, not only to each other, but also to single and repeated testing and study sessions, again for different retention interval. Moreover, a control group with no intervening sessions – besides initial study, of course – is needed in order to assess the effects of single study/JOL/test sessions before final recall. In total, this would call for a
2x2x4 between subjects design, using repetition (single vs. repeated sessions), retention interval (long vs. short) and session type (control vs. study vs. test vs. JOL) as independent variables.

In conclusion, the SFP theory is not supported and neither is it refuted by the findings of the present experiment. Instead, the current data suggest separate domains within the delayed JOL effect in which the SFP theory appears valid or invalid. In terms of memory performance, the SFP theory has ample support. In terms of memory monitoring, however, the SFP theory is not in accord with the effects observed when making delayed JOLs.

References


