What moderates the accuracy of ease of learning judgments?

Andreas Jemstedt¹ · Veit Kubik^{1,2} · Fredrik U. Jönsson¹

Received: 27 July 2016 / Accepted: 3 May 2017 / Published online: 30 May 2017

© The Author(s) 2017. This article is an open access publication

Abstract When people begin to study new material, they may first judge how difficult it will be to learn. Surprisingly, these *ease of learning* (EOL) judgments have received little attention by metacognitive researchers so far. The aim of this study was to systematically investigate how well EOL judgments can predict actual learning, and what factors may moderate their relative accuracy. In three experiments, undergraduate psychology students made EOL judgments on, then studied, and were tested on, lists of word-pairs (e.g., sun – warm). In Experiment 1, the Goodman-Kruskal gamma (G) correlations showed that EOL judgments were accurate ($G = .74$) when items varied enough in difficulty to allow for proper discrimination between them, but were less accurate ($G = .21$) when variation was smaller. Furthermore, in Experiment 1 and 3, we showed that the relative accuracy was reliably higher when the EOL judgments were correlated with a binary criterion (i.e., if an item was recalled or not on a test), compared with a trials-to-learn criterion (i.e., how many study and test trials were needed to recall an item). In addition, Experiments 2 and 3 indicate other factors to be non-influential for EOL accuracy, such as the task used to measure the EOL judgments, and whether items were judged sequentially (i.e., one item at a time in isolation from the other items) or simultaneously (i.e., each item was judged while having access to all other items). To conclude, EOL judgments can be highly accurate ($G = .74$) and may thus be of strategic importance for learning. Further avenues for research are discussed.

Keywords Ease of learning judgments · Monitoring · Metacognition · Cue utilization · Item difficulty

Learning is a central part of everyday life, for example studying a history book for an upcoming exam, learning a new language (e.g., glossary lists), practicing to use a machine at work, or how to operate a new smartphone. One of the first things you may do when you are

✉ Fredrik U. Jönsson
fredrik.jonsson@psychology.su.se

¹ Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden

² Department of Psychology, Humboldt-Universität zu Berlin, Berlin School of Mind and Brain, 12489 Berlin, Germany

about to learn something new is to survey the material to get a feeling for how difficult the task will be. This initial assessment, or metacognitive monitoring, of how difficult a material and its parts will be to learn is referred to as *ease of learning (EOL) judgments*. These judgments are important to investigate, because presumably they guide our study behaviour and our learning (Nelson & Narens, 1990). However, surprisingly few studies have investigated the accuracy of these potentially important pre-study judgments. For example, Dunlosky and Metcalfe (2009, p. 91) noted that “sustained devotion to understanding EOL judgments has been rather minimal and little is known about them.” This set the aim for the present study. We investigated how well EOL judgments predict actual learnability of word pairs, and what factors moderate their accuracy.

EOL judgments are made early in the learning process, before active learning has begun. According to Nelson and Narens (1994, p. 16) they “occur *in advance of acquisition*, are largely inferential, and pertain to items that have not yet been learned. These judgments are predictions about what will be easy/difficult to learn...”. Different labels have been used in the literature for people’s metacognitive monitoring of their learning and memory, depending on whether the monitoring occurs before or after active learning has started, and depending on the task given. As opposed to the less studied EOL judgments, much research has focused, for example, on *judgments of learning (JOL)* (Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011; Sikström & Jönsson, 2005) or *feeling-of-knowing (FOK) judgments* (Hart, 1965; Schwartz, Boduroglu, & Tekcan, 2016). JOLs refer to judgments about future memorability made during or after learning, and FOK judgments are judgments about later retrieval or recognition of presently unsuccessful retrieval attempts.

In the present study, we rephrase the Nelson and Narens (1994) definition of EOL judgments, as it is possibly too strict. For example, that EOLs should *pertain to items that have not yet been learned* is problematic if very easy paired-associates like *bird-song* or *bat-man* is shown - associations that may already be so well learned that no further study is needed to above chance correctly produce the second word if shown the first. Further, even if completely novel items are used (e.g., meaningless symbols), by experimentally eliciting explicit EOL ratings some incidental (or uninstructed but intentional) learning may take place before the judgments are made. We therefore use a slightly more relaxed version of the Nelson and Narens (1994) definition. EOLs are defined in terms of timing and task, as judgments made very early in the learning process, in advance of instructed study (the timing), about what will be easy or difficult to learn (the task).

One common distinction between different measures of EOL accuracy is between relative accuracy (or resolution) and absolute accuracy (or calibration). Absolute accuracy is a directional measure of the difference between confidence and accuracy, but is not the focus of the present study. Relative accuracy refers to the degree that an individual accurately predicts performance on one item in relation to another item. Thereby, it is a measure of the individual’s ability to discriminate between the items within a list, and is typically calculated with the non-parametric Goodman-Kruskal Gamma (G) correlation (Nelson, 1984). In this study, we focus on relative accuracy as calculated with G . More precisely, we correlate the subjectively perceived learnability of the items with the actual learning outcome. Although G has received some criticism lately (e.g., Masson & Rotello, 2009), it is the most commonly used measure in the field of metacognition.

The available research to date suggests that EOL judgments poorly or at best moderately predict actual learnability of the to-be-learned material (Leonesio & Nelson, 1990; Mazzoni, Cornoldi, Tomat, & Vecchi, 1997; McCarley & Gosney, 2005; Nelson & Leonesio, 1988; Son & Metcalfe, 2000; but see Britton, Van Dusen, Gülgöz, Glynn, & Sharp, 1991; Jönsson &

Kerimi, 2011; Lippman & Kintz, 1968). Indeed, Leonesio and Nelson (1990), who directly compared the monitoring accuracy of EOL judgments, FOK judgments, and judgments of knowing (akin to JOLs), found EOL judgments to be poorly, and also the least, related to the performance on a final recall test. However, due to the paucity of research, meaning that it is yet to be systematically investigated, it may be premature to make conclusions about EOL accuracy. For this reason, in the present study we investigated EOL accuracy in the context of paired associates learning, along with some variables that could potentially moderate the relationship between the EOL judgments and actual learnability. These are outlined below.

How does the within-list variability of item difficulty affect EOL accuracy?

Based on the assumption that people's metacognitive judgments are inferential in nature, Koriat (1997) put forward the *cue-utilization approach*. He suggested that people use different sources of available information, so called cues or heuristics, to infer about, for example, how difficult something will be to learn or how well it has been learned. The accuracy of metacognitive judgments should thus depend on how predictive the available information is of what is being judged and how well the information is utilized. Examples of such information are word length (Jönsson & Lindström, 2010), how semantically related words in a word pair are (Dunlosky & Matvey, 2001), if the material is read or listened to (Peynircioğlu, Brandler, Hohman, & Knutson, 2014), the feeling of how easily the material is to process (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989) or to retrieve from memory (Benjamin, Bjork, & Schwartz, 1998).

Koriat (1997) argued that EOL judgments are mainly based on characteristics of the material, such as semantic relatedness between words, whereas subsequent monitoring later in the learning process (e.g., JOLs) increasingly rely on idiosyncratic cues indicating how well the material has been learned, for example how fast an item is recalled from memory. Because EOL judgments are made early in the learning process, idiosyncratic cues indicating how well the material has been learned are likely not yet available to the extent that they play a prominent role in allowing the learner to discriminate between items in a list.

However, this does not necessarily mean EOL accuracy must be poor. Because EOL accuracy should mainly be based on the validity and discriminability of salient cues in the material, EOL accuracy should vary from low to near perfect depending on the characteristics of these two factors. For example, Leonesio and Nelson (1990), who found EOL judgments to be both poorly and the least related to actual learnability, may have underestimated the participants' monitoring ability, as they only used unrelated word pairs as the to-be-learned material. This probably made it particularly difficult for the participants to discriminate between items. Thus, to make conclusions about EOL accuracy it is necessary to investigate EOL judgments in a condition that allows proper discrimination between items in a list, and compare with one that does not. For this reason, in Experiment 1, we manipulated how much item difficulty varied within two to-be-learned lists of word pairs. Both lists were on average equally difficult; however, items in one of the lists varied from high to low semantic relatedness (high-variability condition), whereas the variation was kept low in the other list (low-variability condition). We selected semantic relatedness as the cue to vary because EOL judgments would likely be very sensitive to it (Dunlosky & Matvey, 2001). The material in the low-variability condition corresponds to that used by Leonesio and Nelson (1990).

We expected a clear difference in EOL accuracy as a function of within-list variability of item difficulty. This is in line with the notion that EOLs are strongly guided by salient characteristics of the material and corresponds to previous findings in the literature that were found with regards to other types of metacognitive judgments (Nelson, Leonesio, Landwehr, & Narens, 1986; Schwartz & Metcalfe, 1994). In addition, we were particularly interested in the accuracy level in the high-variability condition. In this condition, participants should be able to better discriminate between items and therefore make very accurate judgments.

Does the presentation format affect EOL accuracy?

The standard experimental paradigm, in the metacognition literature, is to present items for study and/or judgment sequentially on a trial-by-trial basis (i.e., *sequential presentation*; e.g., Nelson & Leonesio, 1988). Each metacognitive judgment is made in isolation from all other items. In an everyday situation, learners may be more likely to have visual access to all of the item material simultaneously (i.e., *simultaneous presentation*), for example a page with a glossary list in a language textbook. This may make comparison between different items easier. Relatedly, Dunlosky and Thiede (2004) demonstrated that presentation format affect what learners decide to study. They instructed participants to have a low performance goal (i.e., to only learn a smaller subset of the items in a presented list). As a result, participants always picked the easier items if all items were shown simultaneously; however, if the items were shown sequentially they instead picked the most difficult items, a poorer strategy given the performance goal. Hence, in a study-regulation task, the presentation format does seem to matter, but whether it affects EOL monitoring accuracy has not previously been investigated.

As previously noted, the EOL accuracy measure under scrutiny in this study was relative accuracy, that is, the participants' ability to accurately discriminate between items that will be easy or difficult to learn. When comparing a sequentially presented item with other items in a list one has to keep previously presented items in mind. Due to limitations in working-memory capacity, EOL judgments for each item may only be made relative to the last presented items (i.e., items that are still in working memory), or to the item or items that happens to come to mind. This could make sequential judgments more sensitive to bias, because in both cases the item sample in working-memory may not be representative of the whole list of items. When making simultaneous judgments the individual not only has an immediate overview of the full sample of items (rather than having to wait until the full list has been presented) but can also directly compare each item with the other list items without being affected by working memory constraints. In Experiment 3 we explored this issue, that is whether the presentation format affects EOL accuracy.

Does the criterion measure used affect EOL accuracy?

Relative EOL accuracy is operationalized as the correlation between EOL judgments and the corresponding values on a criterion measure. The judgments are commonly rated on a scale ranging from least to most difficult, by answering the question "How difficult will it be to learn this item?" (e.g., Kelemen, Frost, & Weaver, 2000; Son & Metcalfe, 2000). The criterion measure usually consists of the dichotomous performance on a memory test (i.e., recalled vs. non-recalled; e.g., Mazzoni et al., 1997). Although a valid way of measuring accuracy,

correlations may be reduced because of a restriction in range in either the EOL judgment or criterion measure (e.g., Schwartz & Metcalfe, 1994). Thus, using a dichotomous criterion measure for learning could potentially lead to underestimation of participants' true EOL accuracy (i.e., their actual monitoring ability).

One criterion measure of learning that would provide a wider range of values is the *trials-to-learn* (TTL) criterion (e.g., Leonesio & Nelson, 1990). The TTL criterion measures on what test, out of multiple study and test cycles, an item was first recalled. Kelemen, Frost and Weaver (2000) argued that this “yields a more sensitive index of performance because the trials-to-learn criterion results in a range of values rather than the binomial outcome” (p.105). In Experiments 1 and 3, we investigated this issue by comparing EOL accuracy calculated on recall at the first test with that calculated with a TTL criterion. Given the above reasoning, we tentatively predicted that using the TTL criterion would lead to a higher EOL accuracy.

Does the type of EOL task affect EOL accuracy?

Finally, we investigated whether the type of EOL judgment task influence EOL accuracy (Experiments 1–3). To this end, we applied two different EOL tasks. In a first, participants judged how difficult each item would be to learn on a scale ranging from easy to difficult (*difficulty task*). In a second EOL task, participants use a trials-to-learning scale (*TTL task*) to judge how many study trials they would need to learn an item. When correlating these EOL judgments with the TTL criterion, the TTL task perfectly matches the criterion measure actually used. This is not to the same extent the case when participants perform the difficulty task, because it is not as clearly tied to either criterion measure. This notion, that monitoring accuracy should be higher, the higher the similarity between the monitoring and criterion measure context is, is referred to as transfer-appropriate monitoring (Dunlosky & Nelson, 1992). Thus, in Experiment 1 and 3, we also investigated whether the match between EOL task and criterion measure benefits the estimated EOL accuracy.

In sum, across three experiments we investigated EOL accuracy, and further whether EOL accuracy is moderated by (i) the extent that items within a list vary in terms of how difficult they are to learn, (ii) the presentation format, that is whether an EOL judgment for a single item is made in the presence or absence of all other items, (iii) the criterion measure used, that is how the actual learnability of the items is calculated, and (iv) the EOL task, that is whether the task was formulated as a difficulty rating or made on a trials-to-learning scale.

Experiment 1

In Experiment 1, we investigated to what extent EOL judgments can be made more accurate by increasing the within-list variability of a salient cue in the material. To this end, we varied the semantic relatedness between words in a list of word pairs (e.g., *Wool–Sock*). Semantic relatedness is both a valid predictor of actual difficulty and a highly salient cue (e.g., Carroll, Nelson, & Kirwan, 1997). We let participants judge and learn one list with high variability and one list with low variability that were on average equally difficult. We also investigated whether the EOL task (difficulty vs. TTL) and the criterion measure (binary criterion vs. TTL criterion) moderated the EOL accuracy, with particular focus on whether EOL accuracy would be better when the EOL task (TTL task) perfectly matches the criterion measure (TTL criterion), compared to the other conditions.

Method

Participants

Thirty-one Stockholm University psychology students (25 women) between 18 to 61 years ($M = 25.38$, $SD = 8.43$) were recruited via flyers at billboards or from classrooms. They all signed an informed consent form and participated voluntarily in exchange for course credit. A further two participants took part in the experiment but were excluded from the analyses due to noncompliance with the instructions.

Materials

Two lists of word pairs were constructed that differed in how much the items *varied* in semantic relatedness (cf., Swedish association norms, Shaps, Johansson, & Nilsson, 1976). A *low-variability list* included 40 similarly related Swedish–Swedish pairs and this to a low degree (e.g., *Baby–North*). The *high-variability list* also consisted of 40 word pairs. The list included 8 Swahili–Swedish word pairs (e.g., *Adha–Problem*), and 8 Swahili–Swahili word pairs (e.g., *Ndege–Fuvu*) having very low semantic relatedness. Further, the list included 8 Swedish–Swedish word pairs with low (e.g., *Foot–Puzzle*), 8 with moderate (e.g., *Hole–Ring*) and 8 with high semantic relatedness (e.g., *Sun–Warm*). Although the variability differed, we aimed for the lists to be equally difficult to learn on average. The experiment was conducted on personal computers running the e-prime 2.0 software (Psychological Software Tools, Pittsburgh, PA; Schneider, Eschman, & Zuccolotto, 2002).

Design and procedure

A 2 (list: high vs. low variability) $\times$ 2 (EOL task: difficulty vs. TTL) $\times$ 2 (criterion: binary vs. TTL) within-group design was used. The experiment consisted of two consecutive blocks with identical procedure. The list variable was manipulated over the two blocks, that is, in one block the procedure was completed with the high-variability list, and in the other block with the low-variability list. In each block, the participants first judged each item in the two EOL tasks (the difficulty task was always followed by the TTL task), after which three consecutive study and test phases were completed. The outcome of the first test in the series of three was used as the binary criterion measure (recalled vs. not recalled). As an item was first recalled either on the first, second or third test, or not at all, this forms a four-graded criterion measure of item learnability. This is the TTL criterion.

We counterbalanced the block order across participants, and the presentation order of all word pairs was randomized in all phases. Up to two participants were tested at the same time with a sound-isolating screen separating them. They first received oral instructions about the procedure and then in text on the computer.

EOL phase Participants were presented one word-pair at a time and for each pair they answered the question: “How difficult will it be to learn the association between the right and left word?” This question was first answered in the difficulty task by selecting either “0% (1)” (most difficult to learn), “20% (2),” “40% (3),” “60% (4),” “80% (5),” or “100% (6)” (easiest to learn). Directly after, the same question was answered in the TTL task by selecting either “1 = I will have learned it by Test 1,” “2 = I will have learned it by Test 2,” “3 = I will have

learned it by Test 3,” or “4 = I won’t have learned it by Test 3.” Judgments were self-paced, and made by pressing the respective key on the keyboard.

Study phase During all three study phases, word pairs were studied one at a time for five seconds each, followed by an interstimulus interval of half a second. After all word pairs were studied, participants made an aggregate judgment, “How many of the word pairs do you think you will remember on the test occurring in approximately 1 minute?”, by typing a number between 0 and 40. The aggregate judgments were not the focus of the present study, and is thus not further analyzed.

Test phase A cued-recall test was completed after each study phase, during which the participants were presented the cue word in each word pair (e.g., *Tabibu* – _____), one at a time, and were asked to type the missing target word on the keyboard. The experiment automatically proceeded to the next item after 15 s or when participants pressed Enter.

A distractor task was completed for 45 s between the EOL, study, and test phases. The participants were to judge whether arithmetic problems were correct or not (e.g., $65 \times 11 = 705$). If the participant failed to give an answer in 15 s, the experimental script proceeded to the next arithmetic problem.

Data processing and statistical analysis

All judgments in the difficulty task were inverted (i.e., a 1 was turned into a 6) before analysis, in order for the judgments in the difficulty task to increase in the same direction as those in the TTL task.

Results and discussion

EOL accuracy (G) was calculated for each participant and condition between EOL tasks and criterion measures. Five participants were excluded from the EOL accuracy analysis.¹

EOL accuracy was high when the to-be-learned material, via a valid and predictive cue, allowed for proper discrimination between the items, which stands in contrast to some studies that have shown EOL accuracy to be low (e.g., Leonesio & Nelson, 1990). EOL accuracy was reliably higher for the high ($M = .74$, $SD = .13$) than for the low variability list ($M = .21$, $SD = .24$), as shown by a main effect of list, $F(1, 30) = 110.82$, $p < .001$, mean difference = .53, 95% CI = [.43, .63], $d = 2.70$. The very large effect size demonstrates the sensitivity of EOL judgments to salient characteristics in the material and item discriminability.

The results also showed that, although the difference was small, EOL accuracy was reliably higher when EOL judgments were made in the TTL task ($M = .50$, $SD = .15$) than in the difficulty task ($M = .45$, $SD = .16$), $F(1, 30) = 4.56$, $p = .041$, mean difference = .06, 95% CI = [.002, .11], $d = .36$. Furthermore, the criterion measure minimally but reliably influenced

¹ One participant was excluded due to perfect recall on one list and near perfect on the other (39 out of 40 items), and four for likely having reverted the scale in the TTL task (i.e., consistently judging easy items requiring many study trials to be learned). The latter four participants were also excluded from the analysis of the magnitudes of EOL judgments.

EOL accuracy, in that gamma was reliably higher when EOL judgments were correlated with the binary criterion ($M = .484$, $SD = .13$) than to the TTL criterion ($M = .465$, $SD = .14$), $F(1, 30) = 9.60$, $p = .004$, mean difference = .019, 95% CI = [.01, .03]. The latter finding was unexpected as correlations in most cases benefit from scales that can take a larger range of values (TTL criterion) than if it is binary. However, the main effect of criterion measure on EOL accuracy was in terms of the effect size very small, $d = .14$.

We hypothesized that the EOL accuracy would be higher when the EOL task perfectly matched the criterion measure. The only condition with such a match was when EOL judgments from the TTL task were correlated with the TTL criterion. If the task match affects EOL accuracy, we should have found a reliable interaction between EOL task and criterion measure, but this was not the case ($p = .119$). Furthermore, there were no statistically reliable interactions between any of the three factors, with the exception of a reliable interaction between list and criterion, $F(1, 30) = 6.73$, $p = .015$, mean difference = .03, 95% CI = [.01, .06], $d = .68$. The difference in gamma in favor of the binary criterion over the TTL criterion was larger for the low-variability list ($M = .04$, $SD = .06$), as compared to the high-variability list ($M = .002$, $SD = .04$). See Fig. 1 for a visualization of the means.

Secondary analyses

Absolute accuracy By subtracting the mean TTL criterion performance from the mean EOL judgment in the TTL task, it was possible to calculate a measure of absolute accuracy for each participant. Participants generally judged that they would need more study trials to learn each word pair than they actually did, and this effect was reliably more pronounced for the low- ($M = -.85$, $SD = .69$) than for the high-variability list ($M = -.58$, $SD = .29$), $t(30) = 2.20$, $p = .036$, mean difference = .27, 95% CI [.02, .52], $d = 0.51$.

Recall The two to-be-learned lists were pre-standardized to differ in how much the items varied in terms of difficulty, but should not differ in terms of mean difficulty. To check that both lists were equally difficult, we analyzed recall performance at the three memory tests and in terms of a TTL score. Table 1 shows that the two lists did not differ at the first recall test ($t(35) = 0.84$, $p = .405$, mean difference = .89, 95% CI = [-1.25, 3.03], $d = .12$), but the low-

Fig. 1 EOL accuracy (G), as a function of list (high vs. low variability) and EOL task (difficulty vs. TTL), separately shown for the two criterion measures, binary (first panel) and TTL (second panel). Error bars represent the 95% confidence intervals

variability list was better learned at the second ($t(35) = 3.91, p < .001$, mean difference = 3.31, 95% CI = [1.59, 5.02], $d = .57$) and at the third test ($t(35) = 5.60, p < .001$, mean difference = 5.06, 95% CI = [2.38, 5.57], $d = .86$); a difference also captured by the TTL criterion. Word pairs in the low-variability list were learned faster than the word pairs in the high-variability list ($t(35) = 3.45, p = .001$, mean difference = .20, 95% CI = [.08, .32], $d = .48$). This likely explains the difference in absolute accuracy between the lists.

Magnitude of the EOL judgments As can be seen in Table 1, mean magnitude of the EOL judgments was similar for both lists, independent of EOL task (difficulty task: $t(31) = 1.20, p = .239$, mean difference = .15, 95% CI = [-.10, .39], $d = .22$; TTL task: $t(31) = .45, p = .657$, mean difference = .04, 95% CI = [-.14, .21], $d = .08$).

Experiment 2

In Experiment 1, EOL judgments were more accurate when formulated as TTL compared to difficulty ratings. However, in Experiment 1 a procedural detail makes interpretation ambivalent. On every trial, the participants always first judged each item in the difficulty task, and then directly judged the same item in the TTL task. Thus, the TTL task advantage could be the result of either the task itself (i.e., it is better to judge TTL than difficulty) or the task order. Perhaps judging the same item twice makes the second judgment more accurate? To investigate whether the superiority holds when the potential order effect is removed, we manipulated the EOL task between groups in Experiment 2.

Method

Participants

Forty-eight psychology undergraduates at Stockholm University (27 female), ranging in age between 19 and 47 ($M = 24.80, SD = 5.94$) were recruited during classes and participated voluntarily in exchange for course credit. All participants signed an informed consent form.

Materials

PsychoPy 1.81.03 (Peirce, 2007), a projector and a projector screen was used for stimulus presentation. The list of 40 word-pairs were the same as in the high variability condition in

Table 1 Means and standard deviations for EOL judgments during the EOL tasks and number of correctly recalled items (maximum 40) at Test 1, Test 2, and Test 3, as well as recall measured as the TTL criterion

List	Recall				EOL task	
	Test 1	Test 2	Test 3	TTL criterion	Difficulty _{Inverted}	TTL
High variability	22.83 (5.57)	29.03 (4.40)	31.92 (3.88)	1.88 (.32)	3.53 (.52)	2.45 (.36)
Low variability	23.72 (9.28)	32.33 (6.85)	35.89 (5.29)	1.68 (.50)	3.67 (.78)	2.49 (.53)

Experiment 1. Each participant received a stapled 8-page pamphlet consisting of an informed consent form, a page for making EOL judgments, three pages of arithmetic problems and three pages used to write down test answers.

Design and procedure

A 2 (EOL task: difficulty vs. TTL) between-groups design was used. The experiment began with the experimenter randomly shuffling the pamphlets and distributing them to the participants, and thereafter presenting the procedure orally and answering questions about the experiment. To also match the TTL task the participants were informed that they would go through three study-test cycles, but the experiment was stopped after only one such cycle.

EOL phase Each item was presented one at a time for 8 s. To help participants know where on the paper each judgment should be written, the sequence number was presented above each word pair on the screen. Also, every time a new word-pair was presented, the computer made a sound to alert the participants about the change of word-pair.

The question “How difficult will it be to learn the association between the right and left word?” was answered for each item on a scale ranging from 1 to 4. In half of the pamphlets, the scale had the ends “1 = very easy,” and “4 = very difficult,” and in the other half of the pamphlets the scale steps was as follows: “1 = I will have learned it by test 1,” “2 = I will have learned it by test 2,” “3 = I will have learned it by test 3,” or “4 = I won’t have learned it by test 3.” Underneath the question and scale, the number 1 to 40 was printed from top to bottom in order for the participants to mark what word pair they were judging.

Study phase After the EOL phase, the participants studied the items one at a time for five seconds each.

Distraction phase After the study phase the participants were informed to turn to the next page of the pamphlet, which consisted of 80 arithmetic problems (e.g., “ $18 \times 6 =$ ”). Participants were given 1 min to solve as many of the problems as possible.

Test phase Thereafter, participants turned the page and the cue word of each word pair was presented one at a time for 10 s by the computer on the projector screen. For each item, participants were asked to try to recall and write down the missing target word on the paper. The paper had 40 rows numbered from 1 to 40. Again, the computer made a sound each time a new word was presented.

Although the items were presented in random order in all phases, the order was the same for all participants in the same classroom. The experiment was completed with four groups.

Results and discussion

The advantage of the TTL task found in Experiment 1 was not replicated when the EOL task was manipulated between groups, that is mean G for the two EOL tasks did not reliably differ (difficulty task: $M = .82$, $SD = .15$; TTL task: $M = .76$, $SD = .15$), $t(46) = 1.23$, $p = .226$, mean difference = $.05$, 95% CI = $[-.03, .14]$, $d = .35$. This indicates that it is not the TTL task per se that improves EOL accuracy, but rather that making a judgment again (albeit differently) may be beneficial.

Secondary analyses

Recall There was no reliable difference in memory performance depending on what EOL task was used (difficulty: $M = 19.30$, $SD = 7.55$; TTL: $M = 20.28$, $SD = 5.56$), $t(46) = .51$, $p = .611$, mean difference = .98, 95% CI = $[-2.85, 4.80]$, $d = .15$).

Magnitude of the EOL judgments The mean judgment of the two EOL tasks did not differ reliably (difficulty: $M = 2.67$, $SD = .30$; TTL: $M = 2.61$, $SD = .35$), $t(46) = .73$, $p = .468$, mean difference = .07, 95% CI = $[-.12, .26]$, $d = .21$).

Experiment 3

The standard experimental paradigm (as in Experiment 1 and 2) is to present items on a trial-by-trial basis, that is, sequentially one item at a time. However, when students study at home, they likely have access to all of the material at once (i.e., simultaneous). We reasoned that in a learning situation where items are presented simultaneously, an EOL judgment for one item is likely made in relation to the other items in the list, whereas in a learning situation in which items are presented sequentially the item can only be compared to items available in working memory. Simultaneous presentation would possibly allow for making more accurate judgments, as participants always have visual access to all other items when making the judgment. To investigate if EOL accuracy is affected by differences in presentation format, one group of participants made sequential judgments, and another group simultaneous judgments. Experiment 3 also included the EOL task and criterion measure manipulations to follow up on the previous experiments.

Method

Participants

Sixty-four psychology students (41 females) at Stockholm University, ranging in age from 18 and 37 years ($M = 25.58$, $SD = 6.43$), participated for course credit. One additional student participated in the experiment but was excluded for non-compliance with the instructions.

Materials

All participants learned two equally difficult lists of word-pairs consisting of 40 pairs each. The material was similar, but not identical, to that used in Experiments 1 and 2. The lists consisted of word pairs divided into four categories of semantic relatedness. Each list consisted of 10 Swedish–Swedish word-pairs with low, 10 with moderate, and 10 with high semantic relatedness, and 10 Swahili–Swedish word-pairs. The experiment was conducted on the same computers as in Experiment 1, running PsychoPy 1.81.03 (Peirce, 2007).

Design and procedure

A 2 (presentation format: simultaneous vs. sequential) $\times$ 2 (EOL task: difficulty vs. TTL) $\times$ 2 (criterion: binary vs. TTL) mixed design was used. Presentation format and criterion were manipulated within groups, and EOL task was manipulated between groups.

All participants were given instructions about the procedure, first orally and then as text on the computer. Similar to Experiment 1, the procedure was divided into two blocks, in which the EOL task was completed and then followed by three consecutive study and test phases.

The EOL phase Participants began each block in the procedure by either judging each item one at a time (sequential), or all at the same time (simultaneous). In the simultaneous condition, word pairs were randomly presented in two columns, with a scale ranging from 1 to 4 to the right of each item. This allowed the participants to judge the word pairs in any order they wanted to. Once all the items had been judged, participants could continue to the next phase of the experiment by pressing the spacebar. The question “How difficult will it be to learn the association between the right and left word?,” was always visible over the two columns of word pairs. In the sequential condition, judgments were self-paced and items were presented one at a time in random order. Because EOL task was manipulated between groups, participants in the difficulty-task condition were presented with the scale ends “1 = very easy,” and “4 = very difficult.” Participants in the TTL task condition were instead presented with the scale ends “1 = I will have learned it by test 1,” “2 = I will have learned it by test 2,” “3 = I will have learned it by test 3,” or “4 = I won’t have learned it by test 3.”

Study phase Each item was presented sequentially for 5 s in random order with a half second long interstimulus interval.

Test phase The left word of each word pair was presented one at a time and the participant was instructed to write down the missing target word. The pairs were presented in random order for a maximum of 15 s or until the enter key was pressed.

Between each phase of the experiment, the participants completed the same distractor task as in Experiment 1, for 45 s.

Results and discussion

Six participants were excluded from the EOL accuracy analysis due to perfect or near perfect recall on the first test (i.e., more than 36, or 90%, of the 40 items were recalled). There was no main effect of presentation format, that is, EOL accuracy was the same independent of whether judgments were made with items presented sequentially ($M = .63$, $SD = .18$) or simultaneously ($M = .60$, $SD = .28$), $F(1, 56) = .48$, $p = .491$, mean difference = .03, 95% CI = $[-.05, .10]$, $d = .12$. Hence, the hypothesis that simultaneous presentation should lead to better EOL accuracy was thus not supported.

We also replicated the nil finding of Experiment 2 in that there was no difference in EOL accuracy between the difficulty ($M = .62$, $SD = .20$) and the TTL task ($M = .61$, $SD = .19$), $F(1, 56) = .03$, $p = .863$, mean difference = .01, 95% CI = $[-.09, .11]$, $d = .05$. However, with respect to the criterion measure, and in line with Experiment 1, the estimated EOL accuracy was slightly and reliably higher when the EOL judgments were correlated with the binary

($M = .64$, $SD = .20$), as compared to the TTL criterion ($M = .59$, $SD = .19$), $F(1, 56) = 38.51$, $p < .001$, mean difference = .05, 95% CI = [.04, .07], $d = .28$. This is further discussed in the General Discussion. None of the factors reliably interacted ($ps > .05$; see Fig. 2).

Secondary analyses

Absolute accuracy Absolute accuracy of the EOL judgments could only be calculated for the group that made judgments in the TTL task. Absolute accuracy did not differ reliably between judgments made sequentially ($M = -.50$, $SD = .48$) compared to those made simultaneously ($M = -.63$, $SD = .44$), $t(31) = 1.30$, $p = .203$, mean difference = .12, 95% CI = [-.07, .31], $d = .26$.

Recall Presentation format did not have any important effect on memory performance. As shown in Table 2, Test 1 recall did not reliably differ ($t(64) = 1.72$, $p = .091$, mean difference = 1.15, 95% CI = [-.19, 2.50], $d = .17$), nor did Test 2 recall ($t(64) = .73$, $p = .471$, mean difference = .34, 95% CI = [-.59, 1.27], $d = .06$), or Test 3 recall ($t(64) = .284$, $p = .777$, mean difference = .11, 95% CI = [-.65, .86], $d = .02$). Likewise, presentation format did not affect the TTL criterion ($t(64) = .04$, $p = .210$, mean difference = .04, 95% CI = [-.02, .11], $d = .10$).

Magnitude of the EOL judgments There was no reliable difference in mean judgment depending on whether EOL judgments were made in the difficulty task or the TTL task, $t(63) = 1.49$, $p = .142$, mean difference = .13, 95% CI = [-.05, .31], $d = .37$. Likewise, the mean judgment did not differ reliably depending on if they were made sequentially or simultaneously, $t(64) = .88$, $p = .386$, mean difference = .04, 95% CI = [-.06, .14], $d = .11$ (see Table 2).

Semantic relatedness as an objective and subjective indicator of item difficulty

In the present study, we reasoned that by varying semantic relatedness in the material, a presumably salient and predictive cue, the participants would use this cue to guide their EOL

Fig. 2 Accuracy of EOL judgments (G), as a function of presentation format (simultaneous vs. sequential) and EOL task (difficulty vs. TTL), separately shown for the two criterion measures, binary (first panel) and TTL (second panel). Error bars represent the 95% confidence interval

Table 2 Means and standard deviations of judgments on the EOL tasks and number for correctly recalled items (maximum 40) at Test 1, Test 2, and Test 3, as well as a TTL criterion

Presentation format	Recall				EOL task	
	Test 1	Test 2	Test 3	TTL criterion	Difficulty	TTL
Sequential	24.22 (7.11)	31.74 (6.16)	35.65 (4.77)	1.68 (.41)	2.34 (.31)	2.17 (.48)
Simultaneous	25.37 (6.67)	32.08 (5.87)	35.75 (4.93)	1.63 (.39)	2.34 (.35)	2.26 (.47)

judgments, that is, to infer about item difficulty. The accurate EOL judgments in the high-variability condition, as opposed to the inaccurate in the low-variability condition (Exp 1), supports this reasoning. However, stronger support for the actual use of semantic relatedness when making EOL judgments would come from (i) a strong relationship between these two factors and (ii) if the predictive validity of the pre-standardized semantic relatedness towards item learnability is high. If the EOL judgments would be uncorrelated with pre-standardized semantic relatedness, it would speak against the notion that the participants use this cue. Further, if semantic relatedness does not predict actual learnability well, but the EOL judgments do, the participants likely base their judgments on something else. To follow up on this we calculated these two correlations for each participant in the three experiments, but only for the high-variability condition. In the low-variability condition (Exp 1 only; Exp2–3 only included the high-variability condition) we cannot compute any correlation because the pre-standardized semantic relatedness was held constant across items.

Interestingly, across all experiments and conditions the EOL judgments were almost perfectly correlated with the pre-standardized semantic relatedness. The mean correlations varied between .87 and .91. This indicates that the participants relied strongly on the semantic relatedness of the word pairs when judging how difficult they would be to learn.

We found support for the assumption that the pre-standardized semantic relatedness, as a proxy for item difficulty, constituted a valid predictor of item learnability. In the Experiment 1 and 2 the mean gamma correlations varied between .78 and .80. In Experiment 3 this relationship was slightly lower, ranging between .58 and .62 across the different conditions, which may be due to that semantic relatedness had a more restricted range in this experiment. The lists consisted of four categories of semantic relatedness, compared to five categories in Experiment 1 and 2.

If the EOL judgments were indeed so strongly guided by semantic relatedness as the near perfect correlations indicate, EOL accuracy (the correlation between EOL judgments and actual learnability) should be capped by the predictive validity of the semantic relatedness towards learnability per se. At least given visual inspection this indeed seems to be the case. EOL accuracy was .74 in Experiment 1, .76–.82 in Experiment 2 and .61 to .62 in Experiment 3. In sum, semantic relatedness is a valid cue for item difficulty on which the participants heavily rely when they make their EOL judgments, and at least for the present material the predictive validity of this cue in predicting actual item learnability seems to limit how high the EOL accuracy can go.

General discussion

The present study investigated how well pre-study judgments, so-called EOL judgments, predict actual learnability of paired associates. These judgments are potentially important as

they presumably guide early study decisions and may thus affect learning. We demonstrated that EOL judgments can be highly predictive of the ease with which word pairs are learned, but EOL accuracy is strongly impacted by the extent that items vary in difficulty within a to-be-learned list (Exp 1). High variability led to a high accuracy, and low variability led to a low accuracy. We further, and somewhat surprisingly, observed a small but stable increase in EOL accuracy, when the judgments were correlated with a binary criterion compared to a TTL criterion (Exp 1 and 3). No other factors reliably influenced EOL accuracy across experiments, such as presentation format (i.e., whether an item was presented simultaneously or sequentially; Exp. 3) or type of EOL task (difficulty vs. TTL task). The results are discussed below.

Although other metacognitive judgments have also been found to rely on item characteristics reflecting item difficulty (e.g., FOK judgments; Nelson et al., 1986; Schwartz & Metcalfe, 1994), the present study shows this is particularly the case for EOL judgments. In line with the cue-utilization framework and Koriat's (1997) argument that EOL judgments are mainly based on cues in the form of characteristics in the material we demonstrated that within-list variability in terms of semantic relatedness between the words in a word pair strongly affects EOL accuracy. Most importantly, the high-variability condition shows that EOL accuracy, in terms of relative accuracy, can be very high. Because metacognitive accuracy is presumed to be closely related to study-regulation efficacy and learning (Kimball, Smith, & Muntean, 2012), EOL judgments may therefore play an important strategic role in controlling study behavior very early in the learning process. Future research should investigate to what extent pre-study metacognitive monitoring impacts learning, as well as the combined effects of pre- and post-study monitoring on learning.

The finding that within-list variability in item characteristics is key to understand EOL accuracy renders us to believe that some previous studies, with very homogenous item materials, may need to be revisited. For example, Leonesio and Nelson (1990; see also Kelemen, Frost, & Weaver, 2000), who compared several different metacognitive judgments, showed that EOL judgments were the least accurate ($G = .22$). They found a gamma correlation similar in magnitude to ours in the low-variability condition ($G = .21$) and it is noteworthy that their material also consisted of word pairs of similar difficulty (unrelated noun-noun word pairs). Given the reliance of EOL judgments on item characteristics it is not surprising that those judgments are less accurate than judgments made later in the learning process, when those characteristics cannot be used to readily discriminate between items. Idiosyncratic cues, such as whether or not a studied item can be retrieved from memory, are only (or mainly) available later in the learning process, for example when making JOLs, and these cues may help to further discriminate between how well items has been learned. Hence, in a low-variability condition it is plausible to find EOL accuracy to be inferior.

However, it is not self-evident that, for example, JOLs would be more accurate than EOL judgments in a high-variability condition, in which EOL accuracy is high (Exp 1). We suspect that when multiple cues are available the effect may not necessarily be additive (i.e., that post-study judgments would always be more accurate than pre-study judgments), but rather that idiosyncratic cues compete with the very salient and variable cue of semantic relatedness later in the learning process. Hence, EOL accuracy may, in a high-variability condition, be on par with JOL accuracy, even though the latter judgments also benefit from the cues available from the learning process. Future studies should investigate this issue.

Although the present study focused on the relative accuracy of EOL judgments, the TTL task and TTL criterion used throughout the study allowed us to calculate a measure of absolute accuracy. Independent of list, students generally underestimated how fast they would learn

(Exp 1). Further, participants were more accurate when judging items in the high-variability list ($M = -.58$), compared to the low-variability list ($M = -.85$), despite mean EOL magnitude being the same for both lists. Although, the two lists were pre-standardized to be equally difficult, and should thus not affect absolute accuracy; in reality, the low-variability list was easier to learn in the long run, as the participants recalled more items on Tests 2 and 3. This likely explains the difference in absolute accuracy between the two lists. The behavioral significance of the general underconfidence in the learnability of the items is unclear, for example, if it transfers to study behavior.

Albeit very small in terms of the effect size, we consistently found (Exp 1 and 3) that the binary criterion led to higher EOL accuracy than the TTL criterion. A possible explanation for this difference is that the two criterion measures differ in how they are affected by idiosyncratic learning that participants could not foresee when they made their EOL judgments, that is *test-potentiated learning* (e.g., Arnold & McDermott, 2013; Kubik, Olofsson, Nilsson, & Jönsson, 2016). Test-potentiated learning refers to the phenomena that during multiple study-test cycles learners tend to focus more on previously unrecalled items than on those that were, improving learning across trials. Because participants may not be aware of the effect of test-potentiated learning, or do not sufficiently adjust their EOL judgments to accommodate the effect (c.f., Koriat, Bjork, Sheffer, & Bar, 2004), this may have decreased EOL accuracy when the TTL criterion measure was used. This is because the TTL criterion was calculated using performance on all three tests. In contrast, EOL accuracy calculated with the binary criterion, cannot be affected by test-potentiated learning as it is measured with a single test only. This relates to previous findings about a stability bias (e.g., Kornell & Bjork, 2009), that is, people tend to underestimate forgetting or how much they will learn with additional study. Because the level of learning at the time of the first test more closely matched the state of learning when the judgments were made, this may have resulted in higher EOL accuracy with the binary criterion.

Although Experiment 1 initially suggested that the TTL task was associated with a higher EOL accuracy than the difficulty task, this was not replicated in Experiments 2 and 3. In Experiment 1, each item was always judged in the difficulty task first and then in the TTL task making it unclear whether the difference was an effect of the task itself or due to the order of the tasks, an experimental flaw that was corrected in the last two experiments. Hence, the TTL task superiority found in Experiment 1 cannot be due to EOL task, but it may instead indicate another intriguing and unexpected effect; if a judgment is made twice, the second judgment will, under at least some circumstances, be better. To clarify, further active elaboration on the EOL judgments, as when doing two (possibly non-identical) judgments in a row, may improve the judgments and thus EOL accuracy. Relatedly, Buratti and Allwood (2013) found this to be the case in another task, but for absolute accuracy. They investigated the confidence in answers to questions about a viewed video clip, and explicitly asked participants to review and adjust their confidence judgments a second time to make them more accurate. As a result, participants became less overconfident. In sum, the EOL task, as operationalized in this study, did not affect EOL accuracy, nor did the consistency between the EOL task given and the criterion measure used (as when the TTL task is correlated with the TTL criterion as opposed to the three other conditions).

Metacognitive judgments are highly comparative in nature (Koriat, 1997). When people judge how difficult every individual item within a material is, each judgment is

made in relation to the other items. Indeed, relative accuracy is a measure of our ability to discriminate between items. We reasoned that simultaneous presentation would lead to higher EOL accuracy than sequential presentation because only simultaneous presentation allows participants to make a proper comparison between all items. When EOL judgments are made in isolation during sequential presentation, the participants will only have full access to the whole list of items when they reach the last item in a list. Further, each judgment can only be made in relation to the recently presented items or those items that happen to come to mind due to working-memory constraints. Hence, we reasoned, this could potentially bias EOL judgments. However, Experiment 3 did not support this notion as EOL accuracy was the same in both conditions. Likewise, there were no reliable differences in absolute accuracy as a function of presentation format. Notably, Dunlosky and Thiede (2004) demonstrated that the presentation format affects study-regulation behavior (i.e., item-selection for restudy), but that this effect was mainly demonstrated for individuals with low-memory span, not for those with a high-memory span. Future studies should therefore include this variable as well. However, potentially, participants do not need to rely on full access to the other list items in order to make reliable EOL judgments; instead, they may rather rely on previous pre-experimental experience of the difficulty of word pairs. The latter explanation is consistent with the present nil-finding.

Concluding remarks

Because EOL judgments have hardly received any attention in the literature, as also noted by others (Dunlosky & Metcalfe, 2009), the main aim of the present study was to alleviate this lack of research. The study clearly demonstrated that variability of a salient cue (i.e., semantic relatedness) within a list strongly affects the accuracy of EOL judgments. That is, EOL accuracy can be high when the material varies sufficiently in item difficulty to allow the participants to discriminate between the items. However, other factors such as the EOL task, criterion measure, and the presentation format, do not seem to have any practically important influence on EOL accuracy.

The present study shows that EOL accuracy co-varies with within-list variability, but how the present finding applies to, for example, glossary lists in language learning textbooks, is unknown, as we have not investigated the real-world prevalence of variability in such glossary lists. Are these lists generally high- or low-variable? Because such lists contain to-be-learned items from to-be-learned texts and not only nouns, but other word classes as well, a speculative guess would be that these lists will generally be more variable than the experimentally designed non-variable lists in some experiments (e.g., Leonasio & Nelson, 1990; the low-variability list in Experiment 1 of this study). Future studies should look into this.

In sum, EOL accuracy seems to be highly dependent on the extent to which items vary in difficulty, but is otherwise stable across a variety of conditions. Importantly it is also largely unknown how strategically important these judgments are for our learning and this remains an important avenue for future research.

Compliance with ethical standards All experiments were conducted in full in accordance with the ethical principles outlined on <http://www.codex.vr.se/> and with the 1964 Helsinki declaration and its later amendments. All participants gave their informed consent before participation.

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 940–945.
- Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, 28(5), 610–632.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55–68.
- Britton, B. K., Van Dusen, L., Gülgöz, S., Glynn, S. M., & Sharp, L. (1991). Accuracy of learnability judgments for instructional texts. *Journal of Educational Psychology*, 83(1), 43.
- Buratti, S., & Allwood, C. M. (2013). The effects of advice and “try more” instructions on improving realism of confidence. *Acta Psychologica*, 144, 136–144.
- Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica*, 95(3), 239–253.
- 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., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks: Sage Publications, Inc.
- Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, 20(4), 374–380.
- Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, 32(5), 779–788.
- Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56(4), 208–216.
- Jönsson, F. U., & Kerimi, N. (2011). An investigation of students’ knowledge of the delayed judgements of learning effect. *Journal of Cognitive Psychology*, 23(3), 358–373.
- Jönsson, F. U., & Lindström, B. R. (2010). Using a multidimensional scaling approach to investigate the underlying basis of ease of learning judgments. *Scandinavian Journal of Psychology*, 51(2), 103–108.
- Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, 28(1), 92–107.
- Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4), 923–954.
- Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370.
- Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one’s own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133(4), 643–656.
- 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.
- Kubik, V., Olofsson, J. K., Nilsson, L.-G., & Jönsson, F. U. (2016). Putting action memory to the test: Testing affects subsequent restudy but not long-term forgetting of action events. *Journal of Cognitive Psychology*, 28, 209–219.
- Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3), 464.
- Lippman, L. G., & Kintz, B. L. (1968). Group predictions of item differences of CVC trigrams. *Psychonomic Science*, 12(6), 265–266.
- Mazzoni, G., Cornoldi, C., Tomat, L., & Vecchi, T. (1997). Remembering the grocery shopping list: A study on metacognitive biases. *Applied Cognitive Psychology*, 11(3), 253–267.
- McCarley, J. S., & Gosney, J. (2005). Metacognitive judgments in a simulated luggage screening task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, pp. 1620–1624). SAGE publications.
- Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 509–527.

- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133.
- 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(4), 267–270.
- Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4), 676.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173.
- Nelson, T. O., & Narens, L. (1994). *Why investigate metacognition?* (pp. 1–25). Cambridge: The MIT Press.
- Nelson, T. O., Leonesio, R. J., Landwehr, R. S., & Narens, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2), 279–287.
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in python. *Journal of Neuroscience Methods*, 162(1–2), 8–13.
- Peynircioğlu, Z. F., Brandler, B. J., Hohman, T. J., & Knutson, N. (2014). Metacognitive judgments in music performance. *Psychology of Music*, 42(5), 748–762.
- Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, 137(1), 131–148.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools Inc..
- Schwartz, B. L., & Metcalfe, J. (1994). *Methodological problems and pitfalls in the study of human metacognition* (pp. 93–113). Cambridge: The MIT Press.
- Schwartz, B. L., Boduroglu, A., & Tekcan, A. İ. (2016). Methodological concerns: The feeling-of-knowing task affects resolution. *Metacognition and Learning*, 11, 305–316.
- Shaps, L. P., Johansson, B. S., & Nilsson, L.-G. (1976). *Swedish association norms*. Uppsala: Dept. of Psychology, University of Uppsala.
- Sikström, S., & Jönsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review*, 112(4), 932–950.
- Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 204.