October, 2013

The Diminishing Criterion Model for Metacognitive Regulation of Time Investment

Rakefet Ackerman

Faculty of Industrial Engineering and Management,

Technion–Israel Institute of Technology, Haifa, Israel

Corresponding Author - E-mail: ackerman@ie.technion.ac.il.

Author note

The study was supported by grants from the Israel Foundation Trustees (2011-2013) and by the Israel Science Foundation (Grant No. 957/13). The ample time conditions of Experiment 3 and Experiment 4 (*N* = 22 in each condition) were also reported in the proceedings of the 35th annual conference of the Cognitive Science Society (Ackerman, 2013), with partial data analyses. I would like to thank Daniel Gopher, Valerie Thompson, Joëlle Proust, Janet Metcalfe, and anonymous reviewers for their helpful feedback on earlier versions; Keith Thiede and Valerie Thompson for kindly conducting new analyses on their previously reported data; and Meira Ben-Gad for editorial assistance.

Abstract

According to the discrepancy reduction model for metacognitive regulation, people invest time in cognitive tasks in a goal-driven manner until their metacognitive judgment, either judgment of learning (JOL) or confidence, meets their preset goal. This stopping rule should lead to judgments above the goal, regardless of invested time. However, in many tasks time is negatively correlated with JOL and confidence, with low judgments after effortful processing. This pattern has often been explained as stemming from bottom-up fluency effects on the judgments. While accepting this explanation for simple tasks, like memorizing pairs of familiar words, the proposed *Diminishing Criterion Model* (DCM) challenges this explanation for complex tasks, like problem solving. Under the DCM, people indeed invest effort in a goal-driven manner. However, investing more time leads to increasing compromise on the goal, resulting in negative time-judgment correlations. Experiment 1 exposed that with word-pair memorization, negative correlations are found only with minimal fluency and difficulty variability, while in problem solving they are found consistently. As predicted, manipulations of low incentives (Experiment 2) and time pressure (Experiment 3) in problem solving revealed greater compromise as more time was invested in a problem. Although intermediate confidence ratings rose during the solving process, the result was negative time-confidence correlations (Experiments 3, 4, and 5), and this was not eliminated by the opportunity to respond by “don’t know” (Experiments 4 and 5). The results suggest that negative time-judgment correlations in complex tasks stem from top-down regulatory processes with a criterion that diminishes with invested time.

Performing cognitive tasks, such as learning, reasoning, problem solving, and decision making, requires representation of information provided within the task, activation of relevant knowledge, and performance of mental operations directed towards goal achievement (Wang & Chiew, 2010). Beyond these, performing sets of such tasks also involves regulation aimed at deciding whether to invest additional effort in the present item or to move on to the next one (Ariel, Dunlosky, & Bailey, 2009; Kruglanski et al., 2012). What guides people as a stopping rule for the investment of cognitive effort in a particular item? The present study considers a metacognitive stopping rule for effort investment in such cognitive tasks.

The metacognitive literature suggests that what underlies decisions regarding the investment of cognitive effort is subjective judgment, or monitoring, of the quality of the current state of performance (Nelson & Narens, 1990). A widely accepted model, known as the discrepancy reduction model, suggests that people set a target knowledge level, or goal, according to their motivation in the given scenario, and that this goal is used as a stopping criterion. For example, in learning tasks, people continue to invest time in an attempt to improve their knowledge until their judgment of learning (JOL) reaches this preset goal, which represents what they consider a satisfactory likelihood of success (Dunlosky & Hertzog, 1998; Nelson & Narens, 1990; see illustration in Ackerman & Goldsmith, 2011) . Support for this notion comes from consistent findings that people invest more time in studying more difficult items (see Son & Metcalfe, 2000, for a review). Additional support comes from the finding that despite monitoring differences in the initial stages of learning texts in two contexts (screen vs. paper presentation) and despite achievement differences, participants stopped studying with equivalent predictions of performance under the two contexts (Ackerman & Goldsmith, 2011). This finding suggests that the participants stopped studying based on their learning goal, which was not affected by the study context. Similar ideas were recently discussed in the literature dealing with reasoning and decision making (J. St. B. T. Evans, 2006; Pleskac & Busemeyer, 2010; Yeung & Summerfield, 2012). In such tasks, the theory suggests that people invest effort in accumulating evidence that increases their confidence in their answer, until they meet their preset threshold. This threshold reflects a level of confidence satisfactory for providing an answer.

If indeed people follow these discrepancy reduction models, we would expect the correlation between invested time and people's subjective judgment of their current performance, either JOL or confidence, to be weak or even nil. This is because people are expected to invest time until they judge they have reached (or surpassed) a preset stopping criterion, regardless of the time it takes them to reach it. Indeed, in line with this prediction, no correlations were found with metacomprehension judgments regarding various reading comprehension tasks (unpublished analyses: Ackerman & Goldsmith, 2011; Ackerman, Leiser, & Shpigelman, 2013; Thiede & Anderson, 2003). However, with many other tasks, persistent inverse relationships between time and metacognitive judgments have been reported over the years. This was the case with JOLs when the task was memorizing paired associates (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Koriat, Ackerman, Adiv, Lockl, & Schneider, in press), and with confidence judgments when tasks involved answering knowledge questions, eyewitness questioning, decision making, reasoning, and problem solving (e.g., Ackerman & Koriat, 2011; Glöckner & Betsch, 2008; Kelley & Lindsay, 1993; Pleskac & Busemeyer, 2010; Robinson, Johnson, & Herndon, 1997;

unpublished analysis for Thompson, Prowse Turner, & Pennycook, 2011; Thompson et al., 2013).

Notably, there seems to be a contradiction here: People's judgments should simultaneously reflect that (a) by investing more time, they enhance their likelihood of success (by improving their knowledge/answer or acquiring more convincing supportive evidence), and (b) responses that take longer to provide have a lower likelihood of success. In an attempt to resolve this conflict, Koriat, Ma'ayan, and Nussinson (2006) distinguished between *data-driven* and *goal-driven* investment of effort as cues for JOL and confidence. They suggested that time and judgments are negatively correlated when the ease of processing, or fluency, of each item provides a heuristic cue for the metacognitive judgment. That is, in a memorization task, for example, items committed to memory quickly are judged after learning as more likely to be recalled than those which take longer to study (see also Begg et al., 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003). In contrast, when learners are especially motivated to succeed in particularly important items, their goal-driven investment of study time in these items enhances both the objective and assessed likelihood of success, with a positive time-judgment correlation. To examine their explanation, Koriat et al. (2006; Experiment 5) assigned lower or higher incentives (point values) to each paired associate, while also using items at various levels of difficulty. As expected, they found a negative correlation between time and JOL for items at the same incentive level (Figure 1, solid lines), and a positive correlation between time and JOL when they analyzed items that were assigned different point values (see Figure 1, dashed lines; see also Koriat et al., in press). This distinction between data-driven and goal-driven effort bears some similarity to the distinction between reactive control and effortful control of emotional self-regulation (see Derryberry & Rothbart, 1997) and between effort investment in a demanding task versus motivated cognition (Kruglanski et al., 2012). By this notion, data-driven effort investment is led by the item in a bottom-up fashion, while goal-driven effort is a top-down process where deliberate effort is invested to improve the likelihood of success. The metacognitive judgment is inferred by a delicate attribution process taking into account the two contradictory sources of effort—one that reduces and another that enhances the likelihood of success (see Koriat et al., in press).

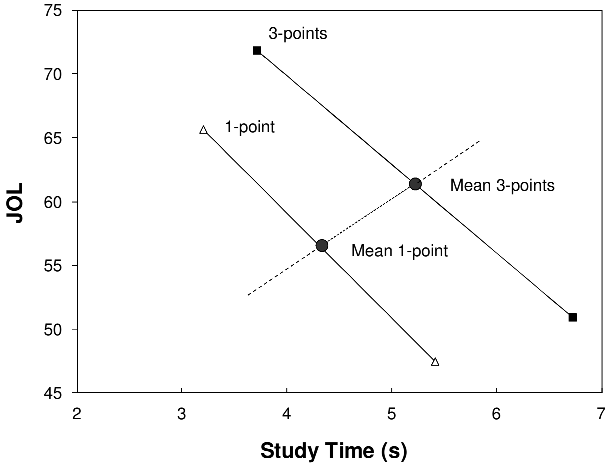

Figure 1. Mean judgment of learning for 1-point and 3-point items as a function of mean study time invested in these items (broken line). Also plotted are judgments of learning for below-median and above-median study time for each incentive level (full lines; adapted from Figure 11, Koriat et al., 2006).

Koriat et al. (2006) developed their theory using a memorization task, but they also replicated their study with a figural problem-solving task (similar to Raven's matrices; Experiment 7). After providing a solution for each problem, the participants rated their confidence in their solutions (i.e., their estimate of the likelihood their solution was correct). Despite the different task, Koriat et al. found a pattern of results highly similar to the memorization results presented in Figure 1. Thus, although there is a good explanation as to why there should be either a positive correlation (goal-driven regulation) or a weak correlation between time and judgment (discrepancy reduction regulation), the negative correlation dominated Koriat et al.'s (2006) results. In their case, the negative correlation reliably reflected the relationship between response time and actual performance. However, Ackerman and Zalmanov (2012) found a persistent negative correlation between problem-solving time and confidence even in cases where response time was non-predictive of accuracy, as solutions produced quickly and slowly were equally likely to be correct.

As with the memorization task, Koriat et al. (2006) explained the negative time-confidence correlation they found in the problem-solving task by referring to the distinction between data-driven and goal-driven investment. In other words, by their theorizing, these findings mean that bottom-up, data-driven processes are more dominant than top-down, goal-driven regulatory processes in informing the problem solver's confidence judgments, as is the case with JOLs.

There is strong evidence that meta-reasoning processes, such as those involved in problem solving, resemble meta-memory processes in some respects, but differ in others (see Ackerman & Thompson, in press, for a review). In this regard, it is useful to consider Funke's (2010) distinction between simple cognitive tasks, like perception or memorization, and complex tasks, like problem solving or decision making, which involve multi-step, goal-directed information processing. In line with this distinction, the dominance of bottom-up effort investment and its utilization as a cue for the metacognitive judgment seem to accord with the nature of simple tasks. Although data-driven effort investment probably plays some role in complex tasks, there is room for questioning its dominance over top-down, goal-driven processes that underlie the regulation of effort and the final metacognitive assessment of performance in these contexts.

Regulation of a large variety of complex tasks has been discussed in depth under the umbrella of the dual-process theory (J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2003; Stanovich & West, 2000). Under this approach, System 1 or Type 1 (T1) processes produce initial responses that come to mind quickly based on default reasoning procedures. System 2 or Type 2 (T2) processes execute deliberate, lengthy, and analytic reasoning in a goal-driven manner. Thompson and her colleagues (Thompson et al., 2011; Thompson et al., 2013) found that high feeling of rightness judgments provided regarding the very first response that came to mind in various reasoning and problem-solving tasks predicted both less reconsideration time and lower likelihood of altering the initial answer, relative to a low feeling of rightness. According to Thompson and her colleagues, feeling of rightness judgments regarding the output of T1 are the trigger according to which T2 processes are activated (or not) before providing a response to a complex task.

As for the time-confidence correlation, Thompson et al. (2011, 2013) found the correlation between initial response time and the feeling of rightness to be negative, and interpreted that in light of the fluency principle, consistently with the data-driven notion. As described above, they also found (although did not report in their paper) a negative time-confidence correlation. However, the processes that determine the time invested in T2 and ultimate confidence are still unknown. By the dual-process conceptualization, the T2 processes required for performing complex tasks are deliberate and goal-directed (J. St. B. T. Evans, 2006). Under Koriat et al.'s (2006) theorizing, this should have led to a positive correlation between time and confidence, while, as explained above, under the discrepancy reduction model it should have led to a weak correlation between them. Thus, the findings of negative correlations between response time and confidence across decision-making, reasoning, and problem-solving tasks that involve T2 processing remain as yet unexplained.

The Diminishing Criterion Model

The proposed model deals with the hypothesized metacognitive stopping rule for complex tasks such as those described above. The consistent negative correlations between time and metacognitive judgments in these tasks have typically been interpreted, including by Ackerman and Zalmanov (2012), in terms of bottom-up fluency (e.g., Kelley & Lindsay, 1993; Koriat et al., 2006; Unkelbach & Greifeneder, 2013). The present study suggests an alternative explanation with dominance of a top-down regulatory process.

According to the *Diminishing Criterion Model* (DCM), the stopping rule for complex tasks is a criterion similar to that suggested by the discrepancy reduction model, where people continue to invest effort until they subjectively reach their target level. In support of perceived progress with time, previous studies found that intermediate judgments rise, with positive correlation with the invested time, during performance of learning, problem-solving, and reasoning tasks (Ackerman & Goldsmith, 2011; Metcalfe & Wiebe, 1987; Thompson et al., 2011; Vernon & Usher, 2003). The novel addition suggested here is that when negative correlations are found between time and final metacognitive judgments, this is because people’s goals are not constant but, rather, shift downward with time. In other words, as processing is prolonged, the respondent becomes increasingly willing to compromise—to provide a response with less confidence that it is correct. See illustration in Figure 2.

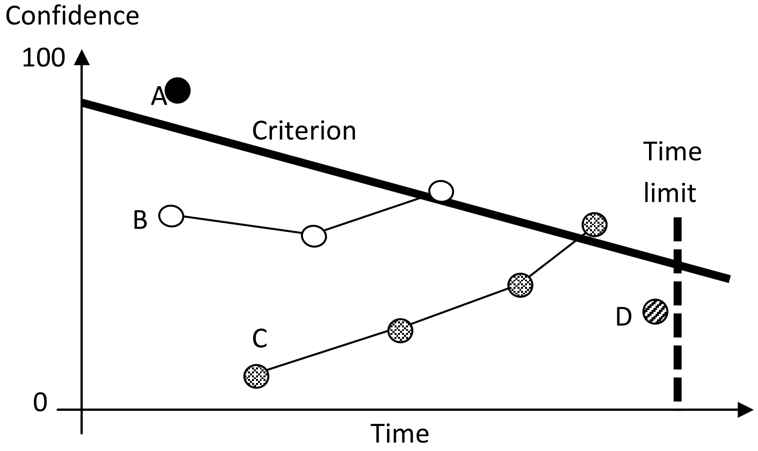

Figure 2. The Diminishing Criterion Model (DCM) with hypothetical confidence ratings for four items and a time limit. The thick black line shows how the stopping criterion diminishes as the time limit approaches.

Figure 2 presents hypothetical confidence ratings for four problems with a time limit, and a diminishing stopping criterion. In all four cases, the solver reads the problem at the beginning of the timeline (the origin of the X axis), and then begins to consider possible solutions. As time passes, solvers evaluate their confidence regarding possible solution options against their stopping criterion for that point in time. The four hypothesized cases will illustrate. First, in case A, a solution comes to mind quickly with sufficiently high confidence to terminate the process for this problem. In case B, a solution similarly comes to mind quickly, but the solver's confidence in this solution is too low relative to the stopping criterion. This process yields reconsideration of the same or another solution candidate that is still not satisfactory, and a third with a confidence rating similar to the initial one, but high enough to meet the stopping criterion for this point in time. In case C, no immediate solution comes to mind, and a slow formulation of potential solutions is developed until the solver's confidence is sufficiently high. In case D, a very slow problem-solving process eventually yields an unsatisfactory solution, and the respondent is not willing to invest further effort in this problem. In this case, the respondent might prefer to withhold the solution, if given the option.

Overview of the study

The five experiments that comprise this study were designed to examine predictions derived from the DCM. In line with Koriat et al. (2006), in the present study, problem-solving tasks were used to examine the model, as representatives of complex tasks. Generalizing beyond Koriat et al.'s particular stimuli, the problems chosen for Experiments 1-4 were Compound Remote Associate (CRA) problems. Each CRA problem consists of three words, and the task is to find a fourth word which forms a compound word or two-word phrase with each one[1](see Bowden & Jung-Beeman, 2003). For example, for the triplet PINE/CRAB/SAUCE the correct answer is APPLE, resulting in PINEAPPLE, CRABAPPLE, and APPLESAUCE. CRA (and RAT) problems are widely used in studies dealing with a range of cognition-related questions, including neurological correlates for cognitive behavior (e.g., Kounios et al., 2006; Sandkühler & Bhattacharya, 2008), the relationship between affect and cognition (see Topolinski, in press; e.g., Topolinski & Strack, 2009), and creativity in work contexts (e.g., Miron-Spektor, Efrat-Treister, Rafaeli, & Schwarz-Cohen, 2011; Probst, Stewart, Gruys, & Tierney, 2007). Besides allowing generalization beyond Koriat et al.'s (2006) stimuli, CRA problems offer another advantage for the current study, in that they require complex processing with only a minimal reading challenge (see Thompson et al., 2013, on the difference between perceptual fluency and

[1] CRA problems are a subset of the Remote Associate Test (RAT; Mednick, 1962). The distinguishing feature of CRA problems is that the solution word is not merely associated with each cue word, as is the case in RAT problems, but forms a compound word or two-word phrase with each one.

processing fluency). As will be seen, this was important for Experiment 1, as it allowed comparison between a (complex) CRA task, which involves reading three words, and a (simple) word-pair memorization task, involving reading of two words per item. Experiment 5 generalized the findings to yet another type of a problem-solving task: misleading math problems of the type often used for studies in the context of the dual-process theory. For both task types, the relevant metacognitive judgment is a confidence rating, which reflects the solver's assessment of the likelihood that the solution is correct.

The detailed predictions derived from the DCM are explained in the introduction to each experiment. In brief, the purpose of Experiment 1 was to expose boundary conditions for the bottom-up effect of fluency and a difference in this respect between simple and complex tasks. This difference calls for an alternative explanation for the negative correlations in complex tasks. Experiment 2 and Experiment 3 examined the prediction that manipulations that have been found to lower the stopping criterion, low incentive for success and time pressure, in fact do not affect all responses equally, but affect instances that involve lengthy processing more than those involving quick processing. Experiment 3, Experiment 4, and Experiment 5 were based on findings that intermediate metacognitive judgments rise (either linearly, logarithmically, or with a sudden spike) during learning, reasoning, and problem solving (Ackerman & Goldsmith, 2011; Metcalfe & Wiebe, 1987; Thompson et al., 2011; Vernon & Usher, 2003). These experiments used intermediate confidence ratings to examine whether this upwards progress coexists with a negative time-confidence correlation. Experiment 4 and Experiment 5 also examined the prediction that a free-report procedure (which allows participants to provide a "don't know" response) does not eliminate volunteering of low-confidence solutions after lengthy thinking, even under conditions of time pressure. This point responds to the question raised by cases C and D in Figure 2, where a slow problem-solving process eventually yields a solution at a low level of confidence. The free-report procedure allows examination of whether lengthy low-confidence solutions indeed satisfy a diminishing stopping criterion, or reflect abandonment of solving attempts which result in responses participants would prefer to withhold, if given the option.

Experiment 1

As explained above, previous studies have suggested that the negative time-judgment correlation seen consistently in both simple and complex tasks stems from fluency, where respondents use ease of processing as a bottom-up cue for their JOL or confidence judgments (e.g., Kelley & Lindsay, 1993; Koriat et al., 2006). The DCM proposes that in fact, this negative correlation stems from processing fluency only in simple tasks, while in more complex tasks, it stems from top-down regulation. The former are tasks in which data-driven regulation dominates, while in the latter, goal-driven regulation dominates.

The aim of Experiment 1 was to contrast a simple and a complex task, both of which show overall negative time-judgment correlations, and expose a difference between them in their association with processing fluency. The chosen tasks were word-pair memorization and CRA problems. As explained above, CRA problems share a minimal reading challenge with paired associates, allowing a focus on task-directed processing efforts.

Sensitivity to processing fluency was examined by including some highly difficult items in both tasks. Studies of memorization processes often combine related and unrelated paired associates (e.g., SOCK-SHOE vs. KITE-POT), and find consistent negative time-judgment correlations (e.g., Koriat et al., 2006; Undorf & Erdfelder, 2011). However, it is well established in the literature that metacognitive judgments are more sensitive to item characteristics, such as difficulty, when there is within-participant and even within-block variability (Koriat et al., in press; Koriat, Bjork, Sheffer, & Bar, 2004; Yue, Castel, & Bjork, 2013), as is the case with many other judgments (see Unkelbach & Greifeneder, 2013 for a review). For instance, Koriat, Ackerman, Lockl, and Schneider (2009) found (with a sample of 4th graders) a stronger negative time-JOL correlation with a mixed list, including related and unrelated word pairs, than with a list containing only unrelated pairs, despite greater time variability in the latter. This finding raises some doubts regarding the strength of the association between time and JOL when processing the unrelated pairs, which are more challenging and thus less fluent. With CRAs, in contrast, Ackerman and Zalmanov (2012) found the negative time-confidence correlation to be as strong for wrong responses as for correct responses, although the former generally took longer to provide. Thus, low processing fluency did not weaken the correlation.

The hypothesis guiding the design of the present experiment was that difficult items in both tasks would expose a boundary condition for the dependency of metacognitive judgments on processing fluency. That is, it was hypothesized that for word-pair memorization, only relatively easy items would show sensitivity of JOLs to invested time, while with CRAs, a negative time-confidence correlation would be seen regardless of the difficulty level. The explanation offered here, and examined in the following experiments, is that for CRAs the negative correlation does not stem from fluency as the main source, but rather from top-down regulation with a diminishing stopping criterion.

How can the challenge imposed by word pairs be increased? Notably, in commonly used word pairs, the stimuli are associations between familiar words. Several studies using paired associates have challenged participants by employing less familiar stimuli, such as foreign vocabulary (e.g., Swahili; Jönsson & Kerimi, 2011; Townsend & Heit, 2011), unknown pictures (Gruppuso, Lindsay, & Masson, 2007), and associations between concepts and categories (e.g., Kornell & Bjork, 2008; Wahlheim, Finn, & Jacoby, 2012). However, these studies did not focus on the time-JOL relationship. For the present study, I also used associations between concepts and categories, but generated large variability in difficulty by using five types of associations, similarly to Gruppuso et al.'s pictures. Type A comprised commonplace concept-category associations that were expected to be known to most participants (e.g., CUBA - ISLAND). Type B comprised familiar concepts and categories whose association is less well known (e.g., BAHRAIN - ISLAND). In types C and D, one member of the pair presented a familiar category (type C) or concept (type D), while the other was designed to be less commonly encountered (e.g., KIRIBATI - ISLAND; RADISH - CRUCIFERAE). Finally, in type E, both members of each pair were intended to be unfamiliar (e.g., ERUCARIA - CRUCIFERAE). While all the concept-category associations were inherent and not arbitrary (as with unrelated word pairs), the associations were designed to rise in difficulty (or to become less easy) from type A to type E. Memorization was self-paced, and the participants provided JOLs immediately after studying each item. The negative time-JOL correlation was expected to weaken as difficulty rose.

For comparison, the data of the CRA problems collected by Ackerman and Zalmanov (2012) were reanalyzed by dividing the problems into difficulty levels parallel to the difficulty gradations among the word pairs. For them, the negative time-confidence correlations were expected to remain strong even for the most difficult items.

Method

Participants. Thirty-two undergraduates participated in the word-pair experiment for course credit (M_{age} = 24.7; 56% females). There were 28 participants in Ackerman and Zalmanov's (2012) CRA study reanalyzed here as described above. The two samples were drawn from the same population.

Materials. The word pair stimuli comprised sixty-two pairs, each made up of a concept and associated category. Two pretests were used to categorize concepts and categories as more or less familiar as a basis for construction of the five types (A through E, detailed above). In Pretest A, participants (*N* = 34) were presented with a list of 74 categories from 6 domains (foods, states/nations, diseases, animals, famous people, and plants; hobbies were used for practice); in each case, they were asked to describe the category in their own words and to name three concepts belonging to it. Those categories for which more than 50% of the participants wrote accurate descriptions and named three appropriate exemplars were categorized as known.

In Pretest B, participants (*N* = 68) were presented with concepts drawn from the lists created by the participants of Pretest A and lesser-known concepts in the same categories. The concepts were presented one by one on the computer screen, and each participant saw 100 concepts chosen randomly from a total of 309. In the first block, the participants rated their subjective familiarity with each concept on a 4-point scale. In the second block, the participants saw each concept again, but this time, they were asked to choose, first, the appropriate domain from a list that appeared on-screen, and then the appropriate category from a list related to the chosen domain. Thus, correct category choice also indicated correct domain choice. As with the categories in Pretest A, those concepts that were accurately identified by most participants were categorized as known.

The types of concept-category pairs were defined based on four parameters collected in the pretests. For each Type A pair, (a) the concept was perceived as familiar in the first block of Pretest B with a mean familiarity rating above 2.5; (b) the category it related to was known to more than 50% of Pretest A participants; it was assigned to (c) the right domain and (d) the right category by more than 80% in the second block of Pretest B. Type B pairs satisfied the first three criteria, but the category assignment in the second block of Pretest B was correct for less than 50% of the participants. In Type C pairs the concepts were familiar, but their category was known to less than 30% of Pretest A participants and, as in Type B, they were assigned the correct category by less than half the participants in block two of Pretest B (domain knowledge was not restricted). Type D concepts were rated as unfamiliar, and their category was familiar (Pretest A) but not assigned correctly (Pretest B, correct category < 50%) (domain knowledge was not restricted). In Type E, both the concepts and categories were rated as unfamiliar (Pretest A), and the concepts were rarely assigned to the correct category (< 30%) or even domain (< 40%). Twelve concept-category pairs from each type were chosen to be included in the study. Two additional pairs were used for practice. The 60 word pairs used for the experiment represented 23 categories, with 1 to 5 ($M = 2.7$) concepts per category. In all pairs, the concept was used as the cue word and the category as the target.

The CRA problems were the 34 problems used in Ackerman and Zalmanov (2012). Two of the problems were used for demonstration and two for self-practice.

Procedure. Participants were presented with one pair or problem at a time, in a random order. They pressed the “Present” button on an empty screen when ready to study/solve an item, and the “Continue” button when they were done. Both tasks were self-paced. Pressing "Continue" exposed a JOL scale for the word pairs and a confidence scale for the CRAs. In both cases the judgment was provided on a horizontal scale, along which an arrow could be dragged from 0 to 100%.

At the test phase of the word-pair memorization task, the concept appeared as the cue, and the task was to recall the associated category. For both tasks, after the experimental session ended, an experimenter reviewed the answers and manually marked answers that were mismarked as wrong by the automated system because of typos and meaningless differences from the formal correct answer (e.g., “islands” instead of “island”). The experimenter also manually differentiated between wrong answers and non-words (e.g., “--” or “no idea”).

Results and discussion

In the memorization task, the participants responded with meaningful words (as opposed to non-words like “--”) for 99% of the items. The mean success rate (*M* = 53.0, *SD* = 12.1) was highly similar to that found by Ackerman and Zalmanov (2012) for the CRAs (*M* = 52.0; *SD* = 9.6). Study time for the word pairs was short (*M* = 3.42 sec., *SD* = 1.43) — about a tenth of the response latency for the CRAs (*M* = 40.4 sec., *SD* = 6.2). The time-judgment slopes were examined by mixed linear regressions (Proc Mixed macro of SAS© 9.2). Effect sizes are expressed by the variance explained by the examined factor (R^2), study time or response latency in the present case. As found in many previous studies, study time predicted JOL reliably, $b = -1.32$, $t(1613) = 8.28$, $p < .0001$, $R^2 = 29\%$.

However, the predictive value of response latency for confidence in CRAs had a stronger effect, $b = -0.80$, $t(812) = 28.6$, $p < .0001$, $R^2 = 51\%$. As a first descriptive glance into the time-judgment relationships in the two tasks, the slope of each participant was analyzed separately. For the word pairs, the slopes were significant ($p < .05$) for 11 participants (34%), marginally significant ($p < .10$) for 4 participants, and not significant for the remaining 17 (53%) participants. For the CRAs, in contrast, the slopes were significant ($p < .005$) for all participants.

As explained above, if fluency as operationalized by time serves as a central cue for JOL, it should take its effect even when analyzing each difficulty level separately. Table 1 presents the means (and SDs) of success rates, study times and JOLs for each of the five types of word pairs, the significance of the differences between them, and the statistical information regarding the regression lines. As can be seen in the table, there were in fact four difficulty levels, as word pairs of types B and C had similar success rates. The CRAs were therefore divided into quarters by global success rates for each problem, allowing 7-8 problems at each difficulty level, with similar mean success rates to those of the word pairs. Table 1 also presents the results for the CRA quarters. Figure 3 presents the regression lines graphically. The lengths of the lines in Figure 3 illustrate the time variability within each difficulty level (see figure caption).

Table 1. Means (SD) for word pairs in Experiment 1, divided by type, and compound remote associate (CRA) problems in Ackerman and Zalmanov (2012, Experiment 2), divided by difficulty level.

Word pairs

Pair type	Success rate	Study time	Judgment of Learning (JOL)	Time-JOL slope				
				b	*df*	*t*	*p* ≤	R^2(%)
A	85.4 (13.2)^{a}	3.2 (1.9)^{a}	77.6 (11.8)^{a}	-2.08^{a}	398	6.30	.0001	4
B	63.5 (18.7)^{b}	4.6 (3.1)^{b}	55.9 (10.3)^{b}	-.95^{b}	434	3.26	.001	2
C	60.8 (17.2)^{b}	5.0 (3.6)^{bc}	50.1 (11.7)^{c}	-.40^{b}	426	1.44	.15	1
D	36.2 (17.3)^{c}	5.4 (4.1)^{c}	38.8 (12.9)^{d}	-.43^{b}	454	1.72	.09	1
E	27.6 (15.3)^{d}	6.1 (4.3)^{d}	28.9 (14.9)^{e}	.45^{c}	447	2.31	.02	1

CRAs

Difficulty level	Success rate	Response time	Confidence	Time-confidence slope				
				b	*df*	*t*	*p* ≤	R^2(%)
Most easy	90.8 (14.4)^{a}	15.6 (8.7)^{a}	92.2 (11.9)^{a}	-.82^{ab}	155	12.81	.0001	50
Somewhat easy	62.6 (17.8)^{b}	37.5 (12.3)^{b}	65.7 (19.0)^{b}	-.94^{a}	273	19.00	.0001	52
Somewhat difficult	41.1 (18.3)^{c}	46.5 (14.4)^{c}	55.0 (17.9)^{c}	-.66^{bc}	135	8.94	.0001	59
Most difficult	20.2 (14.8)^{d}	57.2 (17.4)^{d}	42.6 (19.2)^{d}	-.58^{c}	247	10.62	.0001	52

$^{a\ b\ c\ d\ e}$ – Different letters signify significant differences between types or levels ($p < .05$).

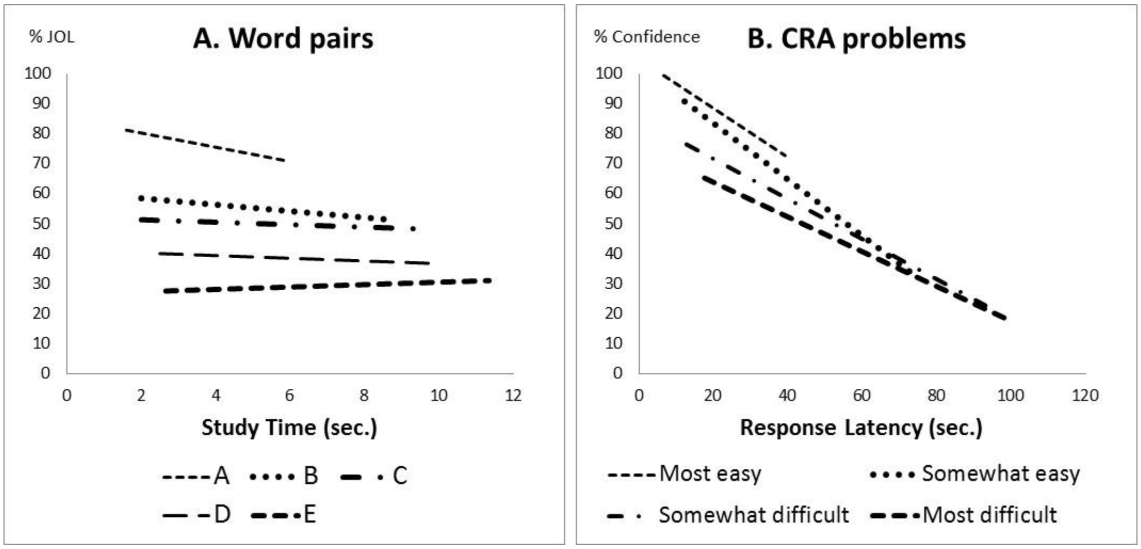

Figure 3. Experiment 1 - Panel A presents the regression lines predicting judgments of learning (JOL) by study time in word-pair memorization for item types in increasing order of difficulty (A to E). Panel B presents the regression lines predicting confidence by response latency in compound remote associate (CRA) problems used by Ackerman and Zalmanov (2012) according to level of difficulty (most easy to most difficult). The start and end points of the lines represent the means for the 25% of responses provided most quickly and the 25% provided most slowly per participant in each type or difficulty level.

An interesting observation when looking at the overall pattern of results in Figure 3 is that in both task types, the predictive lines in Figure 3 draw closer as response times lengthen. This suggests that while metacognitive judgments reliably show a difference among difficulty levels for those responses produced quickly, this differentiation weakens as participants invest more effort. Notably, the lines draw closer not at high levels of JOL or confidence, as one might expect to result from effortful thinking, but at medium-low levels.

A closer look at the findings supports the prediction that the negative time-judgment correlation depends on the type and difficulty of the task. In word-pair memorization, the effect sizes of the regression lines (last column in Table 1) were much weaker when the data were broken down by difficulty level than the overall effect reported above, when the data were considered across difficulty levels. The study time-JOL correlation was negative but weak for types A and B. It was weaker for type B, which involved less-familiar associations, than for type A, where the associations were chosen to be well-known. For types C and D, time did not reliably predict JOL, and for type E the relationship was actually positive, but weak. For the problem-solving task, in contrast, the negative time-confidence relationships had consistent large effects, similar in size to the overall effect across difficulty levels. Nevertheless, in this case as well, the slopes were shallower for the more challenging items.

Why does the predictive value of latency for confidence in CRA answers have such a strong effect? A potential alternative explanation for considering the negative confidence slopes as associated with response times is that confidence is tightly associated with answer accuracy, and correct solutions are provided more quickly than wrong solutions. Indeed, this was the case under all conditions, all *p*s < .0001. This information source was absent for participants in the memorization task when they rated their JOLs, because the judgments were provided immediately after studying, with both words present, without a diagnostic attempt to recall them (the interested reader may refer to explanations of the delayed JOL effect, Dunlosky & Nelson, 1992; see Rhodes & Tauber, 2011 for a review). When solution accuracy was controlled for (partialled out) in the regression analyses, the latency-confidence slopes remained significantly negative, all *p*s < .0001 with residual effect sizes above 18%. This finding assures us that the association between latency and confidence is strong above and beyond the association between latency and accuracy (see additional support in Ackerman & Zalmanov, 2012). To ensure a focus on the effect of latency on confidence, accuracy was controlled for in all the time-confidence regression analyses reported in this paper.

The differences between the tasks in the strength and consistency of the association between time and metacognitive judgment demonstrates that bottom-up fluency does not always provide a satisfactory explanation for the negative time-judgment correlation. For the word-pair memorization task, the negative correlation depends on the items' processing fluency and pronounced difficulty variability among them. The CRA problems, in contrast, show consistent negative time-confidence correlations, regardless of processing fluency. The explanation suggested here is that word-pair memorization is a simple task dominated by bottom-up fluency, which thus falls out of the scope of the DCM, while CRA problems represent a complex task dominated by a top-down regulatory process that nevertheless ends up with a negative correlation, and that fits the scope of the DCM. It should be noted, however, that although by the DCM top-down processes dominate in leading to the negative time-judgment correlations in complex tasks, this does not preclude fluency effects on these tasks as well, as reflected in the slope attenuation with difficulty (Figure 3, Panel B; see General Discussion).

Experiment 2

The aim of Experiment 2 was to begin examining the predictions derived from the DCM by replicating the results of Koriat et al. (2006, Experiment 7) and considering evidence for the proposed alternative explanation for the negative time-confidence correlation. Koriat et al. manipulated incentives for success by allocating low or high point values to each item. In traditional regulation studies, such motivation manipulation was hypothesized to shift the stopping criterion downwards or upwards, respectively (e.g., Koriat & Goldsmith, 1996; Thiede & Dunlosky, 1999). These theories did not take into account the time it takes to perform the task as a factor that affects the criterion level. Similarly, the goal-driven explanation of Koriat et al. (2006) predicts a similar shift in the stopping criterion for both quickly and slowly provided responses (see Figure 1). The prediction under the DCM, in contrast, is that the stopping criterion should not be affected by the motivation manipulation for responses provided quickly. That is, early in the solving process the criterion is expected to be high regardless of the incentive, because until sufficient time has passed there is no reason to compromise. The effect of the motivation manipulation was expected to manifest itself in a more rapid fall in the criterion level for low-incentive items than for high-incentive items. This should lead to a confidence difference between the incentive levels only for responses that take some time to produce. Experiment 2 was designed to distinguish between these two possibilities.

In this experiment, an incentive group and a control group faced CRA problems. For the incentive group, half the problems were assigned 1 point and the others 5 points, indicating higher priority. The overall pattern of results was expected to replicate the findings of Koriat et al., who used Raven-like matrices as their problem-solving task, in that both were expected to end up with negative time-confidence correlations, and both tasks were hypothesized to be dominated by goal-driven regulation of effort. However, in addition, using a mixed linear regression to predict confidence by response latency was expected to expose the pattern of results predicted by the DCM. By this hypothesis, the predicted pattern was concealed in the analysis of Koriat et al. by the median split methodology (see Figure 1). Specifically, the use of a regression model would expose any differences between the incentive levels at the regression line origins (intersects with the Y axis) and in the slopes of the predictive lines. According to Koriat et al. (2006) there should be a difference in the origins but not in the slopes, while according to the DCM, it should be vice versa.

The control group solved the CRA problems with no particular instructions, as in Ackerman and Zalmanov (2012, Experiment 2—the experiment which provided the data analyzed in Experiment 1 above). Their results provided the baseline for this study. An open question was whether in the incentive group, relative to the control group, (a) high incentives would attenuate the slope (less compromise), suggesting stronger motivation for success; (b) low incentives would lead to a steeper slope (greater compromise), suggesting greater readiness to compromise on confidence for the less-important items; or (c) both.

Method

Participants. Sixty undergraduates participated in the experiment for course credit or for payment (M_{age} = 25.6; 28% females). They were randomly assigned to groups such that 20 participants were in the no-incentive group and 40 in the incentive group.

Materials. These were 34 CRA problems, with two of the problems used for demonstration and two for self-practice. The set of problems was based on the 34 problems used by Ackerman and Zalmanov (2012). Ten problems that showed extremely high or extremely low confidence levels were replaced, so as to allow room for variability in confidence ratings for all problems.

Procedure. The experiment was conducted in groups of 2 to 8 participants. For the no-incentive group, the instruction booklet detailed the procedure, explained what comprised a valid solution, and illustrated the procedure using two problems. Pressing a “Start” button on an empty screen brought up each problem. The three words appeared side by side. Respondents had to type the solution into a designated space below the three words and press “Continue” when done. Response time was measured from when participants pressed “Start” to when they pressed “Continue”. Pressing "Continue" exposed the confidence scale. Pressing the “Next” button cleared the screen for the next problem. After the demonstrations, the two self-practice problems appeared first, and the rest were randomly ordered for each participant. The session lasted 30 minutes.

The procedure for the incentive group differed from that described above in that either one star or five stars appeared on the screen two seconds before the problem. Half the problems were preceded by one star and half by five stars. The meaning of the point values was explained in the instructions and demonstrated by the first two problems.

Results and discussion

The participants provided meaningful solution words (rather than “--”, for example) for 97% of the problems, and their overall success rate was 46.7% (*SD* = 16.1). Figure 4 presents the results. There were no significant differences between the control and incentive groups when taking into account both incentives together. A within-participant comparison of response time between the 1-point and 5-point problems revealed a significant difference, with participants investing less time in the 1-point problems (*M* = 34.8 sec.; *SD* = 11.0) than in the 5-point problems (*M* = 40.2 sec.; *SD* = 13.6), *t*(39) = 2.34, *p* < .05, Cohen’s *d* = 0.37. This effect size represents a small-to-medium effect according to the accepted guidelines, where 0.20, 0.50, and 0.80 represent small, medium, and large effects, respectively (Cohen, 1988; Fritz, Morris, & Richler, 2012). As expected, confidence was lower for the 1-point than the 5-point problems, *t*(39) = 2.16, *p* < .05, *d* = 0.34. The means are represented by the ends of the thick continuous line in Figure 4. A mixed linear regression revealed that response time predicted confidence reliably for all incentive levels. The slopes for the incentives after controlling for accuracy were *b* = -0.30, *t*(593) = 11.62, *p* < .0001, R^2 = 40% for no incentive, *b* = -0.40, *t*(583) = 11.97, *p* < .0001, R^2 = 39% for 1 point, and *b* = -0.29, *t*(583) = 10.34, *p* < .0001, R^2 = 36% for 5 points. Up to this point, the results replicate the findings of Koriat et al. (2006).

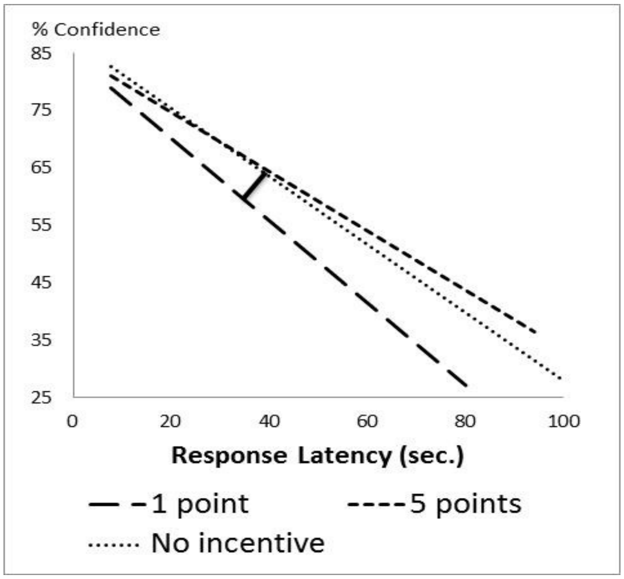

Figure 4. Experiment 2: Regression lines representing confidence predictions by response time for the no-incentive, 1-point, and 5-point conditions. The start and end points of the lines represent the response latency means for the 20% of solutions provided most quickly and the 20% provided most slowly per participant. The thick continuous line connects the mean times and confidence ratings for the 1-point and 5-point conditions.

The predictions guided by the DCM were supported as well. First, the conditions did not differ at their origin, all *t*s < 1. This finding is consistent with the prediction that solutions generated quickly are reported if they are accompanied by high confidence levels, and this satisficing level was similar under all conditions. The difference in mean confidence between the incentive levels stems, then, from problems which took a relatively long time to answer. Second, the slopes of the 1-point and 5-point conditions differed, as indicated by the interactive effect found in the regression, $t(1154) = 3.34$, $p < .001$, $R^2 = 2\%$.[2] This result suggests that the strong effect of latency could be affected, albeit slightly, by the incentive manipulation. This is consistent with the numerically small, though significant, confidence and latency differences between the incentive levels (see Figure 4).

A comparison to the no-incentive condition revealed a significant difference between the slopes for the 1-point condition and the no-incentive condition, $t(1167) = 3.10$, $p < .005$, $R^2 = 1\%$, and no difference between the slopes for the 5-point condition and the no-incentive condition, $t < 1$. These findings suggest that when presented with items that took longer to solve, participants reduced their criterion for the 1-point items rather than increasing their criterion for the 5-point items.

In sum, this experiment generalized Koriat et al.'s (2006, Experiment 7) results by using a different type of problem. The common findings across the task types support the DCM as a general model. This experiment also extended Koriat et al.'s (2006) findings by using a different data analysis approach and interpreting the findings in light of the DCM. The findings suggest that participants were more willing to compromise on their stopping

[2] Throughout the paper, the reported effect size for interactive effects was calculated by deducting the variance explained by latency from the variance explained by the full model that included the manipulation interaction with latency. Accuracy was partialled out in both models.

criterion for the low-incentive items than for the high-incentive items, and this pattern became stronger as it took longer to respond.

Experiment 3

A low-incentive condition, such as that used in Experiment 2, is one way to lower the stopping criterion. Another way to lower it is by making participants work under time pressure (Kruglanski et al., 2012; Thiede & Dunlosky, 1999). The prediction arising from the DCM is that time pressure will increase the slope of the confidence criterion relative to an ample time condition, because people are expected to compromise more quickly on their confidence level when they are motivated to move quickly to the next problem in the set.

The time pressure was manipulated between participants. This contributes to the generalizability of the findings beyond Experiment 2, because, as mentioned above, metacognitive judgments were found to be more sensitive to variability within the list of items than to variability in between-participant manipulations, or even to variability between blocks (Koriat et al., in press; Koriat et al., 2004; Yue et al., 2013). Sensitivity to variability within the list emphasizes the experience associated with each item relative to the experience of processing adjacent items. This experience-based judgment regarding each particular item accords with Koriat et al.'s (2006) data-driven explanation. However, if the same pattern of results is generalized to a between-participants manipulation, this will support the notion that the metacognitive judgment reflects top-down strategic regulation in light of global goal-setting which stems from task characteristics common across items, and that it is this top-down regulation which underlies the slope difference between conditions.

Beyond the slope differences in light of motivational and time constraints, another aim of Experiment 3 was to delve into participants' ongoing metacognitive monitoring while performing the task. The prediction derived from the DCM is based on findings that intermediate metacognitive judgments rise (either linearly, logarithmically, or with a sudden spike) during learning, reasoning, and problem solving, with the final judgments generally higher than the preceding ones (Ackerman & Goldsmith, 2011; Metcalfe & Wiebe, 1987; Thompson et al., 2011; Vernon & Usher, 2003). The novel prediction here is that despite this rise along the process, lengthy processing ends with a lower judgment than quick processing. This pattern was not expected to be affected by the global time frame.

Participants were asked to provide initial and intermediate confidence ratings for the solution options they considered while solving CRA problems. An illustration of such intermediate ratings is provided by cases B and C in Figure 2. The initial ratings were solicited shortly after the problem was presented, and the intermediate and final ratings followed. For simplicity of phrasing, in what follows the term "intermediate ratings" also encompasses initial ratings, unless initial ratings are explicitly specified. Final confidence ratings are distinguished from intermediate ratings throughout the paper.

Similarly to the feeling of rightness collected by Thompson and her colleagues (Thompson et al., 2011, 2013), the initial ratings allowed examination of their predictive value early in the attempt to solve the problem. However, unlike in Thompson et al.'s studies, the participants were not required to provide interim solutions, and were permitted to skip ratings. Thus, a finding that participants did indeed provide intermediate ratings would by itself—i.e., regardless of any predictive power of the ratings—suggest that the solving process can be interrupted voluntarily for progress assessment and then continued. In other words, interruptions to provide ratings without disturbing the time-confidence slope (and its effect size) would argue against fluency as a dominant cue for the final metacognitive judgments in multi-step processes involved in performing complex tasks.

Method

Participants. Forty-two participants drawn from the same population as in the previous experiments (M_{age} = 24.6; 44% females) were randomly assigned to the time conditions, time pressure (*N* = 20) and ample time (*N* = 22).

Materials. The problem set was the same as for Experiment 2.

Procedure. The procedure per item was the same as in Experiment 2, with one exception: intermediate confidence rating scales appeared on the screen while the participant solved each problem. The ends of the scale were marked “I still have no idea” and “I’ve got it.” The scale for the first rating appeared 3 seconds after presentation of the problem. A new scale then appeared below the previous one every 15 seconds, and the previous scale became inactive, even if no rating was entered. The screen could present up to 5 scales. The space for the answer was present in the lower part of the window during the entire process, such that participants could enter their answer, rate their final confidence, and move on to the next problem whenever they wanted. Times were documented when participants entered their intermediate and final confidence ratings.

The instructions for the ample time group stated that the session time the participants were invited for (30 min.) should allow them to work at ease and solve the entire problem set. The instructions for the time pressure group stated that this experiment was similar to tests the students were used to, in which time was limited. It was explained that they were allowed 16 minutes to solve the entire problem set, which meant about half a minute per problem. This time frame was significantly shorter than the time used by the control (no-incentive) group in Experiment 2 (*M* = 41.7 sec., *SD* = 11.9, *t*(19) = 15.6, *p* < .0001, *d* = 0.98). The problems were numbered at the top of the screen, and participants knew the number of problems in the set, so they could track their progress through the problems. Both groups were told that they should manage their time and try to solve all problems before the time elapsed. For both time conditions, the experimenter announced the middle point of the time frame and one minute before the time elapsed. Participants were told to stop working once the time elapsed.

Results and discussion

The participants worked on a mean of 27.5 (*SD* = 2.5) problems under time pressure and 29.3 (*SD* = 1.4) problems under the ample time condition. As in Experiment 2, participants provided meaningful solution words for 97% of the problems they worked on under both time conditions. The overall success rate under the ample time condition (*M* = 48.0%, *SD* = 16.8) was similar to that in Experiment 2. Comparisons of time, confidence, and success rates between the ample time condition of the present experiment and the control condition in Experiment 2 revealed no significant differences, all *t*s < 1. Most importantly, no differences in origins or slopes were found in a time-confidence regression analysis (both *t*s < 1). These findings suggest, as expected, that interruptions did not affect the outcomes.

A between-participant comparison between the time conditions revealed, as expected, shorter response latencies in the time pressure condition (*M* = 28.9 sec.; *SD* = 5.2) than in the ample time condition (*M* = 41.6 sec.; *SD* = 12.6), *t*(40) = 4.23, *p* < .0001, *d* = 1.33. The result was a lower success rate (Time pressure: *M* = 36.7%, *SD* = 11.8) than under the ample time condition (reported in the previous paragraph), *t*(40) = 2.49, *p* < .05, *d*

= 0.79, and lower confidence ratings (Time pressure: *M* = 49.9; *SD* = 14.2; Ample time: *M* = 60.0; *SD* = 15.6), *t*(40) = 2.20, *p* < .05, *d* = 0.69.

The regression results are represented by the dashed lines in Figure 5. In line with the predictions of the DCM, this analysis revealed no difference between the groups in the lines' origin, *t* < 1, an inverse relationship between time and confidence for both groups, *b* = -.54, *t*(543) = 10.69, *p* < .0001, R^2 = 62% for time pressure and *b* = -.35, *t*(640) = 11.86, *p* < .0001, R^2 = 64% for ample time, and a significant slope difference between them, *t*(1167) = 4.46, *p* < .0001, R^2 = 2%. Thus, the effect of the time frame manipulation was highly similar to the effect of the within-participant incentive manipulation of Experiment 2.

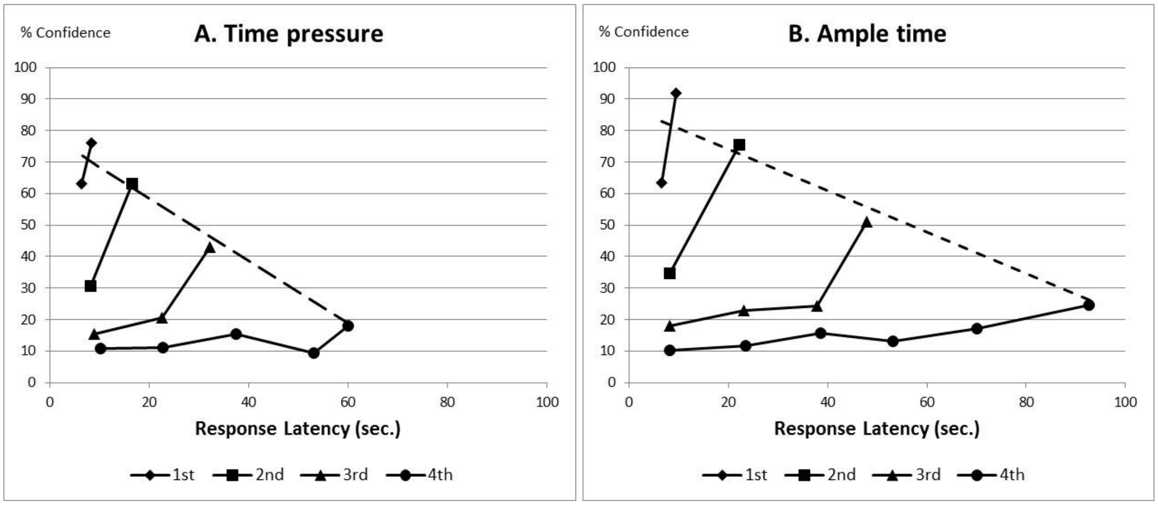

Figure 5. Experiment 3: The dashed lines represent the regression lines of predicting confidence by response time. The solid lines represent the intermediate and final confidence ratings for the four problem solving timelines, divided according to response latency quarters (1 to 4). The data are presented separately for the time pressure (A) and ample time (B) conditions.

To examine the progress of the ratings, the data were split for each participant into his/her own quarters of latencies, with 7-8 problems in each quarter, from the fastest provided solutions to the slowest ones. The mean success rates per quarter were 66.4, 50.0, 24.7, and 6.1 for quarters 1-4 of the time pressure condition, and 85.4, 62.7, 31.0, and 13.7, respectively, for the ample time condition. Figure 5 presents the intermediate and final confidence ratings for each quarter for each of the two time conditions. The standard error of the means for response times per quarter ranged from 0.3 for the 1st quarter of the time pressure condition to 6.2 for the 4th quarter of the ample time condition. The standard error of the means for the final confidence ranged from 2.5 for the 1st quarter of the ample time condition to 5.6 for the 1st quarter of the time pressure condition.

An important observation is that there was an almost perfect linear relationship between latency and final confidence even when using the means per latency quarter. The regression analysis used throughout this study necessarily results in linear predictive lines. This would be the case even for data that is not linear in nature (e.g., data forming a logarithmic curve). The breakdown by latency quarters allows examination of whether the linear regression lines are representative. The proximity between the regression lines and the final confidence means per quarter suggests that linear regressions reliably describe the relationship between latency and confidence.

Initial ratings were provided for 59% of the solutions. The following intermediate confidence ratings were provided for 40%, 27%, 19%, and 13% of the second, third, fourth, and fifth rating scales, respectively. Although the literature is somewhat vague regarding the definition of insight problems, CRAs (or the similar RAT problems) are often accounted as such problems (e.g., Kounios et al., 2006). One characteristic of insight problems is low ongoing confidence for some time followed by a spike upward (Metcalfe & Wiebe, 1987). This pattern is clearly evident in Figure 5. However, previous analyses did not include a detailed time analysis and/or a 0-100% confidence rating scale. Here, the analyses exposed diminishing confidence at the end of the solving process despite this spike pattern.

To examine the general pattern of a rise in confidence from initial to final ratings, a mixed three-way Analysis of Variance (ANOVA) for Time Condition (Pressured vs. Ample) × Quarter (1-4) × Rating (Initial vs. Final) was used. The analysis was based on participants who provided initial ratings under all four quarters (*N* = 28, 67%). The final confidence rating was necessarily provided by all participants. The main effect of the time condition was not significant, *F*(1, 26) = 1.10, *MSE* = 1465.57, *p* > .30, η_p^2 = .04. The main effect of the quarter was significant, *F*(3, 78) = 103.37, *MSE* = 332.98, *p* < .0001, η_p^2 = .80, reflecting that the ratings fell from the first to the fourth quarters. The main effect of the rating was also significant, *F*(1, 26) = 59.50, *MSE* = 617.16, *p* < .0001, η_p^2 = .70, reflecting the increase from the initial to the final confidence ratings. The Quarter × Rating interaction across the two time conditions was significant, *F*(3, 81) = 9.87, *MSE* = 232.60, *p* < .0001, η_p^2 = .27. This interaction can be seen in Figure 5 by the differential increase from the first to the last confidence rating for the four quarters, but all were significant, *p*s <= .001. Importantly, the triple interaction was insignificant, *F* < 1, suggesting a similar pattern of results for both time conditions.

An additional examination using the entire data set was conducted to examine the pattern of a rise in ratings leading to a decision to stop investing effort in the problem. Each problem-solving process was marked as “fit” if the last intermediate rating was the highest among the intermediate ratings and the final confidence rating was still higher, and “unfit” otherwise. The percentage of solutions that fit this criterion averaged 77% (*SD* = 17) for the time pressure condition and 76% (*SD* = 15) for the ample time condition, with no difference between them, *t* < 1. This finding suggests that in most cases, confidence ratings rose consistently during the solving process. Nevertheless, about a quarter of the solving processes showed swerves in the intermediate ratings. This finding is discussed further under Experiment 4, below.

Overall, the results are similar for time pressure manipulated between participants and for the incentives manipulated within participants in Experiment 2. This provides further support for the goal-directed nature of time investment, because the effect was found even without manipulation variability among items, and despite interruptions for providing the intermediate ratings. The goal-directed nature of the process is also supported by the positive relationship between time and confidence ratings during the solving process, together with the negative relationship between time and final confidence ratings in both time conditions. These results are clearly consistent with the DCM (see Figure 2): They accord with the notion that participants progress in a goal-driven manner to achieve a satisfactory confidence level, which diminishes as processing lengthens. Experiment 4 was designed to challenge the diminishing pattern of the stopping criterion by considering an alternative explanation for the findings.

Experiment 4

According to the DCM, people provide answers which satisfy their confidence criterion, and this criterion diminishes as more time is invested in a particular task item. However, what happens when people decide to give up on a problem—i.e., to abandon the effort to solve it? If the diminishing confidence criterion were the only stopping rule, respondents would be expected to continue to invest effort until their confidence reached a level judged as satisfactory for that time point—even if that level of confidence was zero. However, another stopping rule that may take effect in this case is a limit to the effort one is willing to invest in a particular item. Illustrating this idea, Kruglanski et al. (2012) offered an analogy between effort investment in cognitive tasks and physical forces. They suggested the term "potential driving force" to represent the maximal amount of energy an individual is prepared to invest in a cognitive task. This theorizing offers an alternative explanation for the low-confidence responses found in the previous experiments of the present study. By this explanation, it is not that low-confidence solutions became satisfactory after a certain length of time, but, rather, that they reflect cases in which participants reached the limit of the effort they were willing to invest in a problem, even though their confidence did not meet their criterion. Because they gave up on the problem but had to provide a response, they provided their best solution despite being unsatisfied with it, and moved on to the next item. This possibility suggests that the limit on effort overruled the aim of finding a satisfactory solution. Thus, it may be that the effort limit concealed a straight, non-diminishing, confidence criterion, while the found slope stemmed from the lengthy but non-satisfactory solving processes.

In order to examine this alternative explanation, Experiment 4 employed a free-report procedure, rather than the forced-report procedure used in the previous experiments. Under free report, participants are given the option of responding to items with "don't know." Previous studies in the context of knowledge questions have found that people utilize a free-report format to adhere their confidence criterion, by waiving their low-confidence responses and providing those in which their confidence is satisfactory. This was found to be the case for general knowledge questions (Ackerman & Goldsmith, 2008;

Koriat & Goldsmith, 1996), in educational settings (e.g., Krebs & Roebers, 2012), and in eyewitness contexts (e.g., J. R. Evans & Fisher, 2011). Similarly, setting a global low-goal, low-reward, or time-pressure condition to learning tasks led people to waive difficult items, thereby allowing more time for items with a better chance for success (Ariel et al., 2009; Metcalfe & Kornell, 2003; Reader & Payne, 2007; Thiede & Dunlosky, 1999). These findings clearly involve strategic, goal-driven regulation that is guided by considerations related to performance on the entire set of items, beyond the considerations that guide work on each particular item in isolation.

The question is what the confidence-based stopping criterion looks like under a free-report format, when taking into account response time as a predictive factor. If the confidence criterion does not diminish with time, and the main stopping rule is the time limit per item, the free-report procedure should expose it. This is because now participants can provide only solutions which satisfy their straight confidence criterion. On the other hand, if participants provide low-confidence solutions after lengthy thinking even when they can avoid them, this will support the diminishing nature of the stopping criterion, as it will suggest that they find these low-confidence solutions appropriate to provide even when they can withhold unsatisfactory solutions.

Beyond these two possibilities—a straight confidence criterion with a time limit or a diminishing criterion without an additional time-based stopping rule—a third and more likely possibility is a combined diminishing criterion and time limit rule (as illustrated in Figure 2). If this is indeed the case, a free-report format allows people to obey their time limit and move on to the next item without supplying an answer (case D in Figure 2) unless the confidence criterion is met prior to the time limit (case C in Figure 2). This explanation leads to the predictions that the time-confidence slope will remain negative, and that "don't know" responses will be provided after a longer time than solution responses. In addition, this should result in higher confidence for the solutions provided following the longest solving time than are seen under a forced-report procedure. This leads to the prediction of attenuated, though still negative, time-confidence slopes under a free-report relative to a forced-report procedure.

On top of this, there is the consideration of the global time frame allowed for the entire task. If a time limit per item influences respondents' behavior under an ample time condition, all the more should it do so under time pressure. To address these questions, Experiment 4 included time pressure and ample time conditions, as in Experiment 3, but using a free-report rather than a forced-report procedure. Taken together, the considerations raised here should result in a negative time-confidence correlation as before, but with a smaller effect of the time frame on confidence relative to its effect under a forced-report procedure (Experiment 3), because participants can waive responses in which their confidence is insufficiently high when they reach their time limit per item.

Method

Participants. Forty-seven participants drawn from the same population as in the previous experiments (M_{age} = 25.3; 49% females) were randomly assigned to the time conditions, time pressure (N = 25) and ample time (N = 22).

Materials. The problem set was the same as in the previous experiments.

Procedure. The procedure was the same as in Experiment 3, except that here a free-report format was employed. That is, a "don't know" button appeared next to the space for the answer. Clicking this button inserted the words "don't know" into the answer entry space and disabled the confidence rating scale. Participants could then change their mind and type a solution word in the space, replacing “don’t know.” This reactivated the confidence rating scale.

Results and discussion

Most participants (93%) utilized the free-report procedure (i.e., selected “don’t know” as their final answer at least once). Altogether, participants responded “don’t know” to 25% of the problems under both time conditions. Of the solutions provided, 55.4% were correct (*SD* = 21.2). This success rate was higher than in Experiment 3, *F*(1, 85) = 11.04, *MSE* = 343.28, *p* = .001, η_p^2 = .12, although the number of correct solutions did not differ between them, *t* < 1. In line with the prediction that the free-report format allows respondents to waive their lengthiest responses, the time invested in each problem (including both solutions and “don’t know” responses) was shorter here than in Experiment 3, *F*(1, 85) = 5.51, *MSE* = 107.73, *p* < .05, η_p^2 = .06, with no interactive effect with the time condition, *F* < 1. This finding combined with the higher success rate suggests that the free-report format allows more efficient work.

Delving into the comparison between “don’t know” answers and actual solution responses yielded that, as predicted, the response time for the “don’t know” responses was significantly longer than for the solution words for both the ample time (“don’t know”: *M* = 56.9 sec.; *SD* = 26.6; solutions: *M* = 29.5 sec.; *SD* = 13.6, *t*(18) = 5.90, *p* < .0001, *d* = 1.35) and time pressure conditions (“don’t know”: *M* = 34.5 sec.; *SD* = 10.7; solutions: *M* = 22.6 sec.; *SD* = 5.6, *t*(24) = 5.01, *p* < .0001, *d* = 1.02). This finding supports the prediction that participants resorted to “don’t know” only when reaching their own time limit per item after a lengthy attempt to solve the problem, rather than skipping problems quickly.

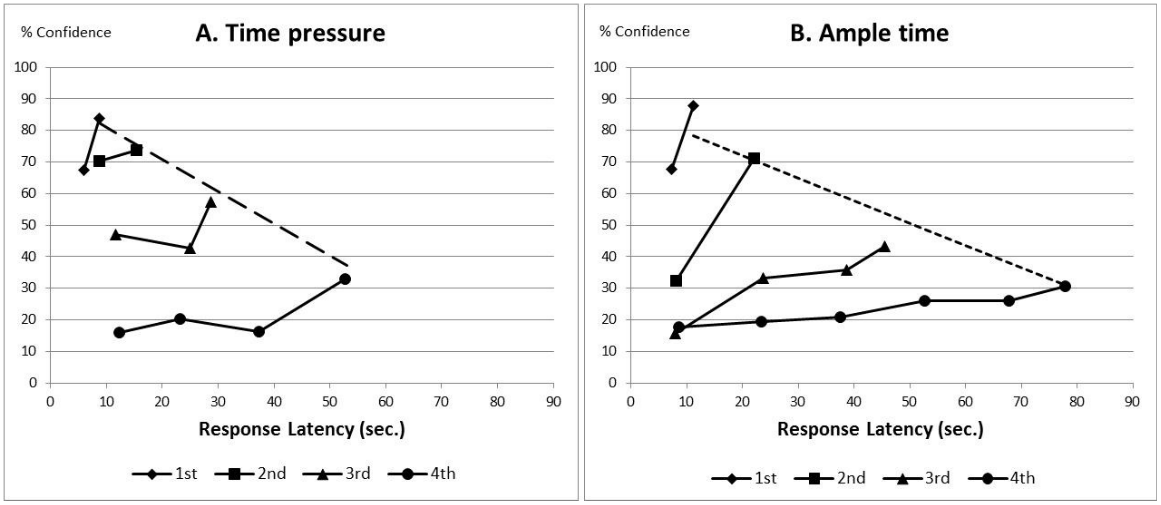

Figure 6. Experiment 4: The dashed lines represent the regression lines for confidence predicted by response time with a "don't know" option. The continuous lines represent the intermediate and final confidence ratings for the four problem-solving timelines, divided according to response latency quarters (1 to 4). The data are presented separately for the time pressure (A) and ample time (B) conditions.

The results of the analyses based on the provided solutions are presented in Figure 6. A between-participant comparison revealed shorter latencies for the time pressure (*M* = 22.7 sec.; *SD* = 5.6) than for the ample time condition (*M* = 29.0 sec.; *SD* = 13.1), *t*(45) = 2.19, *p* < .05, *d* = 0.38. No difference in confidence was found between the conditions in the presence of the "don't know" option (Time pressure: *M* = 68.4; *SD* = 14.1; Ample time: *M* = 65.7; *SD* = 13.5), *t* < 1. Indeed, a comparison between the two experiments in a two-way ANOVA on confidence ratings, Experiment (3 vs. 4) × Time Condition (Time pressure vs. Ample time), yielded a main effect of the experiment, with higher confidence ratings with the "don't know" option than without it, *F*(1, 85) = 15.64, *MSE* = 206.26, *p* < .0001, η_p^2 = .16, and a significant interactive effect, *F*(1, 85) = 4.43, *MSE* = 206.26, *p* <

.05, η_p^2 = .05. Confidence was lower under time pressure only when the participants had to provide a solution for every problem.

Nevertheless, as in Experiment 3, the regression revealed no difference in the groups at the origins, t = 1, with inverse relationships between time and confidence for both groups, b = -0.54, t(547) = 9.18, p < .0001, R^2 = 58% for time pressure and b = -.46, t(493) = 10.69, p < .0001, R^2 = 54% for ample time, and with a significant slope difference between them, t(1040) = 2.02, p < .05, R^2 = 1%. The slope difference, with no reduction in mean confidence, supports the predicted combined effect of a time limit and a diminishing confidence criterion. This combination suggests that the participants worked faster without compromising more on their confidence under the time pressure condition, as they used the “don’t know” response as a preferred regulatory tool which was not available in Experiment 3. An analysis across both experiments revealed, as expected, that the free-report format attenuated the slopes, t(2574) = 3.56, p < .0005, R^2 = 1%. However, there was no triple interaction, reflecting similarity in the overall pattern of results regardless of the report procedure (free vs. forced).

Dividing the provided solutions into response time quarters per participant, as in Experiment 3, provided the data represented by the continuous lines in Figure 6. In this experiment, only 14 participants (30%) provided initial ratings for all four quarters. Nevertheless, a three-way ANOVA comparing these initial ratings to the final confidence ratings replicated the findings of Experiment 3. The main effect of the time condition was not significant, F(1, 12) = 2.84, MSE = 1469.02, p = .12, η_p^2 = .19, while the main effects of the quarter and of the rating were significant, F(3, 36) = 39.57, MSE = 382.05, p < .0001, η_p^2 = .77, and F(1, 12) = 40.12, MSE = 225.12, p < .0001, η_p^2 = .77, respectively. As in

Experiment 3, there was a significant interactive effect for Quarter × Rating, $F(3, 39) = 2.92$, $MSE = 212.36$, $p < .05$, $\eta_p^2 = .18$. The rise from the first to the last confidence ratings was significant for all quarters, *p*s < .005, except the first (highest confidence ratings), $p = .16$. Thus, the ratings were lower for the problems that took longer to solve and generally rose from the first to the final ratings.

Finally, the complementary analysis performed in Experiment 3 was again conducted—i.e., examining the overall pattern of a rise in confidence ratings leading up to a decision to provide a response. Fit rates with the free-report format were high (Time pressure: $M = 94\%$, $SD = 10.8$; Ample time: $M = 88\%$, $SD = 16.1$; $t(45) = 1.54$, $p = .13$)—higher than in Experiment 3, $F(1, 85) = 22.11$, $MSE = 0.02$, $p < .0001$, $\eta_p^2 = .21$, with no significant difference between the time conditions or interactive effect, both *F*s ≤ 1. Thus, under the free-report format, participants primarily provided responses characterized by a gain in confidence as more time was invested.

This experiment replicated the overall pattern of results found in Experiment 3. The free-report procedure certainly did not eliminate the time-confidence slopes: Solutions provided after lengthy consideration were accompanied by low confidence, despite the opportunity to avoid providing low-confidence solutions. As expected, the criterion level rose slightly, though significantly, relative to the forced report in Experiment 3. These findings offer decisive evidence that participants' low confidence in solutions reached after lengthy deliberation nonetheless satisfied their criterion for providing an answer.

Together, the findings of Experiment 3 and Experiment 4 suggest that when participants are forced to provide a solution for every problem, they take into consideration both the diminishing confidence criterion and a time limit per item. But when they are free to respond or not at their own discretion, they avoid providing solutions which remain below their stopping criterion when they reach their time limit.

Experiment 5

The purpose of Experiment 5 was to generalize the findings to another task and to directly compare free- with forced-report conditions. Misleading problems are commonly used in the literature related to dual-process theories as a means of differentiating between fast intuitive (System 1 or Type 1) solutions and the results of more deliberate processing (System 2 or Type 2). For example: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost? ____cents" (Kahneman, 2003). The immediate solution that comes to mind is 10 cents, while the correct solution is 5 cents. These problems have high external validity as they are not confined to three words; they represent various types of problems, decisions, and judgments engaged in by people in daily life (Kahneman, 2003); and they are commonly used in educational contexts such as schools, higher education, and screening tests (e.g., the Graduate Management Admission Test, or GMAT). From a theoretical point of view, these problems allow ecologically valid dissociation between confidence and accuracy in their relationship with response time. This is because the very first solutions considered tend to be accompanied by high confidence but a low chance of being correct, in particular when presented in an open-ended format (Ackerman & Zalmanov, 2012).

As explained in the introduction, Ackerman and Zalmanov (2012) found a consistent negative correlation between time and confidence even with these problems. This might indicate that even with these problems people regard relatively low confidence levels as satisfactory after lengthy thinking, as suggested here. However, as explained in the introduction to Experiment 4, the negative correlation could also stem from the requirement to provide a solution for every problem. In light of this possibility, Experiment 5 examined whether the free-report format eliminates the negative correlation between latency and confidence with a set of misleading problems.

Method

Participants. Fifty-one participants were drawn from the same population as in the other experiments (M_{age} = 24.3; 57% females) and randomly assigned to answering under free-report (N = 27) or forced-report conditions (N = 24).

Materials. The misleading problems used by Ackerman and Zalmanov (2012, Experiment 1) were used for this experiment. They included 12 experimental math problems and a practice problem for demonstrating the procedure. The experimental problems included the three problems used by Frederick (2005; the bat and ball, water lilies cover half a lake, and machines that produce widgets at a certain rate), the drinks version of Wason's selection task (Beaman, 2002), the A-is-looking-at-B problem (Stanovich, 2009), and a conditional probability problem (Casscells, Schoenberger, & Graboys, 1978). The other problems were adapted from personal communications with researchers who use such problems and from preparation booklets for the GMAT. The full list of problems appears in Table S1 in the online supplemental materials.

Procedure. The procedure was similar to that used in the time pressure conditions of Experiment 3 and Experiment 4. The practice problem appeared first, and the rest were randomly ordered for each participant. The participants had 12 minutes to solve the twelve problems. The time allotment of about 1 minute per problem was less than the average time expended per problem (*M* = 70.3 sec., *SD* = 18.3) in the free-entry condition of Ackerman and Zalmanov (2012), but still allowed working on the entire problem set.

In a pilot study (*N* = 20)[3] with the same ample time procedure used for the previous experiments, only 1% of the responses were "don't know." In another pilot study (*N* = 20) with time pressure, 3.5% were "don't know" responses. In the present study, time pressure was used and a sentence was added to the instructions: "Imagine a situation in which you will accumulate points by solving the problems correctly, such that you will receive 5 points for each correct answer and lose 5 points for each incorrect solution." No further incentive was offered.

Results and discussion

The participants provided meaningful responses in all cases and succeeded in providing solutions (correct or wrong) to 11.8 (*SD* = 0.55) out of the 12 problems in the allotted time. The success rate with the forced-report condition was 37.0% (*SD* = 18.2), which was lower than the success rate (45%) found by Ackerman and Zalmanov (2012) in the open-ended test format without intermediate confidence ratings and without time pressure. In the free-report group, a "don't know" answer was provided for 6.6% of the problems and was used by 14 (52%) of the participants. As in Experiment 4, the "don't know" responses followed longer elapsed time (M = 68.6 sec., *SD* = 32.0) than solution responses (*M* = 50.5 sec., *SD* = 10.1). Probably because of the large variability, this difference was marginal, $t(13) = 1.89$, $p = .081$, $d = 0.51$, although the effect size was medium. The presence of the "don't know" option did not affect the success rate, which was 35.1% (*SD* = 17.3), $t < 1$. Confidence and response latency also did not differ between

[3] The results of this pilot study can be seen in Ackerman (2013, Experiment 2).

the groups. Confidence was 79.8 (*SD* = 11.0) for the free-report and 78.3 (*SD* = 10.9) for the forced-report group, *t* < 1. Response latency was 50.3 (*SD* = 8.5) for the free-report and 47.5 (*SD* = 7.8) for the forced-report group, *t*(49) = 1.19, *p* = .24, *d* = 0.35.

The regression analysis of the confidence slopes revealed no difference between the conditions in the origins, *t* = 1, significant slopes for both conditions, *t*(292) = 5.22, *p* < .0001, R^2 = 22% for the free-report and *t*(272) = 6.98, *p* < .0001, R^2 = 29% for the forced-report condition, and a significant slope difference between them, *t*(563) = 2.37, *p* = .018, R^2 = 1%. As found in Experiment 4, although the free-report format attenuated the slope, it did not eliminate it, and a decline in confidence was observed chiefly for the solutions with greater latencies.

Because there were only 12 problems in the set, in examining the intermediate and final confidence ratings the latency data for each participant were split into thirds, so that each line would represent four solutions. Seven participants from the free-report group and five from the forced-report group did not provide intermediate ratings at all. Figure 7 presents the intermediate and final confidence ratings, as in Figure 5 and Figure 6. A difference between the CRAs and the present problems is evident from the figure, in that the pattern for the CRAs was of little progress in the intermediate ratings followed by a sudden spike upwards, while in the present problems the ratings rose more gradually.

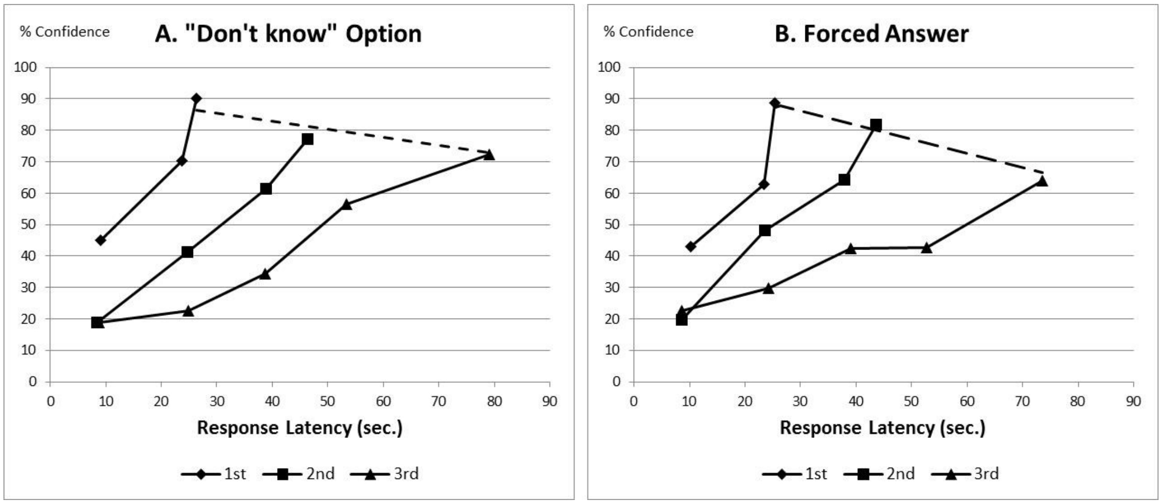

Figure 7. Experiment 5: The dashed lines represent the regression lines for confidence predicted by response time with (panel A) and without (panel B) a “don’t know” option. The continuous lines represent the intermediate and final confidence ratings for the three problem-solving timelines, divided according to response latency thirds (1 to 3).

It is evident in Figure 7 that the confidence levels were all relatively high, even after lengthy thinking. This may explain the low use of the “don’t know” response. In addition, this observation may suggest that all answers were provided above a non-diminishing criterion. However, a deeper look at Figure 7, and in particular at the points plotted at around 70 in Panel A and 65 in Panel B—i.e., the final confidence levels for the problems with the longest latencies—reveals that those confidence levels were not considered satisfactory for solution options found relatively quickly, but were satisfactory enough to warrant providing an answer for problems that took longer to solve, in accordance with the DCM.

In this experiment, 37 participants (73%) provided initial confidence ratings for all three thirds. A three-way ANOVA comparing the initial ratings to the final ratings was performed, as above. The main effect of the report procedure (free vs. forced) was not significant, $F < 1$, while the main effects of the third and of the rating were significant, $F(2, 70) = 27.61$, $MSE = 291.20$, $p < .0001$, $\eta_p^2 = .44$ and $F(1, 35) = 178.51$, $MSE = 846.31$, $p < .0001$, $\eta_p^2 = .84$, respectively. There were no significant interactive effects with the report procedure. There was a Third × Rating interactive effect, $F(2, 72) = 3.06$, $MSE = 242.83$, $p = .049$, $\eta_p^2 = .08$, which stemmed from the differential rises from the initial to the final confidence in the three thirds, but all differences were significant, *p*s < .0001 (see Figure 7, both panels). Examination of the rise in confidence ratings during the solving process revealed overall fit rates of 81% (*SD* = 18.0) with free report and 79% (*SD* = 17.7) with forced report, with no significant difference, $t < 1$, between the conditions.

Experiment 5 provides both generalization and a direct comparison between free- and forced-report conditions. Although, overall, confidence was higher in this experiment than in the CRA task, the pattern of results predicted by the DCM remained, even under the free-report procedure.

General Discussion

The present study was motivated by the puzzling inconsistency between theories and empirical findings across a large variety of cognitive tasks. While the theories lead to the prediction that more effort will increase the perceived likelihood of success, in fact, as people invest longer in a task this likelihood falls. As reviewed in the introduction, this is the case with many memorization, question-answering, problem-solving, reasoning, and decision-making tasks. Koriat et al. (2006) provided an enlightening explanation for this inconsistency by differentiating between data-driven and goal-driven effort investment. In particular, they demonstrated opposite effects for difficulty variability and motivation variability on the relationship between time investment and JOL (in a word-pair memorization task; see also Koriat et al., in press) and confidence (in a problem-solving task). Their explanation suggests that the overall pattern of negative time-judgment correlations stems from bottom-up data-driven effects dominating top-down goal-driven effects. This is consistent with the widely accepted explanation that suggests utilization of processing fluency as a central cue for numerous judgments (Bjork, Dunlosky, & Kornell, 2013; Unkelbach & Greifeneder, 2013). The present study challenges the scope of the data-driven, fluency-based explanation as the dominant basis for the negative time-judgment correlations which are so commonly found.

Boundary conditions for fluency effects on judgments were demonstrated in Experiment 1 by using a particularly large range of item difficulty levels in memorization of concept-category pairs and in CRA problems. Although both tasks showed negative time-judgment correlations and similar success rates and variability, the results revealed clear differences between them (see Table 1 and Figure 3). The memorization task, which showed adequate predictive power for time in predicting JOL over all, resulted in low predictive power when the difficulty variability was reduced by dividing the items into difficulty levels. When considering each difficulty level separately, the predictive power of time weakened, and the correlation even turned into positive as the items became more challenging. For problem solving, in contrast, there were highly predictive negative time-confidence slopes, regardless of the items' difficulty and variability. The mere finding that the predictive power of fluency depends on a combination of task type, difficulty variability, and item difficulty has general importance in challenging fluency as the main source for the commonly found negative correlations between time and metacognitive judgments.

This differential effect of fluency on the two task types, memorization and problem solving, is suggested here to stem from a difference between them in the dominance of data-driven versus goal-driven effort investment. By this explanation, the memorization task is dominated by data-driven effort, and thus JOL is strongly affected by processing fluency. However, only when the stimuli meet a minimal level of processing fluency and there is enough variability among items can fluency take its effect. When these criteria are not met, fluency effects on JOL weaken dramatically and even disappear. Problem solving, in contrast, is dominated by goal-driven effort investment. The sensitivity of confidence regarding problem solutions to time results mainly from top-down regulation of effort, with bottom-up data-driven fluency being only a minor contributor. This goal-driven effort ends in a negative time-judgment correlation that cannot be explained only by fluency and is explained by the DCM.

The prediction of the difference between the tasks in their sensitivity to fluency was based on the distinction between simple and complex tasks. In particular, Funke (2010) demonstrated such a distinction by comparing memorization and problem solving. This distinction is clearly vague, but key characteristics in many definitions of complex cognitive tasks are multi-step and goal-directed effort investment (see Funke, 2010; Quesada, Kintsch, & Gomez, 2005). These features accord with the DCM and characterize many tasks that show consistent negative time-judgment correlations. However, this distinction is certainly not sufficient as an inclusion criterion for the DCM. For example, reading comprehension is clearly a multi-step and goal-directed task that is also intuitively perceived as a complex cognitive task. Still, as mentioned in the introduction, studies of reading comprehension have found no correlation between time and metacomprehension judgments. This could stem from methodological factors, such as an insufficient number of items per participant (typically about 6) or insufficient time variability among items. However, it may also hint at theoretically important differences in regulation of effort between task types, and leaves many open questions for future research. For example: What are the factors differentiating between complex cognitive tasks that show and those that do not show negative time-judgment correlations? Are there complex learning tasks that show a negative correlation? Are there question-answering, problem-solving, reasoning, or decision-making tasks that do not show a negative correlation? Under what conditions? How can we better define the difference between tasks that show negative time-judgment correlations because of data-driven dominance and those for which there is goal-driven dominance? Does this highlight a substantial difference between meta-memory and meta-reasoning processes (see Ackerman & Thompson, in press)?

The proposed DCM is based on several theories—the discrepancy reduction model, goal-driven versus data-driven investment, and the dual-process theory—all of which suggest that the performance of various cognitive tasks progresses in a goal-driven manner (e.g., J. St. B. T. Evans, 2006; Koriat et al., 2006; Nelson & Narens, 1990), with a positive correlation between time and assessed likelihood of success. Collecting intermediate confidence ratings in addition to the final confidence level (Experiment 3, Experiment 4, and Experiment 5) made it possible to delineate changes in this assessed likelihood of success during the solving process, and provided empirical support for a goal-driven rise in confidence during performance of the task.

The unique contribution of the proposed DCM is that the stopping criterion diminishes, reflecting people's increased willingness to compromise on the chance of success they consider satisfactory as they invest longer in a task. This was supported in Experiment 2 and Experiment 3 by the findings that low incentives (within participants) and time pressure (between participants), which are known to reduce the stopping criterion (e.g., Koriat & Goldsmith, 1996; Thiede & Dunlosky, 1999), do not in fact take effect immediately, but only once some effort has been invested in the task. The comparison between forced- and free-report conditions with and without time pressure (Experiment 3, Experiment 4, and Experiment 5) provided assurance that this was not an artifact of a self-set time limit per item. The results suggest that under forced report, both the diminishing criterion and a time limit were jointly used as stopping rules, such that respondents provided a response as soon as their current solution satisfied the diminishing criterion or they reached their own time limit, whichever came first. Under free report, respondents who reached their time limit before arriving at a satisfactory solution could extricate themselves from the situation by providing a “don’t know” response. However, the negative time-confidence slope showed the pattern of a diminishing stopping criterion despite the opportunity to avoid low-confidence responses. Interestingly, time pressure affected the final confidence level only when the participants had to provide a response for each item.

Overall, the evidence clearly supports the DCM by showing upwards progress of confidence during performance of the task together with an ultimate negative time-confidence correlation. Importantly, high motivation for success, ample time allotment, and use of a free-report procedure did not eliminate this negative correlation.

It is important to highlight that the DCM’s explanation for the negative correlation between time and final metacognitive judgments—namely, a willingness to compromise on the likelihood of success—does not disqualify the data-driven fluency-based explanation suggested by Koriat et al. (2006). In the present study, the participants were able to interrupt their thinking process to provide intermediate ratings, although this was not enforced and had no effect on the outcomes. This finding weighs against explanations associated with fluency as underlying negative time-judgment correlations (see also Mueller, Tauber, & Dunlosky, 2013). Nevertheless, there is no reason to categorically reject the data-driven explanation: each judgment may reflect a combination of data-driven and goal-driven effort, but the balance between them differs based on task characteristics. Looking back at Figure 1 and considering word-pair memorization, it is possible that for high-incentive items, people provide an initial judgment based on their data-driven experience with the item, and then manage their time investment by their goal-driven motivation generated by the high incentive. Support for this possibility can be found in a recent study by Koriat et al. (in press) with fifth- and sixth-grade children using well-known words as stimuli. They found that children at this age had trouble reflecting both data-driven and goal-driven effort in their JOLs when difficulty and incentive variability co-existed in the same list of paired associates. The same children provided judgments that appropriately reflected each factor—i.e., variation in difficulty or in importance—when they were isolated within the studied list. In their Experiment 5 (Phase 2), Koriat et al. (in press) presented paired associates at a variety of difficulty levels for self-paced learning without showing their point value first. When the child was done with the item, the point value appeared, and the child could then restudy the item in a self-paced manner. As expected, the children invested longer in restudying the items with high point values. Importantly, with this procedure, the children showed an adult-like pattern of JOLs, similar to that shown here in Figure 1. It is possible that adults spontaneously engage in multi-step processing even when memorizing paired associates. In support for this hypothesis, Ariel et al. (2009) examined with young adults whether selection of items for restudy is dominated by motivational considerations or by the difficulty of the items (i.e, whether their selection of items for restudy is data-driven or goal-driven; Koriat et al., 2006). They found that decisions to restudy paired associates were dominated by the importance of the items for success at test (goal-driven effort) rather than by their difficulty (data-driven effort). Future studies are encouraged to examine whether the negative correlations found with memorization tasks under high motivation for success also stem from goal-driven regulation towards a diminishing stopping criterion, as found here with problem-solving tasks. In this case, the difference between goal-driven meta-memory and meta-reasoning may be less fundamental than appears from the analysis above.

The revelation stemming from the present study is that, at least for complex tasks, each intermediate rating and the final judgment result from a balance between four factors: data-driven effort, goal-driven effort, a comparison against the diminishing criterion, and a comparison against a limit people set for the time they are willing to invest. Thus, the final confidence ratings primarily reflect the level of the criterion at that point in time and the time limit. In sum, while Koriat et al. interpreted the negative time-judgment correlation as stemming from data-driven dominance, by the DCM, it is goal-driven dominance that leads to this result.

This analysis sheds new light on the aforementioned findings that in reasoning and problem-solving tasks the feeling of rightness for the very first answer that comes to mind is negatively correlated with the initial response time (Thompson et al., 2011, 2013) and that final confidence ratings are also negatively correlated with overall response time (present results; Ackerman & Zalmanov, 2012; Thompson et al., 2011, unpublished analysis). The reconciling explanation suggested above corresponds to this case as well: The initial feeling of rightness, like the initial data-driven JOL provided by the children in Koriat et al. (in press), is based on data-driven effort dominated by fluency, while the final confidence rating is determined by the diminishing criterion.

Considering this possibility raises an intriguing implication regarding the dual-process approach and the role of metacognitive processes in it. It seems to follow that data-driven, bottom-up processes underlie the metacognitive processes that guide T1 and the activation of T2 reasoning, as suggested by Thompson et al. (2011, 2013), while goal-driven, top-down metacognitive processes underlie the stopping rule for T2 reasoning, as suggested here. This possibility cannot be examined with the data collected for the present study because the confidence scales appeared at predefined intervals, and many participants responded soon after the appearance of each scale. For example, in Figure 5a the response latencies of the first three judgments across the quarters show clearly that response times were highly similar for all quarters, as if they were externally fixed. This makes response times for the intermediate confidence ratings non-diagnostic of effort. It does not mean that fluency or effort were not used as cues by the participants (Koriat et al., 2006, Experiment 2), only that we, as researchers, cannot use response time as an indicator for these constructs.

A question of interest is whether the required time and the associated target confidence level are set by participants as a preliminary step (i.e., right after encountering the problem), or are updated dynamically after the solving process is initiated. The former, judging the solvability of a problem right after encountering it, is similar to the quick initial feeling of knowing that was found to guide a memory search in question answering (e.g., Reder & Ritter, 1992) and a quick decision to skip the most difficult items when memorizing under time pressure (Metcalfe & Kornell, 2005) (see Ackerman & Thompson, in press). Indeed, looking at Figure 5, Figure 6, and Figure 7, it is evident that respondents' initial confidence ratings have high predictive power for both time and final confidence. Similarly, a feeling of rightness regarding an initial answer option was found to be predictive of later deliberation time (Thompson et al., 2011, 2013). Also, several studies have showed that participants can distinguish quickly between solvable and non-solvable RAT problems (problems very similar to the CRAs used here; e.g., Bowers, Regehr, Balthazard, & Parker, 1990; Topolinski & Strack, 2009) and anagrams (Novick & Sherman, 2003).

Nevertheless, several findings suggest an absence of preliminary planning. First, unpublished data from a pen-and-paper pilot study show that when participants assume that all problems are solvable, they do not distinguish between those that are indeed solvable and those that are not.[4] Similarly, using unsolvable water-jar problems, Payne and Duggan (2011) found that participants only determine that a problem is unsolvable after a lengthy attempt to solve it, and sometimes not even then. Second, in the present study, the participants who worked under the free-report procedure could use their initial rating to

[4] In this pilot study, participants (*N* = 19) rated how many out of 100 of their peers would solve each of a set of CRA problems. The CRA list comprised 42 problems at various difficulty levels, including 4 unsolvable problems (three words with no commonly associated fourth word). The mean rating was 59.1 (*SD* = 10.0) for the solvable problems and 55.9 (*SD* = 15.3) for the unsolvable problems, with no significant difference between them, *t*(18) = 1.26, *p* = .23, *d* = 0.29.

improve their efficiency by quickly skipping problems where they recognized their chance of finding the correct solution would be low even with lengthy thinking. The finding of longer latency for “don’t know” responses than for substantial solutions under free report for both CRAs (Experiment 4) and misleading math problems (Experiment 5) points against this possibility. The improvement in efficiency found in Experiment 4 relative to Experiment 3 did not stem from quick acknowledgement of a low chance for success in particular items, but from giving them up after much effort. Moreover, time pressure did not increase the rate of “don’t know” responses at all (Experiment 4) or increased it only slightly (pilots for Experiment 5). Thus, the accumulated evidence suggests that ongoing adjustment to passing time with a diminishing stopping criterion is more plausible than preliminary setting of the stopping point.

Beyond support for the DCM, this analysis offers a novel interpretation for how the free-report format functions. As explained above, previous studies referred to free report as a way to increase success rates among provided responses (Ackerman & Goldsmith, 2008; J. R. Evans & Fisher, 2011; Koriat & Goldsmith, 1996; Krebs & Roebers, 2012). This is the case here as well, but the present study adds the time limit as an additional factor that is taken into account in the choice between a solution and a “don’t know” response. As found in Experiment 4, this use of the free report can help improve efficiency by both shortening response times and increasing success rates. Furthermore, a monotonic rise characterized the intermediate and final confidence ratings more consistently under free report (Experiment 4) than under forced report (Experiment 3). This finding provides yet another insight regarding utilization of the “don’t know” option—namely, allowing respondents a way out when their confidence drops during the task. Finally, the “don’t know” option was used only rarely for the misleading problems (Experiment 5), even when the incentive policy included a loss of points for wrong solutions that should have led to loss aversion (Tversky & Kahneman, 1991). This may be why some of the effects found with CRAs were not found with the misleading problems. Future studies are called for to improve our understanding of the factors that affect the willingness to admit failure.

Finally, the present study suggests the diminishing criterion as a metacognitive stopping rule for performing complex cognitive tasks, but other stopping rules for similar tasks have also been considered in the literature. For example, Payne and Duggan (2011) focused on the conditions that lead people to give up when facing unsolvable problems. They found that more time is invested in problems as the likelihood that the problem is actually solvable rises and as the number of problem states it allows increases. That study put the emphasis on characteristics of the problem. By looking into the processes involved in training for problem solving, Josephs, Silvera, and Giesler (1996) dealt with the effect of the subjective feeling of skill improvement on the stopping criterion people adopt. Dougherty and Harbison (2007) looked at individual differences in motivation which led participants to terminate attempts to free recall lists of 10 words—a task for which participants had no doubt that at least partial success could be achieved. Future studies are called for to integrate these and other stopping rules in cognitive tasks with the presently proposed DCM, which focuses on metacognitive regulatory processes and the time limit people set.

To conclude, metacognitive studies traditionally focus on memorization of very simple stimuli with well-known words. The present study evolved by considering highly challenging paired associates and problem-solving tasks, which are both understudied from the metacognitive point of view. Investigating more challenging tasks brings to the fore factors that were not previously considered despite having broad ecological validity. For instance, the principle of the diminishing criterion may help people interpret answers they receive to challenging problems presented to others (e.g., in expert consultations in medicine or law). The analysis of the free-report format with and without time pressure has implications for efficient answering (e.g., in educational exams). By proposing the DCM, this study aimed to shed new light on the processes that lead people to end up with low confidence in their chance of success, even when they can potentially avoid it by continuing improvement attempts or admitting failure.

References

Ackerman, R. (2013). A metacognitive stopping rule for problem solving. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 121-126). Austin, TX: Cognitive Science Society.

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting--With and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1224-1245.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18-32.

Ackerman, R., & Koriat, A. (2011). Response latency as a predictor of the accuracy of children's reports. *Journal of Experimental Psychology: Applied, 17*(4), 406-417.

Ackerman, R., Leiser, D., & Shpigelman, M. (2013). Is comprehension of problem solutions resistant to misleading heuristic cues? *Acta Psychologica, 143*(1), 105-112.

Ackerman, R., & Thompson, V. (in press). Meta-reasoning: What can we learn from meta-memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as memory*. Hove, UK: Psychology Press.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review, 19*(6), 1189-1192.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*(3), 432-447.

Beaman, C. P. (2002). Why are we good at detecting cheaters? A reply to Fodor. *Cognition, 83*(2), 215-220.

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.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, 35*(4), 634-639.

Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology, 22*(1), 72-110.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine, 299*(18), 999-1001.

Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Derryberry, D., & Rothbart, M. K. (1997). Reactive and effortful processes in the organization of temperament. *Development and Psychopathology, 9*(4), 633-652.

Dougherty, M. R., & Harbison, J. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(6), 1108-1117.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker (Ed.), *Metacognition in educational theory and practice* (pp. 249-275). Mahwah, NJ: Lawrence Erlbaum Associates 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.

Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology, 25*(3), 501-508.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378-395.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18.

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive processing, 11*(2), 133-142.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 5*(34), 1055-1075.

Gruppuso, V., Lindsay, D. S. , & Masson, M. E. J. (2007). I'd know that face anywhere! *Psychonomic Bulletin & Review, 14*(6), 1085-1089.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22-34.

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. doi: 10.1080/20445911.2011.518371

Josephs, R. A., Silvera, D. H., & Giesler, R. B. (1996). The learning curve as a metacognitive tool. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 22*(2), 510-524.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449-1475.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1-24.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (in press). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General*.

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24*(2), 169-182.

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.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36-68.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585-592.

Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. (2006). The prepared mind neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science, 17*(10), 882-890.

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning*, 1-16.

Kruglanski, A. W., Bélanger, J. J., Chen, X., Köpetz, C., Pierro, A., & Mannetti, L. (2012). The energetics of motivated cognition: A force-field analysis. *Psychological Review, 119*(1), 1-20. doi: 10.1037/a0025488

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*(4), 530-542.

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

Metcalfe, J., & Wiebe, D. (1987). Metacognition in insight and noninsight problem solving. *Memory & Cognition, 15*, 238-246.

Miron-Spektor, E., Efrat-Treister, D., Rafaeli, A., & Schwarz-Cohen, O. (2011). Others' anger makes people work harder not smarter: The effect of observing anger and sarcasm on creative and analytic thinking. *Journal of Applied Psychology, 96*(5), 1065-1075.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*(20), 378-384.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125-173). San Diego, CA: Academic Press.

Novick, L. R., & Sherman, S. J. (2003). On the nature of insight solutions: Evidence from skill differences in anagram solution. *The Quarterly Journal of Experimental Psychology, 56*(2), 351-382.

Payne, S. J., & Duggan, G. B. (2011). Giving up problem solving. *Memory & Cognition, 39*(5), 902-913.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864-901.

Probst, T. M., Stewart, S. M., Gruys, M. L., & Tierney, B. W. (2007). Productivity, counterproductivity and creativity: The ups and downs of job insecurity. *Journal of Occupational and Organizational Psychology, 80*(3), 479-497.

Quesada, J., Kintsch, W., & Gomez, E. (2005). Complex problem-solving: A field in search of a definition? *Theoretical Issues in Ergonomics Science, 6*(1), 5-33.

Reader, W. R., & Payne, S. J. (2007). Allocating time across multiple texts: Sampling and satisficing. *Human-Computer Interaction, 22*(3), 263-298.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451.

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.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*(3), 416-425.

Sandkühler, S., & Bhattacharya, J. (2008). Deconstructing insight: EEG correlates of insightful problem solving. *PLoS One, 3*(1), e1459.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 26*(1), 204-221.

Stanovich, K. E. (2009). Rational and irrational thought: The thinking that IQ tests miss. *Scientific American Mind, 20*(6), 34-39.

Stanovich, K. E., & West, R. F. (2000). Individual difference in reasoning: Implication for the rationality debate. *Behavioural and Brain Science, 23*, 645-665.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129-160.

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

Thompson, V. A., Prowse Turner, J., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107-140.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237-251.

Topolinski, S. (in press). Intuition: Introducing affect into cognition. In A. Feeney & V. Thompson (Eds.), *Reasoning as memory*. Hove, UK: Psychology Press.

Topolinski, S., & Strack, F. (2009). The analysis of intuition: Processing fluency and affect in judgements of semantic coherence. *Cognition and Emotion, 23*(8), 1465-1503.

Townsend, C. L., & Heit, E. (2011). Judgments of learning and improvement. *Memory & Cognition, 39*(2), 204-216.

Tversky, A., & Kahneman, D. . (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039-1061.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1264.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behaviour* (pp. 11-32). Hove, UK: Psychology Press.

Vernon, D., & Usher, M. (2003). Dynamics of metacognitive judgments: Pre-and postretrieval mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(3), 339-346.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition, 40*, 703-716.

Wang, Y., & Chiew, V. (2010). On the cognitive process of human problem solving. *Cognitive Systems Research, 11*(1), 81-92.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310-1321. doi: 10.1098/rstb.2011.0416

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*(2), 229-241.

Journal of Experimental Psychology: General

Supplemental Materials

The Diminishing Criterion Model for Metacognitive Regulation of Time Investment

Rakefet Ackerman

Table S1. A full list of the problems used in Experiment 5.

Problem	**Units**	**Correct answer**	**Misleading answers**	**Reference**
1. A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?	cents	5	10	Kahneman and Frederick (2002)
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?	minutes	5	100	Frederick (2005)
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?	days	47	24	Frederick (2005)
4. If you flipped a fair coin 3 times, what is the probability that it would land "Heads" at least once?	percent	87.5	12.5, 37.5	Frederick (personal communication, Nov. 2009)
5. A frog fell into a hole 30 meters deep. Every day it climbs up 3 m, but during the night it slides 2 m back down. How many days will it take the frog to climb out of the hole?	days	28	30	GMAT practice book

Problem	Units	Correct answer	Misleading answers	Reference
6. Apple mash is comprised of 99% water and 1% apple solids. I left 100 kg mash in the sun and some of the water evaporated. Now the water is 98% of the mash. What is the mash weight?	kg	50	99	Uri Leron (personal communication, Nov. 2009; Attributed to Abraham Arcavi)
7. Jack is looking at Anne, and Anne is looking at George. Jack is married, but George is not. Is a married person looking at an unmarried person? A) Yes B) No C) Cannot be determined		A	C	Stanovich (2009)
8. A certain pub in town serves only whisky and coke. The cards depicted below have information about four people sitting in the pub. Each card shows a person's age on one side and what he or she is drinking on the other. It is a legal requirement that people under 18 drink coke in this pub. Select the card(s) you definitely need to turn over to determine whether anyone is breaking the law (e.g., 1, 2, 3). 17 32 Coke Whisky 1 2 3 4		1, 4	1, 3	Beaman (2002, an easier version of Wason's selection task)

Problem	Units	Correct answer	Misleading answers	Reference
9. If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5% what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's signs or symptoms?	percent	2	95	Casscells et al. (1978)
10. Every day, a bakery sells 400 cookies. When the manager is not there, 20% of the cookies made that day are eaten by the staff. How many additional cookies should be made on the manager's day off to ensure that 400 cookies can be sold?	cookies	100	80, 500	GMAT practice book
11. Steve was standing in a long line. To amuse himself he counted the people waiting, and saw that he stood 38th from the beginning and 56th from the end of the line. How many people stood in the line?	people	93	94, 92	GMAT practice book
12. Ants are walking in a line. A bad-mannered ant cuts in front of the ant walking second. What is the rude ant's place in the line?		2nd	1st	GMAT practice book

Note. The actual experimental materials were in Hebrew. For problems for which the actual version was a translation from an English source, the original English phrasing is presented.