This version is not the version of record

February 2020

Paper in press reference:
Ackerman, R., Yom-Tov, E., & Torgovitsky, I. (2020). Using confidence and consensuality to predict time invested in problem solving and in real-life web searching. *Cognition*, *199*, 104248.

Using Confidence and Consensuality to Predict Time Invested in Problem Solving and in Real-life Web Searching

Rakefet Ackerman[1], Elad Yom-Tov[2], and Ilan Torgovitsky[1]

[1] Technion—Israel Institute of Technology, [2] Microsoft Research Israel

Corresponding author: Rakefet Ackerman, Email: Ackerman@ie.technion.ac.il

Abstract

Understanding processes that lead people to invest a certain amount of time in challenging tasks is important for theory and practice. In particular, researchers often assume strong linear associations between confidence, consensuality (the degree to which an answer is independently given by multiple participants), and response time. The Diminishing Criterion Model (DCM; Ackerman, 2014) is a metacognitive model which explains the stopping rules people employ under uncertainty in terms of the confidence–time association. This model is unique in predicting a curvilinear rather than a linear confidence–time association. Using consensuality as an alternative to confidence for predicting response time offers theoretical and practical opportunities. In four experiments, including replications and variations, we examined confidence (where collected) and consensuality as predictors of the time people invest in three problem-solving tasks and in real-life web searching. The results using consensuality, like those for confidence, fitted the curvilinear time pattern predicted by the DCM, with one exception: at least 30% of the population must endorse a potential answer for consensuality to predict response time based on the stopping rules in the DCM. Beyond examining consensuality as a predictor, the study brings converging evidence supporting the DCM's curvilinear confidence–time association over alternative models. The methodology used for analyzing web searching offers new directions for metacognitive research in naturally-performed tasks.

Keyword: Metacognition; Consensuality; Problem solving; Response time; Stopping rules; Effort regulation

Highlights:

- The Diminishing Criterion Model (DCM) predicts response time based on confidence
- We examined consensuality as a replacement when confidence cannot be elicited
- Consensuality predicted response time in problem solving and real-life web searches
- We found a consistent curvilinear time pattern in line with the DCM
- This pattern was robust across multiple-choice, but not open-ended, tasks

Data for Experiments 1, 2, 3a, 3b, and 3c, are available to authorized users as supplementary material to the online version of the article.

1. Introduction

Many real-life tasks involve the challenge of time allocation in addition to performing the task per se (e.g., learning, a doctor seeing patients, preparing a multi-dish meal). When tasks are easy, people perform them quickly and confidently. But when facing challenging tasks, people may invest substantial amounts of time and nevertheless often do not succeed. This is the classic "labor in vain" (Nelson & Leonesio, 1988).

According to the metacognitive approach, people generally do not know their actual chance of success in a task, but infer their confidence (subjective chance of success) based on heuristic cues (Koriat, 1997). Overall, people express high confidence of success in tasks that are quickly performed, and acknowledge their labor in vain by expressing low confidence in relation to the most challenging items despite high time investment. This negative correlation between confidence and time is robust across many tasks, including memorizing words (e.g., Undorf & Erdfelder, 2013), answering knowledge questions (Kelley & Lindsay, 1993), reasoning and problem solving (e.g., Koriat, Ma'ayan, & Nussinson, 2006; Thompson, Prowse Turner, & Pennycook, 2011), decision making (see Unkelbach & Greifeneder, 2013, for a review; e.g., Walker, Turpin, Fugelsang, & Koehler, 2019), and even searching the web (Risko, Ferguson, & McLean, 2016). Yet the labor-in-vain phenomenon presents something of a puzzle. Its very name

suggests that people should attempt to improve their effort investment efficiency by cutting down the time they invest in items where they are not likely to succeed even after lengthy thinking. The present study deals with conditions under which people do reduce this waste of time in lab tasks as well as in real-life contexts.

1.1. Confidence–time association

The classic explanation for the negative judgment–time correlation is based on fluency as a heuristic cue that underlies the judgment. This explanation emphasizes a bottom-up regulatory process: people spontaneously invest the effort called for by the item, and then use the effort already invested to infer their confidence (see Koriat et al., 2006; Thompson & Morsanyi, 2012; see also Unkelbach & Greifeneder, 2013, for a review). Fluency-based explanations assume a *linear* association between confidence and time. In the context of decision making, a linear judgment–time association is also predicted by drift diffusion with collapsing boundaries, expressed by theoretical and computational models, with some models explaining this pattern in terms of post-decision confidence shifts (e.g., Fleming & Daw, 2017; Moran, Teodorescu, & Usher, 2015; Palestro, Weichart, Sederberg, & Turner, 2018; Pleskac & Busemeyer, 2010; Ratcliff & Smith, 2010). Notably, however, this body of literature discusses the confidence–time association mainly in the context of two-alternative forced-choice tasks. This procedural characteristic limits generalization and theorizing regarding effort regulation in more complex tasks (Funke, 2010). Importantly, in these and other tasks, linear confidence–time correlations have been found to be far from perfect (e.g., -.30 to -.54 in Koriat & Ackerman, 2010, for 7- to 12-years-old children answering general knowledge questions; -.20 to -.60 in Koriat et al., 2006, for memorizing and problem solving; or 0 to -.40 in Moran et al., 2015, and -.07 to -.40 in Pleskac & Busemeyer, 2010, for perceptual decision making). Thus, there is room for improvement in the prediction of response time.

The *Diminishing Criterion Model* (DCM, Ackerman, 2014) offers another theoretical explanation for the negative judgment–time correlation, and a data-analysis method for improving the predictability of response time. Unlike fluency-based explanations, the DCM emphasizes a top-down regulatory process directed by two stopping rules for mental effort investment. First, under the DCM, people gain confidence as they invest more effort in each task item (see also Koriat et al., 2006). They then stop investing effort when they reach a satisfactory level of confidence. This stopping rule is in line with classic metacognitive discrepancy

reduction theories (Nelson & Narens, 1990) and with drift diffusion models. Notably, however, the confidence-based stopping criterion in the DCM drops with time, reflecting a compromise as more time is invested in each item. This drop reflects the fact that spending more time on a problem is indicative of it being a hard problem, and explains the overall negative confidence–time correlation.

Second, in addition to the diminishing confidence criterion, the DCM uniquely includes a time limit. The time limit is the maximum time a person is willing to invest in each item; as such, it is set by people's motivation and characteristics of the task. Its practical importance lies in the reduction of labor in vain when people face challenging items with low chance for improvement despite additional time investment. Under the DCM, people stop investing effort when they reach their time limit regardless of their confidence level. See Fig. 1. The combination of the confidence-based diminishing criterion and the time limit yields a curvilinear confidence–time relationship, which adds explanatory value beyond the standard linear confidence–time relationships.

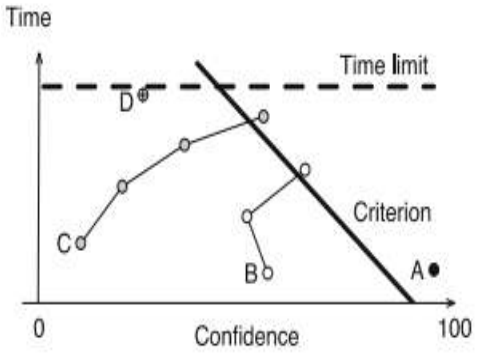

Fig. 1. The Diminishing Criterion Model proposed by Ackerman (2014). The graphical representation is adapted from Undorf and Ackerman (2017, Fig. 3 Panel B). In the figure, circles represent hypothetical confidence ratings for Items A, B, C, and D. The thick line shows how the confidence stopping criterion diminishes as time passes. The dashed line represents a time limit—the maximum time the participant is willing to invest in each item, regardless of his/her confidence at that point. Thin lines show how confidence for Items B and C progresses over time until the stopping criterion is reached. Confidence within each item tends to rise with time, while comparisons across items show a reduction in confidence as more time is invested in each item. Putting the time on the y-axis allows calculating mean time at each confidence level, thus enabling examination of whether similar amounts of time are invested across the lower confidence levels, as predicted by the DCM.

The DCM is a process model, as described below. It should be noted that as the DCM is a relatively new model, some of the assumptions on which it relies have not yet been empirically examined.

The process suggested by the model can be described as follows:

1. Participant either is instructed regarding a specific time frame for the entire task or independently defines a reasonable global time frame.
2. The two stopping criteria, confidence threshold and time limit, are set for the entire task, based on the participant's self-perceptions, motivation, and understanding of the task aims and requirements, along with task characteristics (see Ackerman, 2019; Undorf & Ackerman, 2017). The confidence threshold drops as more time is invested.
3. For each item:
    a. Participant assesses solvability. In a memorization context, as in Undorf and Ackerman (2017), and in the problem-solving and information-search situations in the current study, participants can safely assume that all presented problems are solvable. However, this is not always the case (see Ackerman & Beller, 2017; Lauterman & Ackerman, 2019; Payne & Duggan, 2011, for consideration of discrimination between solvable and unsolvable problems).
    b. Participant generates his/her first potential answer candidate. This might come to mind intuitively or be systematically calculated.
    c. Participant assesses confidence in the considered answer candidate, reflecting his/her subjective assessment of its chance of being correct.
    d. Participant compares this assessed level of confidence with his/her current confidence threshold, and compares the elapsed time (or effort) since problem presentation to his/her time (or effort) limit. If confidence and time (effort) both remain within the threshold defined by their respective stopping rules, participant continues thinking and considers an alternative solution or approach.
    In this case, stages b, c, and d are repeated. Participant (typically) gains confidence as the solving process progresses (see Figures 5, 6, and 7 in Ackerman, 2014).

e. Once either the confidence threshold or time limit is reached, participant stops and reports his/her current answer candidate and confidence.

4. Participant moves on to the next item, taking into account the remaining total time.

The DCM was developed in the context of meta-reasoning—the metacognitive processes that accompany reasoning, problem solving, and decision making (Ackerman & Thompson, 2017). Ackerman (2014) focused her empirical examination of the model on the diminishing confidence criterion in an attempt to advance the alternative explanation it offers for fluency as the main factor generating negative judgment–time associations. In three experiments (Experiments 3, 4, and 5), Ackerman (2014) found strong support for the process described by the theoretical model. In those experiments, initial, intermediate, and final confidence ratings were collected. Initial confidence ratings, provided after a mean of 7 seconds, were consistently predictive of ultimate confidence, RT, and accuracy of the provided solution.

The time limit, although included in the model in its original version, was secondary to her examination. Undorf and Ackerman (2017) tested the robustness of the DCM in the classic paired-associates memorization task, while putting more emphasis on testing factors affecting the time limit. Like Ackerman (2014), they employed a hierarchical regression model to predict the association between Judgment of Learning (JOL) and study time. However, to emphasize the role of the time limit in the regulatory process, their analysis allowed for a curvilinear pattern, extending the linear analyses considered by Ackerman (2014) when examining the DCM. Consistently across manipulations of motivation (by incentive levels) and time frames (pressured vs. loose) within and between participants, Undorf and Ackerman (2017) found the curvilinear pattern predicted based on the DCM. In all cases, regression models showed that the curvilinear pattern added predictive value beyond the linear judgment–time association. See example in Fig. 2. This was also the case when the regression analysis was applied to a computerized application for merging databases. Database merging is a highly challenging task in which database administrators must find matches between two databases with conceptually similar data structures. Including in the application elicitation of confidence ratings and a measure of response time yielded the negative correlation and the unique contribution of the curvilinear confidence–time pattern predicted by the DCM (Ackerman, Gal, Sagi, & Shraga, 2019). The overall negative slope is explained by the confidence-based criterion, and potentially also involves inference based on fluency. However, clearly, people invest similar amounts of time on

items with low to moderate metacognitive judgments—a pattern which cannot be explained solely by a single-factor fluency model or by drift diffusion models.

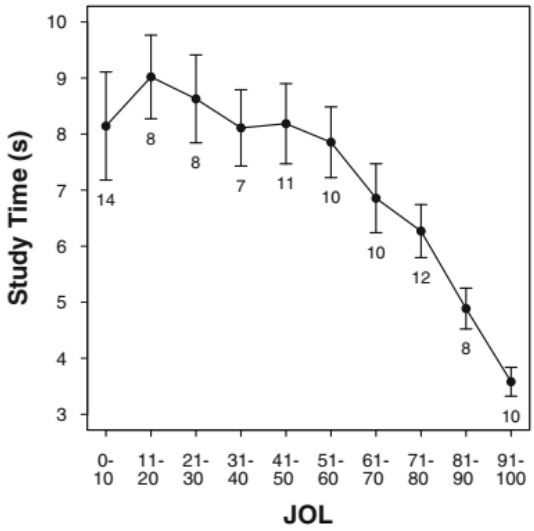

Fig. 2. Adapted from Undorf and Ackerman (2017, Fig. 1). Study time as a function of Judgment of Learning (JOL) in a classic paired-associates memorization task. Error bars represent standard errors of the mean.

1.2. Consensuality–confidence association

In lab tasks, such as those by which the DCM was proved effective for predicting invested time, researchers can collect metacognitive judgments. However, in real life people typically do not provide such explicit judgments. In the present study we considered, in addition to confidence, an alternative method for predicting the allocated time—*consensuality.* Consensuality was defined by Koriat (2008) as the percentage of participants endorsing a particular answer, regardless of its correctness. He (and others) found that the more members of any group independently select the same answer, the higher their individual levels of confidence in that answer (see Koriat, 2012). Notably, as in the decision-making research mentioned above (e.g., Fleming & Daw, 2017; Pleskac & Busemeyer, 2010), Koriat used two-alternative forced-choice tasks to examine the consensuality principle. Jackson (2016) applied consensuality to open-ended general knowledge questions, as a means of controlling for alternatives to the heuristic cue of cardinality (the number of unique answers provided for a question), which he proposed to underlie confidence. Two findings are relevant here: (a) consensuality uniquely predicted confidence, even when other potential predictors were taken into account; and (b)

people's confidence in their answers fell as the number of reasonable answer options for a question increased, as is commonly found in other contexts (e.g., choice overload in consumer behavior, see Chernev, Böckenholt, & Goodman, 2015, for a meta-analysis). However, associations between confidence, consensuality, and response time were not considered in these studies.

1.3. Consensuality–time association

Bajšanski and Žauhar (2019) examined associations between consensuality and response time by using syllogistic reasoning tasks with two-answer alternatives (valid or invalid conclusion). They found that response times for high-consensuality responses were quicker than for low-consensuality responses, indicating a general negative correlation between consensuality and response time. A similar consensus-time pattern can also be inferred from the results reported by Ackerman et al. (2019) in the context of database integration. In this case, consensus was used mainly for predicting the accuracy of the database matches.

In the present study we examined whether consensuality can be used, like confidence, to predict time allocation based on the DCM, beyond the general negative correlation between consensuality and response time previously found. The aim of predicting response time based on the DCM is to understand the considerations that play into decisions to continue or cease investing effort. Knowing the time limit for each task and the range of confidence and consensuality levels associated with various response times can be used to guide applications in numerous domains, including user interface design, educational endeavors, and legal investigations. In all these domains, and others, researchers often consider response times as indicators of confidence and/or probability of a correct response, but there are findings which question the validity of this association (e.g., Bago & De Neys, 2017; Ghazal, Cokely, & Garcia-Retamero, 2014; Hornbæk & Law, 2007; Weber, Brewer, Wells, Semmler, & Keast, 2004). By the DCM, people facing a challenging item stop investing further effort when they reach their time limit regardless of their level of confidence, and so the association between confidence and response time is not a straightforward linear correlation.

Making the transfer from confidence to consensuality as a predictor of response time posed a challenge, since we cannot assume a perfect association between confidence and consensuality. We wished to examine at what levels of confidence and consensuality we can assume similar effort regulation considerations. For instance, Dunlosky and Thiede (2004) considered the

possibility that when memorizing words, under some conditions people quickly skip the most challenging items, so that they invest less time in those items than in intermediate-difficulty items—a phenomenon termed the *shift to easier materials* (STEM; see also Metcalfe & Kornell, 2005; Son & Sethi, 2010). This effort regulation strategy is effective for reducing the labor-in-vain phenomenon. However, Undorf and Ackerman (2017) found that participants did not skip the most difficult items, but rather invested equivalent amounts of time in items with the lowest judgments of learning (JOLs) as in those with somewhat higher JOLs, in line with the DCM's time limit (see Fig. 2). Replacing lab-elicited judgments (such as confidence and JOL) with consensuality might expose situations in which people skip difficult items quickly. Items with quick response times and low-consensuality answers are presumably those in which people quickly guess answers rarely chosen by those who invest more effort in answering the question. In contrast, when confidence is used as a predictor, zero confidence may combine cases of quick skipping with cases of giving up on an item after somewhat lengthier thinking. This explanation may underlie the relatively large variation for JOLs of zero that can be observed in Fig. 2. This analysis leads us to consider the possibility that confidence and consensuality may differ in the patterns by which they predict response time at the lowest levels of their respective scales.

1.4. Study overview

In this study, we adopted the analysis methodology used by Undorf and Ackerman (2017) for memorization and applied it in four experiments to the context of solving challenging problems. This method enabled us to examine whether the DCM can be used to explain and predict time allocation in the context of such tasks. Beyond that, we replaced the judgment of learning measure used by Undorf and Ackerman (2017) with confidence and consensuality, so as to compare the two as predictors of response time. To facilitate generalization across populations, some of the experiments were conducted online and some in the lab. In Experiment 1 (online) and Experiment 2 (lab) we compared confidence and consensuality as predictors of response time in two problem-solving tasks with a multiple-choice test format, with four and eight answer options, respectively. In these tasks, we chose reasonable potential answers as lures, making the task items challenging. As a result, we expected that confidence and consensuality would both fall as the number of options rises. In Experiment 3 we examined a case in which consensuality cannot be used to predict response time, both in the lab (Experiment 3a) and online (Experiment 3b). Then, in Experiment 3c, we demonstrated how to transfer the same task to one

that can benefit from consensuality. In Experiment 4 we used consensuality to predict time-to-first-click in a data set including real-life web searches taken from Microsoft©. This experiment demonstrates how well-controlled experimentation can be generalized to real-life tasks, where collecting confidence ratings is not an option, by using consensuality to predict response times.

2. Experiment 1—Four-Alternative Analogies

Based on Undorf and Ackerman (2017), we expected a similar judgment–time pattern in problem solving as was previously found in memorization. In order to examine the validity of consensuality for predicting response time, we began our investigation with a set of multiple-choice problem-solving tasks containing four answer options. For each answer option, we calculated consensuality as the percentage of participants who chose that answer for the relevant problem. We attached to each response its consensuality. We used individual confidence and group consensuality for the answer option chosen by the participant to predict response times at the individual participant level (see Undorf & Ackerman, 2017). Limiting the number of options to four was designed to produce a high chance of consensus across participants. Moreover, unlike two-alternative forced-choice perceptual or retrieval tasks, problem solving is a complex task involving memory retrieval and data manipulation, which are multi-step processes (Funke, 2010). Such complex tasks allow room for deliberation and low confidence, and thus call for time management as an additional challenge beyond performing the task itself (see Ackerman & Thompson, 2015, 2017, for reviews).

We used analogies as our problem-solving task. The target population was an online sample of English speakers. One group rated their confidence after providing each answer, and the other group did not. This design allowed us to examine whether collecting confidence affects the consensus–time association.

2.1. Method

2.1.1. Participants

Eighty-two participants were recruited from the online survey platform Prolific Academic© (M_{age} = 36.2, *SD* = 11.5; 69% females). All participants reported being at least 18 years old; speaking English as their first language; and having either the UK, USA, Ireland, or Australia—countries where English is the main language of daily life—as their country of birth, nationality, and current residence. In addition, all participants had taken part in at least ten tasks on this platform with at least a 90% approval rate. Participants earned a monetary reward of £1.3. In the

invitation to prospective participants, we stated that the task was expected to take about 20 minutes and that participants should take part only if they could focus on the task for this period of time. The actual time on task averaged 13 minutes. Three participants were excluded for failing the attention check (see below). Among the remaining participants, 40 were in the group that rated their confidence, and 39 were in the group that did not. All participants provided informed consent.

2.1.2. *Materials*

The stimulus pool comprised 53 analogies. One was used to demonstrate the procedure and was not included in the results set. Among the remaining 52 analogies, four were chosen explicitly to be easy, with success rates greater than 90%. These problems were interspersed among the other problems and were used to screen out inattentive participants. These problems remained in the data set. The other analogies were selected to have success rates between 20% (hardest) and 80%. Altogether, including the easy items, the analogies were chosen to have average success rates around 60%. In addition, to allow room for variability in confidence and accuracy across participants and items, the analogies were designed so that confidence ratings would differ from actual success rates by no more than 30%.

Exclusion criteria for participants included: (a) failure in at least three of the four easy analogies; (b) no variability in confidence ratings (e.g., all 75%); (c) response times and success rates below 2SD of the sample; and (d) switching to a different computer window for more than 25% of the task time. Exclusion criteria for specific items were: (a) focus away from the experiment window for more than 50% of the response time; (b) particularly lengthy response times, with no other responses in a similar amount of time for this or other participants, assuming that participants were occupied by other things while the experimental window was open and on focus. The mean remaining number of items per participant was 51.5 (SD = 1.4).

2.1.3. *Procedure*

The experimental session opened with instructions, including a description of the task and the expected time required. Participants were instructed to complete the task based on their own knowledge, without consulting the Internet, books, or other people. The task was first demonstrated by one example not drawn from the stimulus pool: "Chick–Hen is like Calf–_____" (Cow), with a brief explanation of the associative similarity between the two word-

pairs. Participants were then given an easy example to practice on, "Paw–Cat, Hoof–_____," with the answer options Dog, Scorpion, Elephant, and Horse.

The confidence ratings group were shown how to use the confidence scale (0–100%). The scale ranged from "Definitely wrong"[1] to "Definitely correct." During the task, after participants provided their answer for each analogy, the answer screen was replaced by one containing the problem (but not their answer) and the confidence scale, along with a button reading "Sorry, I don't remember my answer."

The problems were presented in a random order generated for each participant, with the caveat that the easy items meant for verification were spaced about 10 items apart.

2.2. Results and discussion

Overall, the mean success rate was 62% (SD = 16.5) and the response time averaged 7.80 seconds (SD = 2.4). There were no differences in success rates or response times between the groups which provided and did not provide confidence ratings, both ts ≤ 1. As found in previous studies, confidence and consensuality were correlated. The mean within-participant correlation between consensuality and confidence was .43 (SD = .14). This correlation is significantly greater than zero, $t(39) = 19.19$, $p < .0001$, Cohen's $d = 3.03$. However, although this is a large effect size, the correlation is clearly far from a perfect match.

Following Undorf and Ackerman (2017), we fitted a multilevel regression model (level 1: items; level 2: participants) using the R packages lme4 and lmerTest (Bates, Mächler, & Bolker, 2015; Kuznetsova, Brockhoff, & Christensen, 2015; R Core Team, 2015). Prior to the analyses, we log-transformed time to reduce the skewness of the time distribution and centered confidence or consensuality at the group mean (Cohen, Cohen, West, & Aiken, 2003). The quadratic predictor, confidence^2 or consensuality^2, denotes a curvilinear relationship with time, with a positive coefficient indicating a U-shaped relation, a negative coefficient indicating an inverted U-shaped relation, and an insignificant coefficient indicating no curvilinearity. We included both the quadratic (confidence^2 or consensus^2) and the linear predictors (confidence or consensus) in the model. Significant curvilinearity in this design points to a unique predictive contribution for the curve above and beyond the linear association between predictor and time.

[1] We used this phrase as the low end of the scale consistently across experiments, rather than a phrase conveying guessing by chance. This was done for starting the scale in zero in all experiments, regardless of the answering type, open-ended or multiple-choice.

We start by examining the contribution of confidence curvilinearity as a predictor of response time, which is the unique marker of the DCM, within the group which rated their confidence. Fig. 3 Panel A graphically presents the results. The figure was generated by dividing the confidence ratings for all responses into eight bins and calculating the mean response time for each confidence bin. For theoretical reasons, the two bins at the ends of the scales contain, respectively, responses with confidence values of 100% (18% of the items) and values lower than 15% (4% of the items); the remaining data were divided into six bins with roughly equal number of responses for each participant (13% of the items).

Similarly to the findings with memorization, the negative confidence–time slope was significant, $t(2034.3) = -14.31$, $p < .0001$, but confidence^2 (squared confidence) had significant explanatory value on top of the linear slope, $t(2045.0) = -6.38$, $p < .0001$. Table 1 in the Appendix presents the Akaike Information Criterion (AIC) for the models with confidence only (linear slope) and with the addition of squared confidence (so as to capture both the linear and curvilinear patterns).

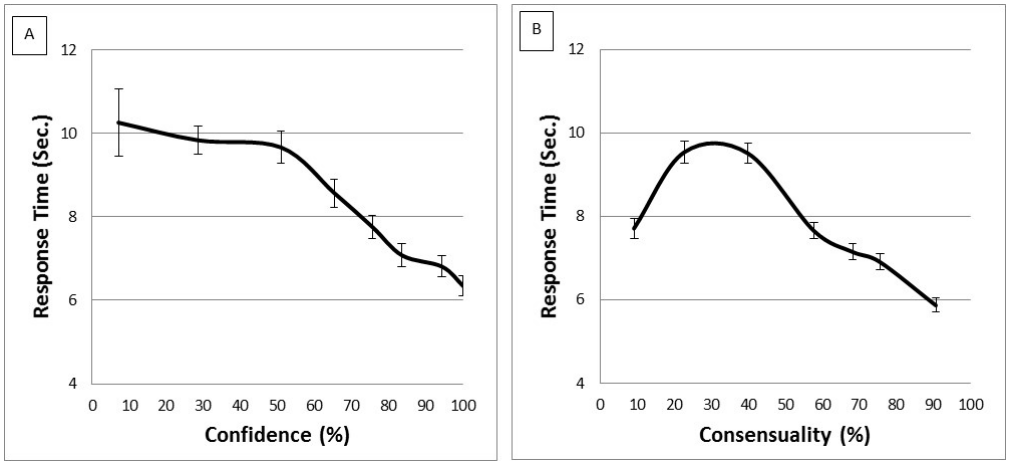

Fig. 3. Experiment 1—Analogies. The two panels represent predictions of response time by confidence (A) and by consensuality (B). Error bars represent standard errors of the mean after dividing the predictor data (confidence or consensuality) into eight bins.

We now consider consensuality as a predictor of response time. Altogether, participants provided 198 different responses to the 52 problems (out of a possible 208), which means that almost all answer options were chosen at least once. In this respect, this task is similar to the two-alternative forced-choice tasks used in previous studies. Of the full set of response options, 18 answers (1%) were chosen only once, and thus had consensuality levels of zero. No

participant chose more than one response with zero consensuality. The mean consensuality was 26.8 (SD = 17.8) for incorrect responses and 68.5 (SD = 17.6) for correct responses.

Consensuality predicted response time with a negative correlation, with no difference between respondents who did and did not provide confidence ratings: M = -.32, SD = .18 with confidence and M = -.28, SD = .21 without confidence, both ps < .0001 for the difference from zero and t < 1 for the difference between them. Comparing the linear and curvilinear predictions of response time by consensuality also reveals no significant differences between the confidence and no-confidence groups. Thus, we report the data for the group which provided confidence ratings, so that the results we present for confidence and consensuality are based on the same data set.

Fig. 3 Panel B graphically presents consensuality as a predictor of response time. The figure was generated by breaking consensuality into eight bins and calculating the mean response time for each, as was done with the confidence bins for Panel A. Similarly to the findings with confidence as a predictor, the negative consensuality–time slope was significant, $t(160.7)$ = -8.37, $p < .0001$. Consensuality^2 (squared consensuality) had significant explanatory value on top of the linear slope, $t(161.0)$ = -3.48, $p < .001$. See Table 1. Although the pattern is somewhat weaker than found for confidence, the pattern predicted by the DCM holds even when consensuality is used as a predictor of response time.

When comparing confidence and consensuality as predictors of response time, as presented visually in the two panels in Fig. 3, it is evident that the two deviate at the lowest levels of confidence and consensuality, with longer response times at the lowest levels of confidence compared to the lowest levels of consensuality. It is possible that low-consensuality responses represent quick wild guesses, in which participants barely attempted to solve the problem, while very low confidence ratings combine cases of giving up quickly and after some effort. At the high end, there were no responses with 100% consensus, while there were responses with 100% confidence. Thus, it seems that to produce similar response time patterns for confidence and consensuality, responses must be characterized by at least 20% consensuality and involve some uncertainty (i.e., not 100% confidence).

As for explaining the time people invest in solving problems, clearly, negative correlations between confidence (or consensuality) and response time are not enough to describe the association between them. This finding implies that one cannot be inferred from the other in a

straightforward manner. The curvilinear pattern predicted by the DCM better explains the pattern of association and can be generalized from memorization to solving analogies.

3. Experiment 2—Eight-Alternative Raven's Matrices

Research into reasoning and problem solving, as well as meta-reasoning research, typically makes use of verbal tasks, like conditionals (e.g., Markovits, Thompson, & Brisson, 2015; Trippas, Handley, Verde, & Morsanyi, 2016), verbally phrased math and logic problems (e.g., Sidi, Shpigelman, Zalmanov, & Ackerman, 2017; Thompson et al., 2013; Toplak, West, & Stanovich, 2014), and the analogies used in Experiment 1. Studying metacognitive aspects of solving visual problems is quite rare. Lauterman and Ackerman (2019) used Raven's matrices to study the initial judgment of solvability and its predictive value for later solving attempts. This task fits the purpose of the current study to examine the association between confidence, consensuality, and response time in challenging tasks with more than two answer alternatives, as its answering format involves an eight-alternative forced choice. While this task still has the benefit of predefined and limited answer options, we expected the average consensuality per answer option to be lower in this task than in Experiment 1, with four alternatives.

3.1. Method

3.1.1. Participants

Forty-one undergraduate students (Mean age = 26.2, 73% females) participated in the study for course credit.

3.1.2. Materials

We used thirty Raven's Progressive Matrices (RPM), in both standard and advanced versions (Raven & Court, 1998). Three additional matrices were used for instructions and four for practice (for more details see Lauterman & Ackerman, 2019).

3.1.3. Procedure

The experiment was administered in small groups of up to seven participants in a small computer lab. At the start, participants read instructions for solving Raven's matrices using three matrices, and then practiced on four matrices. After these matrices, the thirty matrices were presented in a random order. After choosing an answer, participants reported their confidence that their answer was correct on a 0 to 100% scale. The entire procedure, including instructions and practice, took about 30 minutes.

3.2. Results and discussion

The mean success rate was 74% (*SD* = 13.5), and mean response time was 28.9 seconds (*SD* = 6.8). Overall performance in the task was better than in Experiment 1 (62%). However, the lengthier response times (28.9 seconds compared to 7.8 seconds for the analogies) raise the question of whether consensuality predicts response time by way of the DCM's stopping rules even in a more time-consuming task.

The mean within-participant correlation between confidence and consensuality was .51 (*SD* = .21). This correlation was significantly greater than zero, $t(40) = 16.1$, $p < .0001$, Cohen's $d = 2.51$. Fig. 4 Panel A and Panel B present the results graphically. In a regression model conducted as in Experiment 1, the negative confidence–time slope was significant, $t(1097.4) = -4.68$, $p < .0001$. Confidence^2 had a significant effect on top of the linear slope, $t(1081.56) = -2.9$, $p = .004$. See also Table 1 in the Appendix. Although the curve is somewhat weaker than in the previous experiment, Fig. 4 Panel A shows clear adherence to a time limit as posited by the DCM. This finding provides additional evidence that the negative linear association between confidence and time predicted by other models does not fit the data.

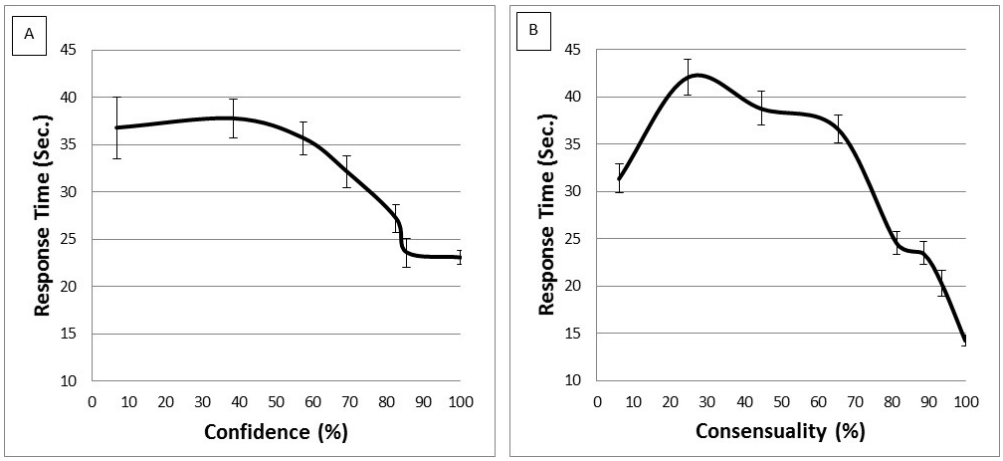

Fig. 4. Experiment 2—Raven's Matrices. The two panels represent predictions of response time by confidence (A) and by consensuality (B). Error bars represent standard errors of the mean.

As in Experiment 1, we calculated consensuality as the percentage of participants who chose a given answer option. Participants provided 145 different responses to the 30 problems. Mean consensuality was 13.3 (*SD* = 10.1) for incorrect responses and 80.3 (*SD* = 19.1) for correct

responses. Among the 145 responses, 56 answers (39%) were chosen only once, and thus had consensuality of zero. The majority of participants (68%) provided at least one response with zero consensuality, with a mean of 2.0 such responses. These cases reflect the larger number of answer options relative to the four-alternative analogies used in Experiment 1.

Fig. 4 Panel B presents consensuality as a predictor of response time. Consensuality predicted response time with a negative correlation (M = -.47, SD = .16, $t(40) = 18.6$, $p < .0001$, $d = 2.90$ for the difference from zero). The negative consensuality–time regression slope was significant, $t(57.45) = -10.67$, $p < .0001$. Consensuality^2 had a significant effect on top of the linear slope, $t(114.54) = -8.72$, $p < .0001$. See Table 1. Clearly, the pattern predicted by the DCM is valid here as well.

Here, as in Experiment 1, visually comparing the two panels in Fig. 4 reveals that the response time pattern differed between the two predictors at the scale extremes. Response times were shorter at low consensuality levels relative to somewhat higher consensuality, but the means of response times were similar to the parallel confidence means. At the high end, in this task there were some responses with full (100%) consensus and their response times were shorter than in the responses with 100% confidence. The range between 20% and 90% shows very similar patterns across the two predictors. When comparing Fig. 3 to Fig. 4, the similarity in the patterns is striking, although the present task took much more time (15–43 seconds here vs. 6–10 seconds in Experiment 1).

4. Experiment 3—Open-Ended Problems

Ackerman (2014) examined the stopping rules described in the DCM with two problem-solving tasks. In Experiment 2 in Ackerman (2014), the task was solving 30 Compound Remote Associate (CRA) problems. Each CRA problem consists of three words, and solvers must find a fourth word which forms a compound word or two-word phrase with each word of the triplet (e.g., for the triplet PINE/CRAB/SAUCE the correct answer is APPLE). As mentioned above, Ackerman (2014) analyzed the confidence–time relationships in her data using linear regressions and found negative correlations between them. For Experiment 3a, we reanalyzed data from the control group of Experiment 2 in Ackerman (2014). This group answered the CRA items without any manipulation. In order to verify the replicability of our findings, in Experiment 3b we analyzed a new data set from an online sample performing the same task.

Notably, the CRA is an open-ended task. Our aim was to examine the level of consensuality needed in order for consensuality to predict time with the pattern expected by the DCM.

4.1. **Experiment 3a**—*Reanalysis of data from Ackerman (2014)*

4.1.1. Method and procedure

Twenty undergraduate students participated in the study (M_{age} = 25.6, 28% females), which took place in a computer lab with eight seats. The materials included 30 randomly ordered Hebrew CRA problems, along with two problems used for demonstration and two as self-practice. Participants' task was to solve each problem and rate their confidence in the freely entered solution word. We measured response time from when participants clicked "Start" on an empty screen to when they clicked "Continue" after entering the solution word into an empty space below the three words. Clicking "Continue" disabled the solution entry space and exposed a continuous confidence rating scale (0–100%). Clicking "Next" after rating confidence initiated the next item. The entire experimental session took 30 minutes.

4.1.2. Results and discussion

The mean success rate was 50% (*SD* = 11.6) and mean response time was 40.1 seconds (*SD* = 9.7). Thus, the task was harder and took longer than in the previous two experiments. For the 30 CRA problems the 20 participants provided 263 different answers. Unlike in Experiment 1 with the multiple-choice analogies, 189 answers (72%) were unique, with consensuality of zero. The mean consensuality was 5.9 (*SD* = 9.1) for incorrect responses and 53.7 (*SD* = 20.2) for correct responses. The mean within-participant correlation between confidence and consensuality was .68 (*SD* = .11), $t(19) = 27.9$, $p < .0001$, Cohen's $d = 6.27$. These findings suggest that the unique responses tended to be incorrect and to be provided with low confidence.

Given that most items show no variability in consensuality, there is not much value in examining the applicability of consensuality for predicting response time in this experiment. For the sake of completeness, we report the regression results and provide a graphical representation as in the previous experiments.

Fig. 5 presents the results of the regression analysis. As in the previous experiments with confidence as the predictor, the DCM pattern was replicated again. The confidence–time slope was significantly negative, $t(572.0) = -19.4$, $p < .0001$, and the curvilinear contribution was significant as well, $t(591.7) = -3.39$, $p = .0007$. See Table 1 in the Appendix. The task was clearly challenging, as the slowest responses were provided after about 80 seconds. However, we

still see the pattern predicted by the DCM—that is, a negative confidence–time correlation until the time limit, with response times almost constant across the lowest levels of confidence.

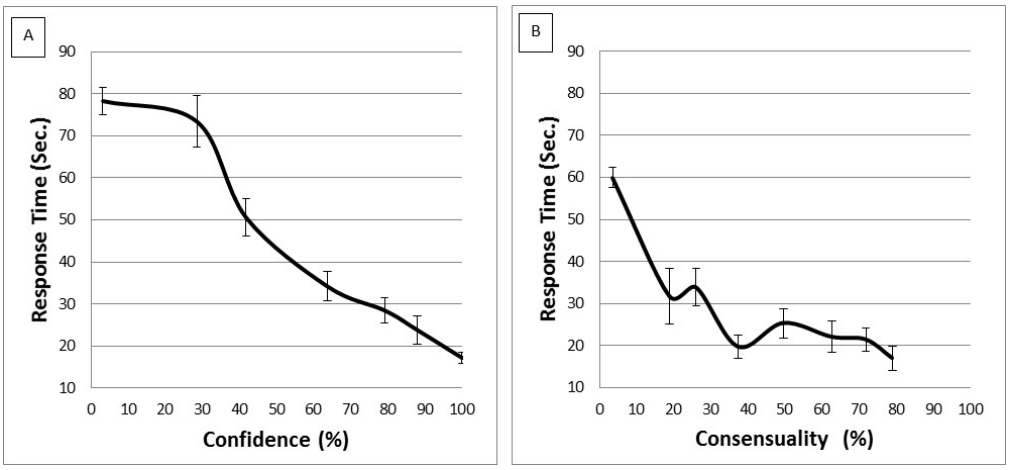

Fig. 5. Experiment 3a—Open-ended compound remote associate (CRA) problems. The two panels represent predictions of response time by confidence (A) and by consensuality (B). Error bars represent standard errors of the mean.

With respect to consensuality, given that most items had zero consensuality, the regression analysis based on consensuality revealed a different pattern of results than in the previous experiments, where the limited set of answer options generated more consensus across participants. The consensuality–time slope was negative as before, $t(528.41) = -16.13$, $p < .0001$. However, as can be seen in Fig. 5 Panel B, the curvilinear consensuality–time association was now significantly positive, $t(201.28) = +4.48$, $p < .0001$. Notably, this means again that the linear association which can be revealed by correlations is not enough to describe the data and considering the possibility of curvilinearity provides additional predictive value. See Table 1. The results of this experiment highlight that when there is consistently low consensuality on answers, response time cannot be predicted by it in terms of the stopping rules incorporated in the DCM.

4.2. **Experiment 3b**—*A replication*

In Experiment 3a, 20 participants solved 30 CRA problems. It seemed logical that to produce sufficient consensuality for analysis, more participants would be required. In Experiment 3b we used a similar number of participants as in Experiment 1 and Experiment 2, anticipating that this would allow a fair basis for comparison between the tasks.

4.2.1. Method and procedure

In Experiment 3b, 44 online participants were recruited as in Experiment 1 (M_{age} = 34.5 years, *SD* = 11.9, 73% females). The method and procedure were as in Experiment 3a. The only differences were that the 30 CRAs were in English (based on Bowden & Jung-Beeman, 2003) and the confidence ratings were collected on a separate screen, presented immediately after participants solved each problem. This enabled us to collect response times for each step separately. On the confidence ratings screen, the problem itself appeared as a reminder, without the solution entered on the previous screen. Participants could indicate that they did not remember their answer instead of rating their confidence.

4.2.2. Results and discussion

Two participants indicated that they did not remember their answer for one response each. These two responses were removed from the data set. The overall success rate was 48.6% (*SD* = 18.0), highly similar to that found in Experiment 3a, and the mean response time was 29.3 seconds (*SD* = 14.8), which is shorter than the 40.1 seconds in Experiment 3a. For 30 CRA problems the 44 participants provided 463 different answers. Despite the sample sizes not being comparable, 330 answers (71%) were unique, approximately the same percentage as found in Experiment 3a (72%). Mean consensuality was 4.3 (*SD* = 5.2) for the incorrect responses and 58.2 (*SD* = 20.6) for correct responses. Also similarly, confidence and consensuality were strongly correlated (*M* = .76, *SD* = .11, which was greater than zero, $t(43) = 47.0$, $p < .0001$, $d = 7.06$), and consensuality and response time were negatively correlated (*M* = -.53, *SD* = .16, $t(43) = 22.1$, $p < .0001$, $d = 3.33$).

Regression analyses as above, for both confidence and consensuality as predictors, revealed only significant negative linear slopes, both *p*s < .0001, and no curvilinearity, with both *t*s < 1. Table 1 in the Appendix shows that the fit of the model was slightly reduced when squared confidence and consensuality were included, unlike the case in the previous experiments. This finding highlights that improvement of the model fit is not a necessary outcome of adding predictors. In terms of the DCM, this situation deviates from the pattern expected by the DCM even for confidence as a predictor of response time. By the DCM this pattern means that the participants came up with answers which satisfied them before reaching their time limit. However, this pattern is not distinguishable from the linear negative correlations predicted by the other models we considered in the introduction.

As for consensuality, even with a comparable number of participants, it appears there is no point in examining its applicability for predicting response time based on the DCM with the open-ended and challenging CRA problems, beyond the simple effects suggested by the fact that accuracy, confidence, and consensuality are positively correlated, and all are negatively correlated with response time.

4.3. **Experiment 3c**—*Transferring an open-ended task to a multiple-choice one*

To examine our surmise that consensuality cannot be used to predict response time in open-ended tasks, we needed either to examine the same open-ended task in a multiple-choice format, or to examine the multiple-choice tasks used in Experiment 1 and Experiment 2 in an open-ended format. The three tasks used so far differ not only in their answer mode, but also in many other task characteristics. Analogies (Experiment 1) and Raven's matrices (Experiment 2) cannot be transferred to an open-ended format without changing the task substantially. In contrast, the CRA task can be easily transferred to a multiple-choice test format.

We expected that limiting CRA answers to a specific set of options would produce sufficient consensuality to allow predicting response time by way of the DCM, similarly to the other tasks we used. Thus, we constructed a four-alternative forced-choice CRA task by choosing for each CRA problem the three incorrect responses in Experiment 3b with the highest consensuality to be the lures. In the data from Experiment 3b, each problem had at least three incorrect responses with consensuality greater than zero (meaning the answer was provided by at least two participants), although some problems had more than 15 different answers. We used an online sample for this experiment, as in Experiment 3b.

4.3.1. Method and procedure

In Experiment 3c, 49 online participants were recruited as in the previous online experiments (M_{age} = 35.0 years, *SD* = 11.8, 62% females). The method and procedure were as in Experiment 3b. The only difference was that the problems appeared in a multiple-choice test format with four alternatives, rather than in an open-ended test format. The four alternatives were presented in a random order for each participant.

4.3.2. Results and discussion

Two responses were removed from the data set because the participants did not remember their answer when they were asked to rate their confidence. The overall success rate was much higher than in Experiment 3b—83.5% (*SD* = 15.0), compared to 48.6%—and the mean response

time was much shorter, 14.8 seconds (*SD* = 5.6; compared to 40.1 in Experiment 3a and 29.3 in Experiment 3b). For 30 CRA problems the 49 participants provided 94 (out of 120 possible) different answers. Among them 19 (20%) were unique. The mean consensuality was 14.2 (*SD* = 9.5) for incorrect responses and 86.0 (*SD* = 12.4) for correct responses. Confidence and consensuality were again correlated within participants (*M* = .55, *SD* = .23, which was greater than zero, $t(48) = 16.6$, $p < .0001$, $d = 2.38$), and also consensuality and response time were negatively correlated (*M* = -.37, *SD* = .21, $t(48) = 12.2$, $p < .0001$, $d = 1.74$).

Notably, now that participants had to choose the solution words from among four options, the regression analyses for both confidence and consensuality as predictors again revealed significant negative linear slopes and significant curvilinearity, as seen in Experiment 1 and Experiment 2: for confidence, $t(1433.3) = -15.0$, $p < .0001$ for the negative linear slope and $t(1431.7) = -6.47$, $p < .0001$ for curvilinearity; for consensuality, $t(46.5) = -6.32$, $p < .0001$, for the negative linear slope and $t(64.1) = -4.15$, $p = .0001$ for curvilinearity. Table 1 in the Appendix shows that the model fit patterns resemble those found in the other experiments in which allowing curvilinearity in the models yielded a significant contribution, in line with the DCM. Thus, limiting the answer options brought back consensuality as a valuable predictor of response time. Fig. 6 presents the results graphically. As found in the previous experiments, the lowest levels of confidence and consensuality deviated in how they predicted response time. In this case, the two predictors show a similar pattern of response times from 40% confidence and consensuality and upwards.

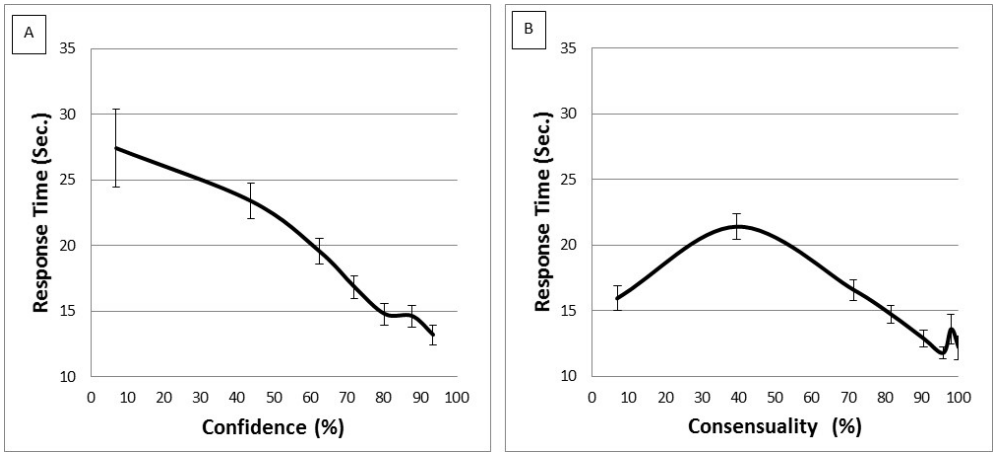

Fig. 6. Experiment 3c—Compound remote associate (CRA) problems in a multiple-choice test format. The two panels represent predictions of response time by confidence (A) and by consensuality (B). Error bars represent standard errors of the mean.

These robust findings, across populations, task types and task durations, are important for defining the scope within which consensuality is (and is not) a valuable predictor of response time. Our conclusion based on the results from Experiment 1 to Experiment 3 is that some level of consensuality is needed if consensuality is to be valuable for predicting response time when considering the pattern predicted by the two stopping criteria incorporated in the DCM.

5. Experiment 4—Real-life web searching

The ultimate goal of behavioral research is to explain and predict people's behavior in real-life scenarios. In Experiment 4 we examined whether consensuality is useful in one such case: predicting the time people invest when searching the web. In this context we do not have confidence ratings, but we can derive consensuality. We defined consensuality as the proportion of people who chose to click a particular link for a particular query. Although search engines respond to most queries with many potential links, users tend to begin by clicking a link on the first page of query results. Thus, web queries can be considered equivalent to multiple-choice questions, where searchers must choose from among a limited number of presented options, as in Experiment 1 and Experiment 2. To create the cleanest measure of response time in this noisy environment, we examined whether consensuality allows predicting time-to-first-click.

5.1. Data selection

We retrieved English-language queries submitted by people in the US to the Bing search engine (https://www.bing.com/) during three randomly chosen months in 2017–2018. We included queries that satisfied the following criteria:

1. Repeated queries—Queries had to have been made at least 100 times in a given month (necessary to allow calculation of click frequencies for each displayed link).

2. Uncertainty—We focused on informational queries involving uncertainty and exploration (Duarte, Oliveira, Côgo, & Pereira, 2015), rather than navigational queries intended to find a particular target web site well-known to the user (e.g., Facebook; Broder, 2002), which do not challenge effort regulation. This was done by using only queries in which the link clicked first was, on average, displayed in the list of results at position 3 or below (where position 1 is the topmost result).

3. Consensuality—The top-consensuality response, that is, the link with the maximal proportion of first clicks for the particular query, had to have a consensuality rate greater than

30%. This criterion is based on our conclusion from the previous experiments that a certain level of consensuality is required to predict response time in line with the DCM's explanation of behavior.

To illustrate our procedure, consider a search with the hypothetical query "using gamification to increase online sales." This search satisfies the above three criteria. In response to this query, the Bing search engine presented ten links on the first results page, with over 30,000 results altogether. Let us assume in our demonstration that result number 3, "Using gamification to boost business performance," received 32% of the first clicks; result number 4, "how to use gamification to engage employees," received 2% of the first clicks; and result number 7, "Top 10 best examples of gamification in business," received 47% of the first clicks. Thus, since 47% is larger than 30% this query is included in our data set, and result number 7, which has the highest click rate, is included in our data set.

A total of about 8000 discrete queries matched these criteria, with an average of 4400 different queries per month (some queries were common across the months). We included in our data set the top-consensuality result for each query. We defined response time as the average time which elapsed from presentation of the results until users clicked the top-consensuality result.

The chosen queries were classified into 61 topics (e.g., shopping, travel, health) using a proprietary classifier developed by Microsoft Corporation©. Each query could be associated with more than one topic. For instance, a query about flight tickets was assigned to commerce, transportation, and tourism. Assuming that topics may differ in time investment patterns, we analyzed the data in a hierarchical manner, within each topic and across topics. A topic in this analysis is analogous to a participant in lab experiments, and the specific queries are analogous to individual items (a word pair or a problem). Fifty-five topics with at least 20 queries each were included in the analyses.

5.2. Results and discussion

We examined the linear and curvilinear consensuality–time relationship predicted based on the DCM. The mean time-to-first-click predicted by consensuality is presented in Fig. 7 Panel A. We fitted a multilevel regression model (level 1: queries; level 2: topics) as above predicting time-to-first-click based on consensuality and consensuality^2. Allowing for variance in slope and

curvilinearity across participants improved the model fit (which was not the case in the experiments reported above). Thus, consensuality and consensuality^2 were added as random effects (level 1: queries; level 2: topics; level 3: consensuality; level 4: consensuality^2). We found a negative consensuality–time slope, $t(42730) = -24.7$, $p < .0001$, similar to the common negative judgment–time correlations found in many metacognitive lab studies. We also found negative curvilinearity, $t(42700) = -34.3$, $p < .0001$. The linear slope was negative for all analyzed topics but six, and the curvilinear relationship was negative for all but one.

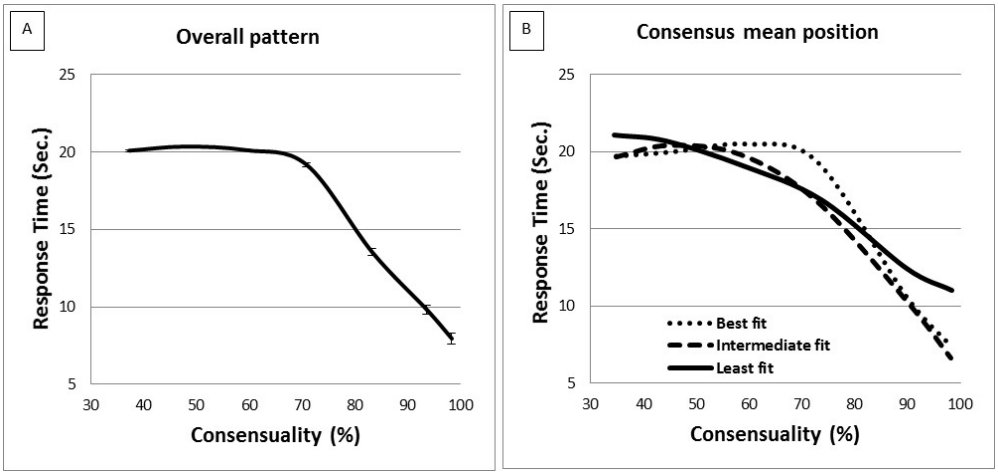

Fig. 7. Experiment 4—Web searching. Average time-to-first-click on results of web queries as a function of the level of consensuality. Panel A shows the overall pattern (with almost invisible error bars). Panel B presents the same data divided by the average position of first-clicked links with a given consensus level. The 'best fit' line reflects the highest positions in the set of query results.

To verify the robustness of this global pattern, we considered several potential moderating factors, and examined whether the linear relationships predicted by other models emerge under particular conditions. Some of the factors were considered as control variables and others as having potential to shed more light on effort regulation. Query month, query word count, average word length in the query, various divisions by topic contents, and number of queries per topic did not affect the consensuality–time pattern.

The only factor we considered that did affect the consensuality–time relationship was the position of the consensuality link among the displayed results. The positioning of query results presumably reflects the relevance of the response option to the query, as assigned by the search engine's ordering algorithm. We explored the association between time-to-first-click, first click

position, and level of consensus. We divided the data into three sets based on the relative position on the page of the link with the highest consensuality (top, intermediate, and bottom). The click frequencies were 44% in the top positions, 43% in the intermediate positions, and 13% in the bottom positions. There was a difference in consensuality–time curvilinearity across first click positions, $t(42730) = 7.9$, $p < .0001$. Despite this difference, curvilinearity remained significant regardless of the position, all ps $< .0001$. Quite surprisingly, the time limit did not change across positions. See Fig. 7 Panel B. Moreover, no single level of any of the considered factors generated the linear negative judgment–time association expected by other models.

In sum, the results consistently support the pattern of effort regulation suggested by the DCM across all considered moderating factors. Time-to-first-click fits snugly with a combination of the two stopping criteria encompassed in the DCM.

6. General Discussion

In this study we examined conditions under which, despite the moderate correlations between confidence and consensuality, the latter can be useful as a predictor of response time. We focused on an effort regulation pattern that can be explained by the stopping rules of the DCM, namely the diminishing confidence criterion and the time limit (Ackerman, 2014). Unlike other models which associate metacognitive judgments and time, and which predict a linear pattern, the DCM is unique in predicting a curvilinear judgment–time pattern. Using four different tasks—analogies, Raven's matrices, CRA problems, and real web searches—we examined conditions that consistently support using the DCM to explain patterns of time investment and considered several boundary conditions. In particular, across the experiments we examined the number of potential answer options as a factor affecting the predictive value of consensuality.

6.1. Using confidence and consensuality for predicting response time

We found both confidence and consensuality to fit the pattern predicted by the DCM in most conditions, despite the moderate correlations between them. The consistent curvilinear pattern found when predicting response time cannot be explained by bottom-up fluency alone, nor by constant thresholds. There must be a combination of factors that underlies such a curved pattern. While the difficulty of the task could explain the shared variance between confidence and time in linear models, the time limit added by the DCM shows that difficulty is not the sole

source for judgment–time associations, and emphasizes the top-down regulatory processes which take place beyond bottom-up processes.

This combination of factors cannot be exposed by the classic data analysis approaches, such as median split and correlations (e.g., Koriat et al., 2006), nor by linear regressions (e.g., Ackerman, 2014). Using the hierarchical regression analyses we employed allows exposing curvilinear judgment–time patterns and supports the diminishing confidence criterion and the time limit. This unique curvilinear signature enabled us to support the DCM to the point of ultimate generalizability, bringing converging evidence from different methods, across populations and contexts, and using tasks that differ largely in subjects' individually-set time limits (8–80 seconds). This is so despite natural "noise" involved in analyzing real-life behavior, online participation in the experiments over the noisy Internet, and when considering a large set of moderating factors (Experiment 4 here; see also Undorf & Ackerman, 2017).

Focusing on confidence, we showed here generalizability of the findings by Undorf and Ackerman (2017) and Ackerman et al. (2019) in support of the theoretical grounding of the DCM (Ackerman, 2014). Notably, Undorf and Ackerman applied the DCM to predictive judgments of learning—judgments provided after learning an item regarding success in a future recall test. Ackerman et al. (2019) used retroactive confidence in an administrative database task involving choosing pairs of attributes with the same content from among many attribute options. In the present study we used three different problem-solving tasks with retroactive confidence in finding a well-defined solution, along with web searching, which is an absolutely open task with no single correct response.

An important conclusion from this line of research, which is relevant for both theory and practice, is that time cannot be used as a direct predictor of confidence, as their relationship is not linear. In particular, invested time was not associated with confidence at the latter's lower range. A central research topic in metacognitive research is to expose heuristic cues which underlie metacognitive judgments. We call for future research to differentiate between low and high confidence levels, as it seems that the current common explanations (e.g., various types of fluency) reflect relatively high confidence levels, which show the negative confidence–time association, while low levels of confidence seem to be based on other cues. This is of particular relevance when considering two-alternative forced-choice tasks. These tasks are often taken as

representing generalizable principles, where in fact they seem to wholly conceal processes that underlie low confidence levels.

6.2. Imperfect confidence–consensuality association

Clearly, the logic underlying the diminishing confidence criterion within the DCM cannot simply be transferred to consensuality. Specifically, it is not clear whether people have a stopping rule for time investment which reflects consensuality, and whether this stopping rule is compromised as time passes. As mentioned above, robust findings from contexts of consumer behavior and decision making suggest that as people consider more options, their confidence goes down and the time they invest goes up (Chernev et al., 2015; Jackson, 2016). A question for future research is to determine the psychological principle behind the similarity in the patterns of response times between the two predictors. To the best of our knowledge, the curvilinear pattern we found has not been considered in the contexts of consumer behavior and decision making. We hope that our study provides directions for future research in these contexts, just as metacognitive research clearly can be enriched by considering choice overload and post-decision processes that are discussed in those domains. Unlike the confidence-based stopping criterion, the time limit within the DCM is not built upon confidence or consensuality, but directly on time. Thus, this theoretical idea can be considered relevant to other predictors as well.

The consensuality-based time prediction deviated from the pattern predicted based on the DCM in two respects. First, the open-ended version of the CRA problems showed negative consensuality–time correlations, as found previously with two-alternative tasks (e.g., Bajšanski & Žauhar, 2019), but also positive curvilinearity (Experiment 3a) and no curvilinearity (Experiment 3b). Clearly, the fact that 70% of the responses in the open-ended task format had no consensuality (i.e., they were unique answers) makes this predictor almost meaningless.

Low consensuality (below a minimum of about 30% of respondents) produces another deviation from the pattern predicted by the DCM. Consistently across all the multiple-choice tasks used in this study, we found quick responses with very low consensuality, unlike in responses with very low confidence, as is evident when comparing response times at the lowest ends of the predictors in the two panels in Fig. 3, Fig. 4, and Fig. 5. As we expected (see introduction), the low extremes of both scales yield differences in effort regulation, as seen in the larger variation indicated by the error bases in confidence compared with consensuality in those three figures and in Fig. 2. In particular, consensuality generated the exact pattern predicted by

memory researchers who suggested the shift-to-easier-materials principle (Dunlosky & Thiede, 2004), with relatively quick decisions to skip the items with the lowest judgments. We are not aware of similar models in the meta-reasoning domain for reasoning, problem-solving, and decision-making tasks. We call for future research to examine whether consensuality is relevant in explaining the conditions that lead to quick skipping over the hardest items in these domains, as this strategy is an effective way to reduce labor in vain.

An initial indication of skipping behavior in problem solving is apparent in the recent study by Lauterman and Ackerman (2019) mentioned above. They manipulated solvability of Raven's matrices such that half were solvable and half were unsolvable. Participants were informed that half the matrices were unsolvable, but not which ones. The experiment had two phases. In the first phase participants had to quickly decide whether each matrix was solvable or not. In the second phase, participants attempted to solve the matrices, providing an answer or a "not solvable" response. Lauterman and Ackerman found that participants invested longer in problems which they had assessed as solvable in their initial judgment, compared to those which they had assessed as unsolvable, regardless of objective solvability. On the one hand, this finding points to skipping behavior when participants had the impression that the problem was unsolvable. On the other hand, participants were determined in their solving attempts when they initially thought that a problem was solvable, even if this was objectively untrue. When all problems were solvable, Thompson et al. (2011) found that low initial Feeling of Rightness (FOR) was associated with longer solving attempts than higher FOR. Thus, it seems that highlighting the possibility that problems may be unsolvable makes a difference in strategic skipping behavior.

Reducing labor in vain is of particular importance under time pressure, as is the case in many work and education contexts (e.g., exams). Future research should also consider whether other stopping rules, beyond the two included in the DCM, guide allocation of time and contribute to the observed time patterns. In particular, future work should seek to determine what guides people to provide quick responses in situations characterized by low consensuality but not necessarily when confidence is very low.

6.3. Consensuality as a useful research and practice tool

We used consensuality as confidence is used in lab experiments (Jackson, 2016; Koriat, 2008). The study demonstrates that consensuality, despite being clearly not identical to

confidence, may be a useful tool for future metacognitive research in cases where collecting confidence is not appropriate under the experimental methodology or under conditions in which confidence elicitation is known (or suspected) to affect performance (e.g., Double & Birney, 2017). It is also useful for analyzing from a metacognitive angle data already collected for other purposes, both in controlled experiments and in existing data sources (e.g., over the Internet or in organizational databases). However, since confidence and consensuality are only moderately correlated, we call on researchers to consider factors that affect the strength of associations between consensuality, confidence, response time, and other behaviors. It is important to note that while we used the association between confidence and consensuality to make predictions based on metacognitive models, clearly consensuality can be a useful tool in other contexts as well.

The wide-scope approach we adopted in this research reveals that some level of consensuality is needed for confidence to follow the DCM (see Experiment 3b). If this finding is proved to be robust, it should be incorporated in future versions of the DCM with a proper process description. Future research should also address various questions raised by this finding. For instance, is the time limit waived when the participant has little knowledge associated with the task? What is the role played by the perceived social legitimacy of not knowing the answer (e.g., the perception that it is normal or commonplace to know/not know)? Or is there another stopping rule that guides people to stop trying before reaching their time limit despite having very low confidence in certain items?

Experiment 4 is unique within this study in directing the spotlight to a real-world application. Our findings with real-life web search data may have implications for search engine designers. In particular, our finding that some very high-consensuality links were presented relatively far down the list of results (though still on the first page) raises questions regarding the adequacy of the algorithms used to sort search results. Thus, our analysis methodology might be of value for research into bases for determining how search engines sort query results (e.g., Pan et al., 2016). For instance, subjective relevance rankings by users are often used as tool for evaluating search engine outcomes (e.g., Mao et al., 2016). Based on the results of Experiment 4, we suggest using consensuality and time-to-first-click as additional tools for evaluating and improving search outcomes. Moreover, designers might benefit from our findings regarding people's self-set time limit for selecting the first option to click on. We found that people are

willing to invest about 20 seconds in choosing which result option to click on first, but not more than that. We also call on search designers to take advantage of the robust curvilinearity across a large set of potential moderating factors.

Many additional considerations are expected to affect effort regulation in web searching. We list here some examples. First, developments in web-search technologies require adaptation of time investment patterns, as for other information-seeking contexts (Pirolli & Card, 1999). Second, there might be changes in web-search times after major news events or health scares, like a beginning of an epidemic (Yom-Tov, 2015). Third, associations between background knowledge, beliefs, and search behavior probably make a difference (Çoklar, Yaman, & Yurdakul, 2017; White & Horvitz, 2015; Yom-Tov, Marino, Pai, Harris, & Wolf, 2016). Fourth, people may change their effort regulation after gaining experience with web searching, as seen in contexts of explore–exploit dilemmas (Navarro, Newell, & Schulze, 2016). Fifth, hardware type (e.g., cellphone versus desktop machine) may be a moderating factor in web-search behavior (Ong, Järvelin, Sanderson, & Scholer, 2017), as found in learning and problem solving (Ackerman & Lauterman, 2012; Sidi et al., 2017). The research community remains at a nascent stage in understanding effort regulation in complex tasks in general, and (as these examples highlight) in web searches in particular. We call for future research to address these gaps.

6.4. Conclusion

The present study generalized the DCM to tasks not yet used to examine it. The study also extends the model by suggesting consensuality as an alternative to eliciting explicit judgments for predicting response time. At the theoretical level, we showed that both confidence and consensuality predict response time with a curvilinear, rather than linear, association. At the practical level, we showed that people indeed employ a time limit as a stopping rule, meaning there comes a point beyond which they are reluctant to invest farther effort regardless of their confidence and the consensuality of any proposed answer. We showed that this pattern is consistent across a variety of tasks, populations, and contexts. A caveat for this pattern is that the consensuality of a response – even those in which people are least confident – must reach a certain level (around 30%) for the DCM-based pattern to be valid.

Beyond the particulars of effort regulation, this study demonstrates how the validity of insights from lab experiments can be examined in real-life scenarios. We hope that our approach will provide food for thought, trigger the development of other new ideas that can be informative

for understanding human behavior, and support bringing new insights to psychological theorizing.

Acknowledgements: We thank Valerie Thompson, Ido Erev, and Taly Bonder for comments on earlier manuscript versions and Meira Ben-Gad for editorial assistance.
Funding: This work was supported by the Israel Science Foundation [grant No. 234/18].

Appendix

Table 1 complements the data analyses with the Akaike Information Criterion (AIC) for the regression models when including confidence or consensuality as linear predictors only, and when also including the squared predictor, which allows exposing the curvilinear time pattern. With the AIC, smaller values indicate a better model (Akaike, 1974).

Table 1

Akaike Information Criterion (AIC) values for the regression models.

	Confidence		Consensuality	
Model type	Linear	Curvilinear	Linear	Curvilinear
Experiment 1 group with confidence	2694.3	2656.3	2788.8	2779.6
Experiment 2	1786.1	1780.0	1790.2	1740.2
Experiment 3a	1298.5	1289.6	1448.8	1431.2
Experiment 3b	575.4	577.3	608.4	609.5
Experiment 3b	1965.0	1926.1	2153.3	2139.7
Experiment 4			18893.9	17737.8

References

Ackerman, R. (2014). The Diminishing Criterion Model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*(3), 1349-1368.

Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psychological Topics, 28*(1), 1-20.

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgments during reasoning and memorization. *Thinking & Reasoning, 23*(4), 376-408. doi:10.1080/13546783.2017.1328373

Ackerman, R., Gal, A., Sagi, T., & Shraga, R. (2019). A cognitive model of human bias in matching. *Pacific Rim International Conference on Artificial Intelligence (PRICAI)*.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816-1828.

Ackerman, R., & Thompson, V. A. (2015). Meta-Reasoning: What can we learn from meta-memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (pp. 164-182). Hove, UK: Psychology Press.

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607-617.

Akaike, H. (1974). A new look at the statistical model identification *Selected Papers of Hirotugu Akaike* (pp. 215-222): Springer.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90-109.

Bajšanski, I., & Žauhar, V. (2019). The Relationship between Consistency and Consensuality in Syllogistic Reasoning. *Psihologijske teme, 28*(1), 73-91.

Bates, D. M., Mächler, M., & Bolker, B. (2015). lme4: Linear mixedeffects models using S4 classes (R package version 1.1-6) [Software].

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

Broder, A. (2002). A taxonomy of Web search. *SIGIR forum, 36*(2), 3-10.

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology, 25*(2), 333-358.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression and correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

Çoklar, A. N., Yaman, N. D., & Yurdakul, I. K. (2017). Information literacy and digital nativity as determinants of online information search strategies. *Computers in Human Behavior, 70*, 1-9.

Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's progressive matrices. *Thinking & Reasoning, 23*(2), 190-206.

Duarte, E. F., Oliveira, E., Côgo, F. R., & Pereira, R. (2015). Dico: A conceptual model to support the design and evaluation of advanced search features for exploratory search. In J. Abascal, S. Barbosa, & M. Fetter (Eds.), *Human-Computer Interaction* (pp. 87-104). Cham: Springer.

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

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review, 124*(1), 91-114.

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

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making, 9*(1), 15-34.

Hornbæk, K., & Law, E. L.-C. (2007). *Meta-analysis of correlations among usability measures.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.
Jackson, S. A. (2016). Greater response cardinality indirectly reduces confidence. *Journal of Cognitive Psychology, 28*(4), 496-504.
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. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349-370.
Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945-959.
Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological review, 119*(1), 80-113.
Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13*(3), 441-453.
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.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (Producer). (2015). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) (R package version 2.0-6) [Software].
Lauterman, T., & Ackerman, R. (2019). Initial Judgment of Solvability in non-verbal problems–A predictor of solving processes *Metacognition and Learning, 14*(3), 365–383.
Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., . . . Luo, H. (2016). *When does Relevance Mean Usefulness and User Satisfaction in Web Search?* Paper presented at the Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.
Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & cognition, 43*(4), 681-693.
Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463-477.
Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cognitive Psychology, 78*, 99-147.
Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology, 85*, 43-77.
Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676-686.
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.
Ong, K., Järvelin, K., Sanderson, M., & Scholer, F. (2017). *Using information scent to understand mobile and desktop web search behavior.* Paper presented at the Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.
Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychonomic Bulletin & Review, 25*(4), 1225-1248.

Pan, Y., Li, H., Sarma, J., Soukal, D., Signorini, A., Gerasoulis, A., & Imielinski, T. (2016). Method and system for determining confidence in answer for search: Google Patents.
Payne, S. J., & Duggan, G. B. (2011). Giving up problem solving. *Memory & cognition, 39*(5), 902-913.
Pirolli, P., & Card, S. (1999). Information foraging. *Psychological review, 106*(4), 643-675.
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.
Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: the temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*(1), 70-94.
Raven, J., & Court, J. (1998). Manual for Raven's progressive matrices and vocabulary scales. *Oxford: Oxford Psychologists, 12*, G60p.
Risko, E. F., Ferguson, A. M., & McLean, D. (2016). On retrieving information from external knowledge stores: Feeling-of-findability, feeling-of-knowing and Internet search. *Computers in Human Behavior, 65*, 534-543.
Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61-73.
Son, L. K., & Sethi, R. (2010). Adaptive learning and the allocation of time. *Adaptive Behavior, 18*(2), 132-140.
Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: do you feel like it? *Mind & Society, 11*(1), 93-105.
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107-140.
Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L., 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.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147-168.
Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1448-1457.
Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review, 24*(6), 2003-2011. doi:10.3758/s13423-017-1261-4
Undorf, M., & Erdfelder, E. (2013). Separation of encoding fluency and item difficulty effects on judgements of learning. *The Quarterly Journal of Experimental Psychology, 66*(10), 2060-2072.
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.
Walker, A., Turpin, M. H., Fugelsang, J., & Koehler, D. (2019). Intuition speed as a predictor of choice and confidence in point spread predictions. *Judgment and Decision Making, 14*(2), 148-155.
Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: the unruly 10-12-second rule. *Journal of Experimental Psychology: Applied, 10*(3), 139–147.
White, R. W., & Horvitz, E. (2015). Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS), 33*(4), 18.
Yom-Tov, E. (2015). *Ebola data from the Internet: An opportunity for syndromic surveillance or a news event?* Paper presented at the Proceedings of the 5th international conference on digital health 2015.

Yom-Tov, E., Marino, B., Pai, J., Harris, D., & Wolf, M. (2016). The Effect of Limited Health Literacy on How Internet Users Learn About Diabetes. *Journal of health communication, 21*(10), 1107-1114.