Routledge
Taylor & Francis Group

Discourse Processes

ISSN: 0163-853X (Print) 1532-6950 (Online) Journal homepage: www.tandfonline.com/journals/hdsp20

Confidence and knowledge calibrations after reading an introductory text on a complex topic

Mandy M. Withall, Michael C. Mensink & David N. Rapp

To cite this article: Mandy M. Withall, Michael C. Mensink & David N. Rapp (2025) Confidence and knowledge calibrations after reading an introductory text on a complex topic, Discourse Processes, 62:4, 229-256, DOI: 10.1080/0163853X.2025.2470034

To link to this article: https://doi.org/10.1080/0163853X.2025.2470034

Published online: 07 Mar 2025.

DISCOURSE PROCESSES
2025, VOL. 62, NO. 4, 229–256
https://doi.org/10.1080/0163853X.2025.2470034

Routledge
Taylor & Francis Group

Confidence and knowledge calibrations after reading an introductory text on a complex topic

Mandy M. Withall [a], Michael C. Mensink [b], and David N. Rapp [a,c]

[a]Department of Psychology, Northwestern University; [b]Department of Psychology, University of Wisconsin-Stout; [c]School of Education and Social Policy, Northwestern University

ABSTRACT

Introductory texts can make complex and unfamiliar topics accessible and understandable. But after reading introductory texts, to what degree are people's understandings calibrated with confidence in their knowledge? Calibration is the degree of fit between confidence and knowledge: Accurate calibration reflects confidence judgments that align with actual knowledge, and inaccurate calibration reflects confidence judgments that are incommensurate with knowledge, manifesting as overestimations (i.e., overconfidence) or underestimations (i.e., underconfidence). We explored whether content and task-based features of introductory text experiences could support accurate calibration. In three experiments, participants read an introductory text on the potentially familiar but complex topic of natural selection, or an unrelated control text. Participants' confidence in their understandings of natural selection increased after reading the introductory text relative to before reading and relative to reading the unrelated text. Moreover, these confidence increases were calibrated with knowledge gained from reading the text. Warnings about topic complexity did not influence confidence, but taking a test that offered participants pre-reading feedback about their knowledge was beneficial for calibration. These results highlight factors that can usefully support calibration, and show that inaccurate calibrations are not a necessary consequence of introductory text experiences.

KEYWORDS

metacognition; reading; text comprehension; learning; confidence

Introduction

People can be introduced to new information from a variety of text sources including news articles, textbook chapters, and blog posts. During such experiences, people acquire knowledge as well as form metacognitive judgments about how much they know. These judgments can reflect how confident people are in their ability to make decisions, to provide explanations, and to accomplish tasks using the newly acquired information. The relationship between people's confidence in their knowledge and what they can actually demonstrate knowing is termed *calibration*, which can be measured by comparing confidence judgments and performance on a test (e.g., Bol & Hacker, 2012; Glenberg & Epstein, 1985; Keren, 1991). People are well calibrated when confidence judgments closely match test performance (i.e., increased or accurate calibration), in contrast to when confidence judgments do not match performance (i.e., decreased or inaccurate calibration). The latter mismatch can take two forms, with overconfidence reflecting confidence judgments that exceed performance and underconfidence reflecting confidence judgments that are lower than actual performance.

CONTACT Mandy M. Withall mwithall@u.northwestern.edu

Decreased calibration is a possible consequence of experiences with introductory texts, particularly if they offer a simplistic overview of a complex topic that positions the ideas as easy to understand. However, introductory texts could also foreground a topic as complex, which could make people aware that they are a novice and encourage accurate calibration between confidence and knowledge. The current study examined whether and how confidence or calibration might change as a function of reading introductory texts. We tested for this by asking participants to read a basic-level introductory text that covered a potentially familiar but scientifically complex topic (i.e., natural selection). We also examined whether features of the content and task, grounded in contemporary research on learning and text comprehension, would beneficially influence calibration after reading an introductory text. Thus, an important goal of the study was to begin to identify potential interventions that might encourage accurate calibration. The findings indeed provide insight into conditions that can usefully support people's calibrations as they learn from introductory texts.

Metacognition and its relationship to reading comprehension

Metacognition can be defined as "thinking about thinking" (e.g., Cromley, 2023; Serra & Metcalfe, 2010) or "cognition about cognitive phenomena" (Flavell, 1979). Definitions like these highlight the role of conscious cognitive processes for monitoring cognition (e.g., Hacker et al., 2009; Nelson & Narens, 1990). This can include the evaluative processes used to determine whether and how knowledge is acquired, how much knowledge has been retained after a learning experience, and how successfully one might retrieve that knowledge in the future (Cromley, 2023). These issues have been of specific interest for accounts of reading comprehension, in which examinations of metacognitive monitoring processes and their relationship to understanding text content are termed *metacomprehension* (Dunlosky & Lipko, 2007; Griffin et al., 2019; Hacker et al., 2009; Hildenbrand et al., 2024; Wiley et al., 2005). Assessments of metacomprehension often involve the collection of confidence ratings to measure people's judgments or beliefs about what they know or can do before, during, or after reading a text (e.g., Fleming, 2024; Griffin et al., 2019; Thiede & Anderson, 2003; Thiede & Dunlosky, 1999; Thiede et al., 2011; Wiley et al., 2016, 2005). These confidence ratings are then compared to actual performance on a test to determine whether people's judgments are accurate or inaccurate (Dunlosky & Lipko, 2007; Wiley et al., 2005). Some metacomprehension research calculates correlations between people's judgments and performance (i.e., relative metacognitive accuracy; Hildenbrand et al., 2024; Schraw, 2009) to determine whether people can discriminate between what they do or do not know. Other projects have focused on calibration, measuring the degree of fit between people's confidence judgments and their performance, irrespective of precision. The current study focused on whether and how people might be thinking about their own knowledge on a topic in general, and therefore assessed calibration rather than relative metacognitive accuracy.

Metacomprehension has been invoked in many accounts of text comprehension, foregrounding the critical role of readers' perceptions about their understandings alongside what they actually know. People's emerging beliefs and expectations about their knowledge can guide comprehension, such as when they become aware that some information is not well understood and make accompanying strategic changes to their reading activity to try to address the issue (e.g., Dunlosky & Rawson, 2012). When confidence and knowledge are well calibrated, people may realize they need to expend more effort to truly comprehend ideas, or in the best of situations, correctly believe they understand what they have read. But research has highlighted the many circumstances during which calibration is inaccurate, with the extant literature often focused on overconfidence as occurs when confidence levels exceed performance (e.g., Alba & Hutchinson, 2000; Dunning, 2011; Fischhoff et al., 1977; Kruger & Dunning, 1999; Lichtenstein & Fischhoff, 1977; Moore & Healy, 2008; Plohl & Musil, 2018; Sanchez & Dunning, 2018, 2020).

Overconfidence can increase as a function of a variety of contextual contributors that inform people's inferences about their comprehension, including the content of texts and the nature of a reading task or goal (e.g., León et al., 2019; Vössing et al., 2017). For example, recent research suggests that experiences with interesting texts can lead participants to make more overconfident judgments when reporting how much they believed they learned from the reading, in contrast to when they read less interesting texts (Senko et al., 2022). This may be due to a misperception of processing ease as a signal of deep understanding (Carpenter et al., 2020; Finn & Tauber, 2015). Relatedly, the inclusion of irrelevant but familiar details can increase confidence in what was learned from a text, as compared to when those details are omitted (i.e., seductive details; Hall et al., 2007). Other factors that encourage overestimates include the amount of material being studied (Tauber & Rhodes, 2010) and the assistance participants receive when trying to solve problems (e.g., letter cues to encourage word recall; Fisher & Oppenheimer, 2021).

Despite the attention that overconfidence has received in the extant literature, reading experiences can also support accurate calibration, including when they motivate people to carefully monitor their learning and performance. For example, the spacing effect is when spacing study efforts or practice problems out over time, as compared to studying everything in one sitting, leads to increased learning and calibration when estimating one's performance on a final test (e.g., Emeny et al., 2021). Explicitly asking participants to actively monitor their comprehension and to consider the risk of overconfidence also increases their evaluative behaviors and attenuates confidence judgments, in comparison to the judgments and behaviors of participants who are not asked to monitor their comprehension (Roelle et al., 2017). Additionally, asking people to recollect times they underperformed on a task can improve calibration between their metacognitive judgments and their actual performance on a subsequent task (Salovich & Rapp, 2021). In sum, empirical findings demonstrate that confidence-knowledge calibrations can be positively and negatively affected by features of learning tasks and the to-be-learned content.

The current project

The current project examined whether participants exhibit accurate or inaccurate calibrations between what they know, and their confidence in that knowledge, after reading an introductory text about a complex topic. We also examined whether warnings about that topic's complexity and information about what one actually knows (as derived from tests assessing actual knowledge on the text topic) could support accurate calibration. For this study, we adapted an introductory text on the topic of natural selection based on information obtained from an open-source biology textbook (Clark et al., 2018). We selected this topic because it is associated with an existing, validated conceptual inventory (the Conceptual Assessment of Natural Selection [CANS]; Kalinowski et al., 2016), which allowed for testing the objective knowledge participants might have on the topic. Additionally, people often hold misunderstandings about even basic aspects of natural selection (Gregory, 2009; Shtulman, 2006). This means that participants may have surface-level familiarity with the theory, including naïve conceptions or preconceived notions about its processes, but are less likely to be familiar with or able to explain those processes in any depth.

In three experiments using this text, we assessed whether particular text and task features included as part of the reading experience would support accurate calibration. A host of such features could be examined, and we selected three given their prominence in work on learning and given their direct relevance for accounts of text comprehension and metacomprehension. As a first feature, we examined whether including a warning about topic complexity would influence participants' confidence judgments. Pre-reading instructions such as warnings can direct people's attention toward relevant text elements or indicate what should be avoided in a text to enhance learning (e.g., Peshkam et al., 2011). They therefore can influence metacognitive processes including how people decide to engage with a text, what they focus on during reading, what learning goals a person might develop, or what to do with text information later

(Donovan & Rapp, 2020; McCrudden, 2019; McCrudden et al., 2010; Narvaez et al., 1999). These instructional foci have been shown to beneficially enhance learning outcomes relative to when no such instructions are provided (McCrudden et al., 2005, 2006). But while warnings seem to influence knowledge acquisition, it remains uncertain whether they might also influence confidence judgments.

In previous work, participants encouraged to monitor their comprehension and who were warned about the consequences of failing to do so showed decreases in their estimates of how much they learned from a text relative to participants who did not receive such instructions (Roelle et al., 2017). This suggests that monitoring processes, as measured by confidence judgments, can be directly impacted by warnings or pre-reading instructions. But in other work, providing warnings that people frequently make overconfident judgments has failed to reduce participants' overconfidence on answers to general knowledge questions relative to receiving no such warnings (Schall et al., 2017). These discrepant findings could be a function of whether people opt to follow or ignore instructional warnings (Salovich & Rapp, 2021), with more evidence needed to determine whether and when warnings might impact confidence judgments. Our first experiment addressed this issue by presenting participants with a post-reading warning about topic complexity to examine whether warnings can impact confidence ratings. Our goal was to test whether the warning would influence people's confidence with respect to what they had already encoded during their reading of the text.[1]

The second feature we explored was the use of pre-reading tests as a means of encouraging participants to contemplate their prior knowledge on the text topic. Tests have received substantial attention as a means of supporting memory performance, such as when people activate and recall relevant information through retrieval practice (James & Storm, 2019; Karpicke, 2009; Richland et al., 2009; Roediger & Karpicke, 2006b). Pretest questions can also impact metacognitive processes by encouraging participants to focus on relevant aspects of a subsequent text, which consequently influences learning outcomes (e.g., McCrudden et al., 2005). As pertains to confidence judgments, tests may provide people with "self-feedback" about their performance, as answering questions could make them aware of information they lack or possess (e.g., Glenberg et al., 1987; Maki & Serra, 1992). This could therefore prompt metacognitive monitoring as people reflect on their actual knowledge (Kornell & Son, 2009; Rivers, 2021). Pre-reading tests would support calibration if participants adjusted their confidence ratings based on their experience taking the test (i.e., the ease or difficulty they experience answering questions). Prior work supports this idea: Retrieval practice helped participants increase calibration by affecting both their confidence ratings and actual knowledge (Little & McDaniel, 2015). However, people are often unaware of the learning benefits of retrieval practice (Karpicke, 2009, 2012; Karpicke et al., 2009; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a), which might suggest that pre-reading tests could fail to affect confidence even when they support knowledge acquisition. Given these differing possibilities, we examined whether taking a test on the topic of natural selection prior to reading the text would support calibration.

As a third feature, we explored the utility of providing explicit feedback about participants' test performance before they read the text. Feedback is used extensively in classroom settings, and can be used to enhance learning (e.g., Gan & Hattie, 2014; Heubusch & Lloyd, 1998; Monteiro et al., 2019; Van den Bergh et al., 2013). For example, feedback can help students become more calibrated by indicating how well they have performed, potentially influencing both confidence judgments and learning outcomes (Hattie & Gan, 2010; Hattie & Timperley, 2007). Indeed, feedback in classroom settings can lead to changes in confidence judgments (e.g., Callender et al., 2016). Students can additionally use feedback to adjust self-regulated learning strategies and consequently support knowledge acquisition (Butler & Winne, 1995). Therefore, telling participants about their performance could inform calibration decisions via learning or confidence changes. To date, many feedback interventions have involved providing large amounts of feedback, but even the provision of less prominent feedback has shown benefits (Butler et al., 2007; Butler & Roediger, 2008). Therefore, the current project

examined whether providing explicit, global feedback on a pre-reading test impacted participants' confidence, knowledge, and/or calibration.

Projects that have assessed accurate and inaccurate calibration, the latter mostly with respect to overconfidence, have examined various kinds of learning experiences (e.g., predicting future events; Sanchez & Dunning, 2018). The current project tested for calibration after reading an introductory text on the complex topic of natural selection. We assessed the influence of each of the above features (i.e., warnings, pretests, and pretest feedback) in separate experiments. The project was thus intended to enrich accounts of metacomprehension by testing whether short, actionable interventions could influence readers' confidence, knowledge, and confidence-knowledge calibrations.

Experiment 1

To begin this line of inquiry, the first experiment focused solely on participants' confidence judgments about what they knew on the topic of natural selection after reading an introductory text about it, in comparison to participants who read an unrelated control text on the topic of comets. This first experiment did not measure actual knowledge, and therefore did not assess calibration, given our goal of first testing for confidence increases. Half of the participants who read the natural selection text also read a warning indicating that the topic was complex, which was intended to discourage overestimating confidence in their knowledge of natural selection by increasing metacognitive monitoring. We expected that participants who read the natural selection text would report greater confidence in their knowledge of natural selection after reading than would participants who had read the control text. This comparison offered an important examination of baseline confidence increases, which if not obtained would obviate the need for subsequent examination of calibration with knowledge. Our prediction was that participants would gain confidence from exposure to the natural selection text because they were likely to recognize that they had learned something from reading it. Participants who read the control text would also learn new information, but not about natural selection, and thus would be unlikely to show confidence gains when judging their knowledge of natural selection. We also predicted that a warning about the complexity of the topic would attenuate confidence ratings, as some prior work suggests warnings might do (Roelle et al., 2017).

Methods

Participants

This study was evaluated and approved by Northwestern University's Institutional Review Board. We calculated the number of participants necessary using G*Power (Faul et al., 2007). A minimum of 207 participants was required to detect a small-to-medium effect size. We recruited 303 MTurk workers, each receiving $5 in compensation. Seventy-four participants were removed for failure to follow instructions, for failing RelevantID checks on Qualtrics, or for self-reporting expertise on the topic of natural selection, as determined by their identified college major. This left 229 participants for analysis. Participants ranged in age from 20 to 73 years old (M = 40.0, SD = 11.6) and 39.7% identified as women. With respect to participants' self-reported identities, 3.5% identified as Asian/Pacific Islander, 4.4% identified as Black or African American, 2.6% identified as Hispanic/Latinx, 6.1% identified with more than one group or outside of these groups, and 82.5% identified as White. All participants self-reported being fluent English speakers residing in the United States and had previously completed 100 MTurk tasks with a 95% approval rating.

Materials and design

The experiment was a mixed 2 (confidence questionnaire: confidence ratings collected before reading, confidence ratings collected after reading; within-subjects) × 3 (text content: natural selection with a warning, natural selection without a warning, or comet control text; between-subjects) design. The experiment was conducted using Qualtrics' survey software (https://www.qualtrics.com/) and was

preregistered (https://aspredicted.org/4LJ_MSH). All materials for this experiment, including R code for analyses and de-identified data, can be found at https://osf.io/jvszt/?view_only=881ca23e01e94d8c9c38106e95fc160e.

Texts. The natural selection text was adapted from a biology OpenStax textbook (Clark et al., 2018) and the control text about comets was adapted from an astronomy OpenStax textbook (Fraknoi et al., 2022; see Appendix A for the texts). Each text was 1141 words long with Flesch-Kincaid grade levels ranging from 11.2 to 11.4. This indicates that both texts were similarly complex in terms of sentence and word length. Both texts were expository and were intended to educate the reader about their respective topics. The two versions of the natural selection text were identical save for the final paragraph, which included either a warning or filler text. The warning stated that "most non-experts do not have a good grasp of what natural selection is" and that "this article has only given a casual, introductory summary to the core ideas." The filler, in contrast, stated facts about Charles Darwin and his research. The control text was about comets in outer space, and contained information about their structure and how new comets are discovered.

Natural selection confidence questionnaire. We created a confidence questionnaire to assess participants' confidence in their knowledge of natural selection. The six questions on the questionnaire were designed to target different types of confidence that participants may have had with respect to their knowledge of natural selection (based on Moore & Healy, 2008). Three of the questions measured participants' estimates of their performance, considerations about their knowledge relative to others' knowledge, and how precise their beliefs were. The other three questions were related to participants' perceptions of how difficult it was to understand the topic and the amount of knowledge they possessed on the topic (see Appendix B). Participants answered each question using a scale from 1 to 100.

Procedure

After obtaining informed consent and ensuring participants were fluent English speakers who resided in the United States, participants were asked to complete the natural selection confidence questionnaire. After completing the questionnaire, participants were randomly assigned to one of the three text conditions requiring them to read the natural selection text including the warning, the natural selection text omitting the warning, or the control text. After reading, participants again answered the natural selection confidence questionnaire, followed by a set of demographics questions (e.g., age, political views, college major). Finally, participants were debriefed and thanked for their participation.

Results

Statistical analyses for this experiment were conducted using R (version 4.2.2) and the stats (R Core Team, 2023), lmtest (Zeileis & Hothorn, 2002), sandwich (Zeileis, 2004; Zeileis et al., 2020), and psych (Revelle, 2024) packages.

We correlated the six confidence questions with each other to determine if they targeted the same construct. Five of the questions were positively correlated, $r(222) = > 0.20$, $p < .01$ (see Appendix C), while a sixth question (i.e., "On a scale of 0–100, where 0 is easy and 100 is hard, how understandable is natural selection?"; question 5 in Appendix B) was not consistently correlated with the other five. We therefore omitted this question from all subsequent analyses (correlation tables for the confidence questions before and after reading appear in Appendix C, Tables C1 and C2). We averaged the numeric responses to the remaining five questions for each participant, separately for both before and after reading their assigned texts, to create confidence ratings for pre- and post-reading. The reliability of this confidence questionnaire was high (pre-reading confidence: $\alpha = 0.87$; post-reading confidence: $\alpha = 0.89$). We then calculated difference scores by subtracting pre-reading confidence ratings from

post-reading confidence ratings. All confidence ratings and confidence difference scores across experiments were calculated in this way for consistency.

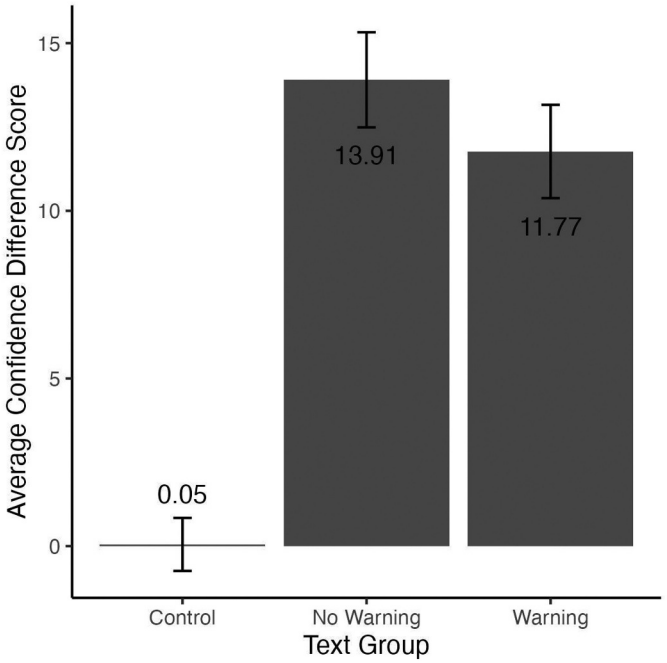

Figure 1. Confidence difference scores by condition in Experiment 1.

We fitted a linear model (estimated using OLS) to predict confidence difference scores with two contrast codes as predictors. One contrast examined the difference between reading the natural selection text (collapsed across both warning and no warning conditions) and the control text, and the other contrast examined the difference between reading the warning and no warning natural selection texts. We ran a Breusch-Pagan test which was significant, $BP(2) = 12.83$, $p = .002$, indicating that there was heteroskedasticity in the model. We ran the regression with robust standard errors in response, which was done each time the Breusch-Pagan test was significant for a model in all three experiments (see Zeileis, 2004).

After reading the natural selection text, participants' confidence scores increased by an average of thirteen points ($M = 12.8$, $SD = 12.3$), in contrast to participants in the control condition whose confidence increased by less than a point ($M = 0.05$, $SD = 6.9$). Reading the natural selection text therefore led to greater increases in confidence about the topic of natural selection than did reading the control text, $b = 17.05$, 95% CI [14.71, 19.39], $t(226) = 10.07$, $p < .001$. However, there was no difference in confidence scores from pre- to post-reading as a function of whether the natural selection text included a warning ($M = 11.8$, $SD = 12.4$) or not ($M = 13.9$, $SD = 12.3$), $b = 2.14$, 95% CI [−0.61, 4.89], $t(226) = 1.08$, $p = .283$ (see Figure 1).

Discussion

In Experiment 1, participants showed increases with respect to self-reported confidence in their understandings of natural selection after reading an introductory text on the topic. In contrast, participants who read a control text on the unrelated topic of comets did not show changes with

respect to self-reported confidence in their understandings of natural selection. It is unsurprising that our participants should believe they learned something about a topic after reading a text on that topic, but these results help exemplify how that belief is reflected in confidence gains for topic knowledge. The results also provide a conceptual baseline for considering confidence gains based solely on exposure to the natural selection text, as relevant for examinations in the subsequent experiments. Perhaps more surprising, given some previous work, is that an included warning about the complexity of the topic failed to influence participants' confidence ratings. This is not without precedent: Schall et al. (2017) presented participants with warnings about their being overconfident as they answered questions and reported their patterns of confidence were similar to the patterns of participants who had not received warnings.

While the results of Experiment 1 provide an example of confidence increases, we did not assess participants' actual understandings of natural selection. Thus, we do not know whether participants were inappropriately calibrated (e.g., their judgments did not match their performance) or appropriately calibrated (e.g., their judgments were close to their actual performance). We took this issue up directly in Experiment 2 and replaced the ineffective warning condition with a pretest condition that could provide participants with information about their knowledge when making decisions about their confidence.

Experiment 2

While the warning about topic complexity in Experiment 1 did not influence participants' confidence, other task-driven features might prove influential. One such feature, used in work on memory and text comprehension, involves presenting participants with a test on a topic prior to reading about it. Attempts to answer test questions could provide participants with an opportunity to reflect on and/or acknowledge what they do or do not know, which in turn may impact confidence. Experiment 2 examined this possibility by asking some participants to complete a test on natural selection prior to reading the natural selection or control text. After reading their assigned text, all participants completed the same test. This design served two purposes. First, it allowed for examining whether the pre-reading test was useful for tempering confidence judgments, which was expected if it encouraged participants to reflect on their actual knowledge. Second, the post-reading test afforded a means of assessing participants' objective knowledge, which was necessary for examining calibration between confidence and knowledge. Other assessment instruments have been used similarly in previous work (e.g., Maki & Serra, 1992; Nehm et al., 2022; Stoen et al., 2020).

To test participants' actual knowledge about natural selection, we used the CANS (Kalinowski et al., 2016). The CANS is a multiple-choice test that was developed to test understanding in five core concepts of natural selection: evolution, mutation, inheritance, selection, and variation. It aims to detect misconceptions related to the process of natural selection and provides multiple contexts in which the core concepts play a role (i.e., giant anteaters, bowhead whales, saguaro cacti, and mosquitos). The CANS also functioned as a test of text comprehension, as it required participants to apply what they learned from the experimental text to a variety of other contexts. As reported during CANS development work by Kalinowski et al. (2016), the average score on the CANS before any instruction was 47%, or around 11 questions correct, with the instrument's empirical reliability at 0.88. (The CANS can be found online in the supplementary materials for Kalinowski et al., 2016). In the current experiment, the CANS functioned both as a pre-reading test and a final test, allowing us to examine whether participants' knowledge changed after reading an introductory text. Coupling the CANS with the confidence questions from Experiment 1 therefore afforded a direct assessment of confidence-knowledge calibration.

We predicted, as observed in Experiment 1, that participants would report increased confidence in their knowledge of natural selection after reading the related introductory text, as compared to the

confidence ratings of participants who read the control text. We also predicted that these confidence gains would be incommensurate with any increases in knowledge that resulted from reading the text. If reading the brief introductory text led to overconfidence, we expected to see decreased calibration between confidence ratings and actual knowledge, in line with previous work (e.g., Sanchez & Dunning, 2018, 2020). However, if the reading experience helped participants to become aware of their limited knowledge on the topic, despite their general and surface-level familiarity with it, we expected to see increased confidence-knowledge calibrations after reading rather than decreased calibration. The experimental design specifically examined whether receiving the CANS prior to reading might help make participants aware that the topic was complicated when, given their lack of expertise, they did not do particularly well on it. If so, we expected participants would exhibit attenuated pre- to post-reading changes in their confidence, in contrast to the confidence gains of participants who did not take the CANS prior to reading the natural selection text. This prediction was based on findings indicating that the experience of taking a test can encourage participants to generate self-feedback about their performance (Glenberg et al., 1987; Maki & Serra, 1992). This might consequently lead to smaller confidence gains for participants in the pretest group relative to participants who did not take a pretest.

Methods

Participants

We again calculated the number of participants necessary to recruit for this experiment using G*Power (Faul et al., 2007). A minimum of 434 participants was necessary to detect a small-to-medium effect size for two between-subjects conditions. We recruited 560 MTurk workers for the experiment, each receiving $7.25 in compensation for participation. One hundred twelve participants were removed for the same reasons as described in Experiment 1. This left 452 participants for analysis, with at least 110 participants in each experimental group. Participants ranged in age from 21 to 78 years old (M = 42.2, SD = 10.9) and 47.3% identified as women. All self-reported being fluent English speakers residing in the United States and had completed at least 10,000 tasks with a 95% MTurk approval rating. With respect to participants' self-reported identities, 6.6% identified as Asian/Pacific Islander, 7.3% identified as Black or African American, 3.5% identified as Hispanic/Latinx, 5.1% identified with more than one or identified outside of these groups, and 77.4% identified as White. The experiment was conducted using Qualtrics' survey software and was preregistered (https://aspredicted.org/VT8_MS7). All materials for this experiment, including R code for analyses and de-identified data, can be found at https://osf.io/jvszt/?view_only=881ca23e01e94d8c9c38106e95fc160e.

Materials and design

We used the natural selection text (without the warning), the control text, and the natural selection confidence questionnaire from Experiment 1. The experiment was a mixed 2 (confidence questionnaire: confidence ratings collected before reading, confidence ratings collected after reading; within-subjects) × 2 (text content: natural selection or control text; between-subjects) × 2 (test condition: received the pre-reading CANS or did not receive the pre-reading CANS; between-subjects) design.

Procedure

The procedure for Experiment 2 was largely the same as Experiment 1, this time leaving out any potential warning information in the natural selection text. All participants were asked to complete the natural selection confidence questionnaire from Experiment 1 prior to and after reading their assigned text (i.e., the natural selection or control text). In contrast to Experiment 1, participants also completed

the CANS test. Specifically, half of the participants were asked to complete the pre-reading CANS after rating their confidence the first time but prior to reading their assigned text, and half did not receive the pre-reading CANS. All participants completed the post-reading CANS after reading their assigned text and rating their confidence for a second time. Thus, half of the participants received the CANS twice, once before reading and once at the end, and half received it only at the end.

Results

Statistical analyses for this experiment were conducted using R (version 4.2.2) and the stats (R Core Team, 2023), lmtest (Zeileis & Hothorn, 2002), sandwich (Zeileis, 2004; Zeileis et al., 2020), and psych (Revelle, 2024) packages.

Confidence scores

We fitted a linear model (estimated using OLS) to predict confidence difference scores (calculated in the same manner as in Experiment 1) using text condition and test condition as predictors. We ran a Breusch-Pagan test which was significant, $BP(3) = 20.94$, $p < .001$, so we conducted the regression with robust standard errors. The confidence questionnaire's reliability was again high (pre-reading confidence: $\alpha = 0.86$; post-reading confidence: $\alpha = 0.89$).

Participants showed greater increases in their confidence scores from pre- to post-reading after reading the natural selection text ($M = 11.9$, $SD = 12.8$) than after reading the control text ($M = 1.1$, $SD = 9.3$), $b = 12.59$, 95% CI [9.86, 15.32], $t(448) = 9.05$, $p < .001$. This replicated the findings of Experiment 1. There was no significant change in confidence scores between participants who took the pre-reading CANS ($M = 6.5$, $SD = 12.5$) and participants who did not ($M = 6.6$, $SD = 12.4$), $b = 1.78$, 95% CI [−0.68, 4.25], $t(448) = 1.42$, $p = .155$. There was also no interaction between text condition and test condition, $b = -3.66$, 95% CI [−7.80, 0.47], $t(448) = -1.74$, $p = .082$ (see Figure 2).

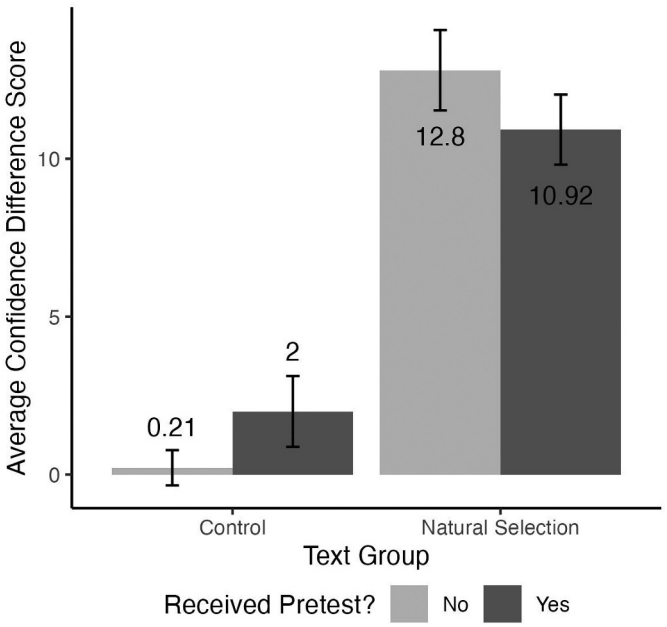

Figure 2. Confidence difference scores by condition in Experiment 2.

CANS scores

We fitted a linear model (estimated using OLS) to predict participants' post-reading CANS scores using text condition and test condition as predictors. Only post-reading CANS scores were examined, since pre-reading CANS scores were available for only half of the participants. CANS scores could range from 0–24, with 0 indicating absolutely no knowledge about the core tenets of natural selection, and 24 indicating substantial knowledge on the topic. Participants in this experiment scored about 10 points on average ($SD = 5.8$) for the pre-reading CANS, with these scores comparable to the average of 11 points reported in prior work (Kalinowski et al., 2016). The post-reading CANS' reliability was high ($\alpha = 0.90$).

Participants who read the natural selection text scored higher on the post-reading CANS ($M = 12.04$, $SD = 5.9$) than did participants who read the control text ($M = 10.03$, $SD = 6.0$), $b = 2.65$, 95% CI [1.12, 4.19], $t(448) = 3.40$, $p < .001$. Additionally, participants who took the pre-reading CANS performed better on the post-reading CANS ($M = 11.6$, $SD = 6.01$) than did participants who only took the post-reading CANS ($M = 10.5$, $SD = 6.01$), $b = 1.76$, 95% CI [0.20, 3.32], $t(448) = 2.22$, $p = .027$. There was no interaction between text condition and test condition, $b = -1.28$, 95% CI [−3.47, 0.92], $t(448) = -1.14$, $p = .253$ (see Figure 3).

Calibration bias scores

To examine the relationship between confidence and knowledge, we next analyzed calibration bias scores. These scores were calculated similarly to the bias score calculation outlined by Schraw (2009). However, the current scores were composite judgments, with participants' overall confidence rating being compared to overall performance, rather than item-by-item judgments for which participants' confidence would be assessed for each item of the CANS. Our method was nevertheless conducted in line with previous work (Gutierrez & Price, 2017; Gutierrez & Schraw, 2015). The overall CANS score for each participant was converted to a percentage that placed it on a 0–100 scale, which mapped onto confidence ratings that were already collected on that scale. Calibration bias scores were calculated by subtracting a participants' overall CANS performance percentage from their collapsed confidence score. Higher positive or negative values indicate a lack of calibration, while values closer to zero

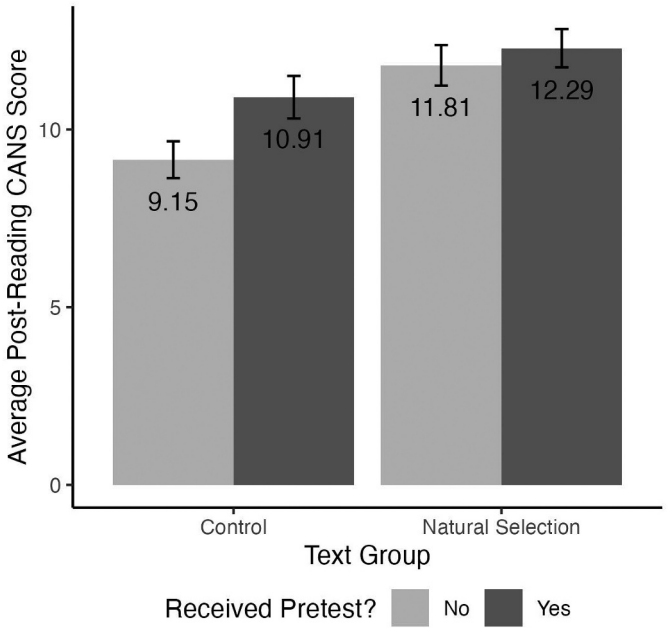

Figure 3. Post-reading CANS scores by condition in Experiment 2.

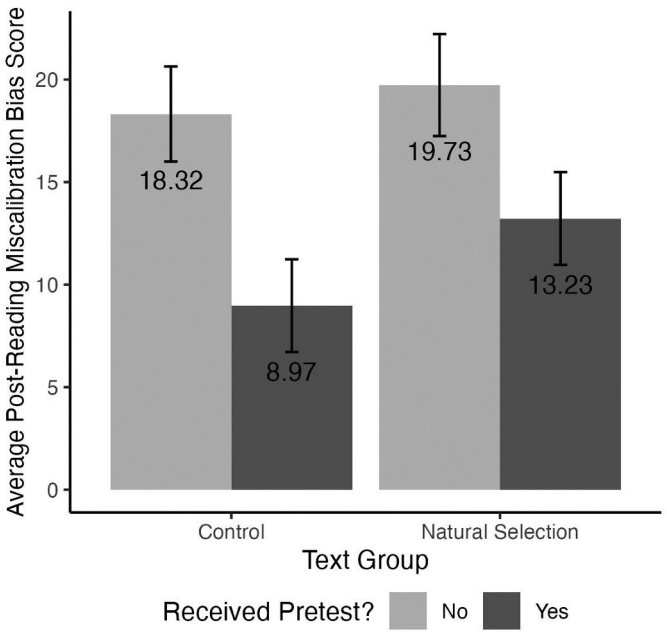

Figure 4. Post-reading miscalibration bias scores by condition in Experiment 2.

indicate calibration. These scores were only calculated based on post-reading CANS and confidence scores, since pre-reading CANS data were only available for half of the participants.

For this analysis, we fitted a linear model (estimated using OLS) to predict post-reading calibration bias scores with text condition and test condition as predictors. The results indicated that there was no difference in calibration bias scores as a function of whether participants read the natural selection text (M = 16.6, SD = 25.9) or control text (M = 13.7, SD = 24.6), b = 1.41, 95% CI [−5.02, 7.85], $t(448)$ = 0.43, p = .666. Reading the natural selection text affected both confidence and knowledge but not calibration, which suggests participants reported increased confidence in line with the knowledge they actually gained from reading the text. Additionally, participants who took the pre-reading CANS exhibited better post-reading calibration (M = 11.1, SD = 23.8) than did participants who did not take the pre-reading CANS (M = 19.0, SD = 25.9), b = −9.34, 95% CI [−15.89, −2.80], $t(448)$ = −2.80, p = .005. This result is informative when coupled with the previous analyses that showed the pre-reading CANS affected post-reading CANS scores but not confidence ratings. It suggests that calibration was specifically affected by changes in knowledge likely due to taking the pre-reading CANS. Finally, there was no interaction between text content and test condition, b = 2.84, 95% CI [−6.36, 12.04], $t(448)$ = 0.61, p = .545 (see Figure 4).

Discussion

Participants reported increased confidence in their knowledge of natural selection after reading the natural selection text, while participants who read the control text showed no confidence change. Additionally, after reading the natural selection text, participants showed increases in their actual knowledge as measured with the CANS. These confidence increases replicated the results of Experiment 1, with the CANS gains reflecting knowledge acquisition after participants read the introductory text. Taken together, the findings demonstrably show that exposures to introductory texts can be beneficial for comprehension, and that people can be aware of these benefits as reflected in confidence in their acquired knowledge.

Informative findings also emerged through examination of the effects of taking the CANS. Greater calibration was observed for participants who took the CANS prior to reading than for participants who did not. We predicted that the CANS might help participants become aware of their lack of knowledge by influencing their confidence considerations, but the results suggest a different possibility. Taking the CANS prior to reading did not seem to impact participants' confidence ratings, but rather increased their knowledge on the topic. These knowledge increases thus afforded calibration. While this pattern of results runs counter to our expectations about confidence influences, they are nevertheless in line with prior work that has shown that taking tests can enhance subsequent test performance by supporting retrieval practice and knowledge acquisition (Karpicke, 2009, 2012; Karpicke et al., 2009; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a). People also seem to be relatively unaware of these testing benefits (Karpicke & Roediger, 2008). This lack of awareness could be one reason why participants' confidence ratings were not affected by taking the pre-reading CANS but their objective knowledge increased (as measured with the post-reading CANS). In addition, participants did not receive explicit feedback on their CANS performance, meaning confidence considerations had to be based on their experience of completing what should have been a challenging test. They may not have felt the CANS was particularly challenging despite obtaining an average score of 10 on it. Participants also might not have thought much about their CANS performance in general, or even misinterpreted it, making it less likely to inform their subsequent confidence judgments.

Even so, and most importantly for this project, participants' confidence ratings were in line with the knowledge they gained from reading the text. In fact, there was no difference between calibration bias scores regardless of whether participants read the natural selection text or control text. Participants therefore did not seem to exhibit increased overconfidence as a function of reading the natural selection text. This runs counter to findings that have suggested people's overconfidence is likely to increase after brief learning experiences (e.g., Sanchez & Dunning, 2018). In addition, because the pre-reading CANS did not provide participants with explicit feedback, it remains unclear whether and how participants actually reflected on their understandings, as the CANS was intended to motivate. Explicit feedback might be more effective at helping participants think about their knowledge. In Experiment 3, we provided participants with explicit feedback on their CANS performance to see if making them aware of their knowledge on the topic would influence confidence ratings and further support calibration.

Experiment 3

Experiment 3 examined whether participants' confidence judgments would change as a function of receiving pre-reading CANS feedback. We predicted, as in the previous experiments, that participants would overall provide higher confidence ratings in their knowledge of natural selection after reading the natural selection text as compared to after reading the control text. Experiment 3 allowed for replicating whether any such increase would reflect appropriate calibration between confidence and knowledge, as observed in Experiment 2. We also predicted that feedback on the pre-reading CANS would increase calibration by encouraging participants to adjust their confidence ratings toward their actual knowledge. Participants who did not receive feedback, in contrast, would have no such motivation and remain similarly calibrated before and after reading. Recall that the warning in Experiment 1, and the experience of taking the pre-reading CANS without feedback in Experiment 2, both failed to reduce confidence ratings. The inclusion of feedback might help draw participants' attention to and awareness of a lack of knowledge on the topic, potentially reducing those ratings. However, if that feedback was ignored, undervalued, or discounted, confidence ratings would be unlikely to change as in Experiments 1 and 2.

Methods

Participants

We aimed to recruit the same number of participants as in Experiment 2 given the same number of conditions. Six hundred two prolific workers completed the experiment and received $9.34 in compensation for their participation, and 152 participants were removed for the same reasons as in Experiments 1 and 2. This left 450 participants for analysis. The participants ranged in age from 18 to 83 (M = 40.8, SD = 12.9) and 52.7% identified as women. With respect to participants' self-reported identities, 4.7% identified as Asian/Pacific Islander, 7.1% identified as Black or African American, 5.6% identified as Hispanic/Latinx, 0.4% identified as Native American or American Indian, 7.3% identified with more than one or identified outside of these groups, and 74.4% identified as White. All participants self-reported that they were fluent English speakers residing in the United States. The experiment was conducted using Qualtrics' survey software and was preregistered (https://aspredicted.org/SDF_6D9). All materials for this experiment, including R code for analyses and de-identified data, can be found at https://osf.io/jvszt/?view_only=881ca23e01e94d8c9c38106e95fc160e.

Materials and design

Experiment 3 used the natural selection text, control text, natural selection confidence questionnaire, and CANS test from Experiments 1 and 2. This time, the CANS was administered both before and after reading the assigned text for every participant. The experiment was a mixed 2 (confidence questionnaire: confidence ratings collected before reading, confidence ratings collected after reading; within-subjects) × 2 (text content: natural selection or control text; between-subjects) × 2 (feedback condition: pre-reading CANS feedback or no pre-reading CANS feedback; between-subjects) design.

Procedure

The procedure for Experiment 3 was similar to the pre-reading CANS condition in Experiment 2. All participants completed the natural selection confidence questionnaire both before and after reading the natural selection or control text. All participants also took the CANS twice, immediately after they provided their confidence ratings (i.e., both before and after reading their assigned text). Experiment 3 also included a feedback manipulation, with half of the participants receiving feedback on their pre-reading CANS performance immediately after taking it. This feedback provided a participant's score in both absolute terms (i.e., the number correct out of 24 questions) and as a percentage. Participants were also told the average score on the test is 10 out of 24, which we computed based on the average observed in Experiment 2. The other half of the participants in the feedback condition, in contrast, did not receive feedback on their pre-reading CANS performance. None of the participants received feedback on the second, post-reading CANS.

Results

Statistical analyses for this experiment were conducted using R (version 4.2.2) and the stats (R Core Team, 2023), lmtest (Zeileis & Hothorn, 2002), sandwich (Zeileis, 2004; Zeileis et al., 2020), psych (Revelle, 2024), lme4 (Bates et al., 2015), and lmerTest (Kuznetsova et al., 2017) packages.

Confidence scores

For this analysis, we fitted a linear model (estimated using OLS) to predict confidence difference scores using text condition and feedback condition as predictors. We ran a Breusch-Pagan test which was significant, $BP(3) = 24.99$, $p < .001$, so we conducted the regression with robust standard errors. The confidence questionnaire's reliability was high (pre-reading confidence: $\alpha = 0.86$; post-reading confidence: $\alpha = 0.88$).

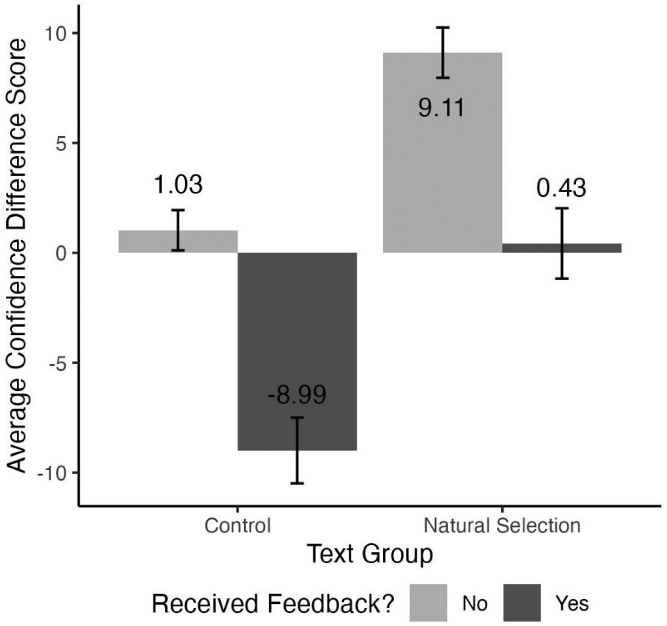

Figure 5. Confidence difference scores by condition in Experiment 3.

As in Experiments 1 and 2, participants who read the natural selection text showed increases in their confidence scores from pre- to post-reading (M = 4.8, SD = 15.2), while participants who read the control text showed reductions their confidence scores from pre- to post-reading (M = −4.0, SD = 14.2), b = 8.08, 95% CI [5.20, 10.96], $t(446)$ = 5.51, $p < .001$. In contrast to Experiments 1 and 2, the task-manipulation (i.e., providing explicit CANS feedback prior to reading) influenced confidence. Participants who received pre-reading CANS feedback showed reductions in their confidence scores from pre- to post-reading (M = −4.4, SD = 17.0), while participants who did not receive pre-reading CANS feedback showed increases in their confidence scores from pre- to post-reading (M = 5.0, SD = 11.6), b = −10.02, 95% CI [−13.47, −6.57], $t(446)$ = −5.71, $p < .001$. There was no interaction between text content and feedback, b = 1.34, 95% CI [−3.84, 6.52], $t(446)$ = 0.51, p = .612 (see Figure 5).

CANS Scores

We fitted a linear model (estimated using OLS) to predict quiz difference scores using text condition and feedback condition as predictors. We again ran a Breusch-Pagan test which was significant, BP (3) = 45.43, $p < .001$, so we conducted the regression with robust standard errors. The CANS' reliability was high (pre-CANS: α = 0.87; post-CANS: α = 0.89).

Participants who read the natural selection text showed greater CANS score increases pre- to post-reading (M = 2.9, SD = 3.7) than did participants who read the control text (M = 0.3, SD = 2.1), b = 2.54, 95% CI [1.84, 3.24], $t(446)$ = 7.13, $p < .001$. There was no effect of feedback condition, b = 0.17, 95% CI [−0.37, 0.72], $t(446)$ = 0.61, p = .539, and no interaction between text content and feedback condition, b = 0.23, 95% CI [−0.90, 1.35], $t(446)$ = 0.39, p = .694 (see Figure 6).

Calibration bias scores

We examined calibration bias scores as in Experiment 2. Analyses were run using mixed effect modeling in the lme4 R package. We switched from using linear models to a linear mixed effect model to avoid analyzing the difference between difference scores and to clarify our results.

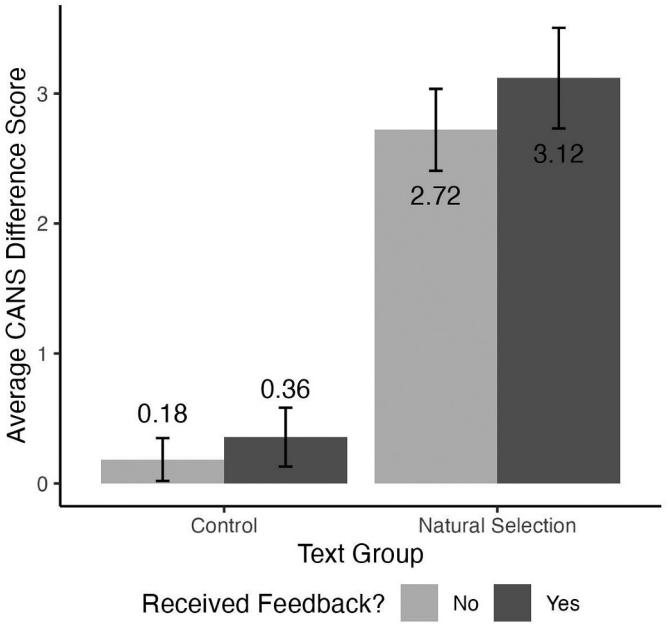

Figure 6. CANS difference scores by condition in Experiment 3.

We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict calibration bias scores using text condition, feedback condition, and calibration bias time (pre- or post-reading) as predictors. The model included participant number as a random effect. There were no main effects (all *p*s > .05), but we observed an interaction between feedback condition and when participants provided confidence ratings: Participants who received pre-reading CANS feedback had higher calibration bias scores before reading their assigned text (M = 15.3, SD = 24.0) as compared to after reading their assigned text (M = 3.8, SD = 21.9). This means that participants who received feedback were more calibrated after receiving feedback than before receiving it. In contrast, participants who did not receive pre-reading CANS feedback showed similar bias scores both before reading their assigned text (M = 18.4, SD = 23.5) and after reading it (M = 17.5, SD = 25.0), b = −10.74, 95% CI [−15.65, −5.83], $t(446)$ = −4.28, p < .001. There were no other interactions (all *p*s > .05; see Figure 7).

Discussion

Participants who read the natural selection text showed confidence increases from pre- to post-reading and, as reflected in their CANS scores, demonstrated knowing more after reading the text as compared to before reading it. As a new finding, pre-reading CANS feedback led to reductions in participants' confidence ratings from pre- to post-reading. The results indicated that the pre-reading CANS feedback did not affect subsequent quiz scores, meaning it impacted participants' beliefs about what they knew, but not necessarily their actual knowledge as other feedback interventions have tended to do (e.g., Gan & Hattie, 2014; Heubusch & Lloyd, 1998; Monteiro et al., 2019; Van den Bergh et al., 2013). This was most likely due to the short feedback experience, which did not attempt to train participants to change how they may have read and learned from the text. Additionally and importantly, participants who received pre-reading CANS feedback showed increased calibration between their confidence and knowledge from pre- to post-reading. Participants who did not receive such feedback showed a relative lack of calibration. These results thus replicate and extend the findings from Experiment 2: Confidence ratings increased for participants who read the natural selection text commensurate with what was learned, and feedback reduced participants' confidence ratings to

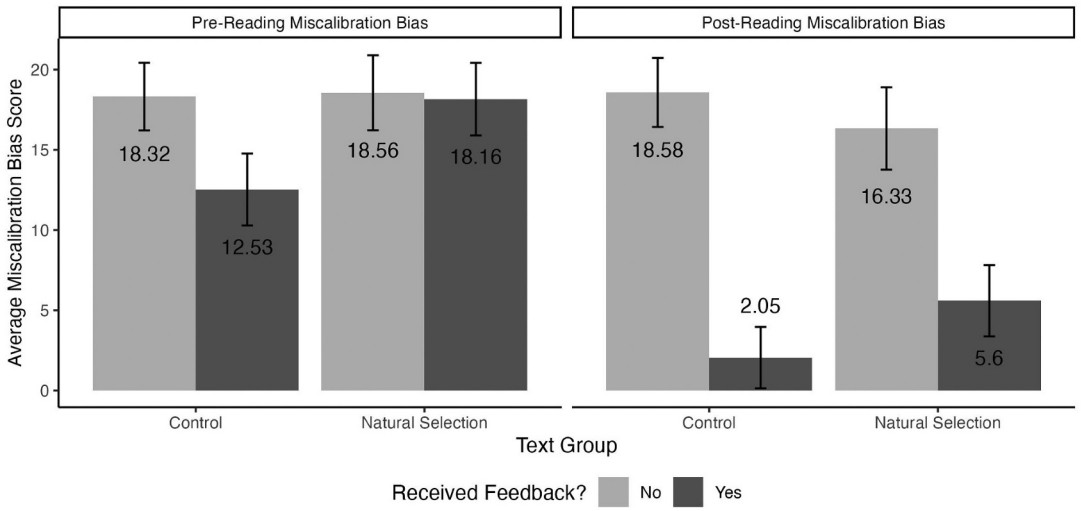

Figure 7. Miscalibration bias scores by condition in Experiment 3.

support calibration. These results are in contrast to other work on feedback, which has shown benefits to both calibration and learning (e.g., Butler et al., 2008; Carpenter et al., 2019; Mertens et al., 2022). Those differential effects may be a function of the amount or type of feedback given, as many projects in the extant literature have provided a large amount of feedback in formal instructional settings. The findings here showed calibration benefits specifically as a function of confidence changes rather than knowledge changes following a relatively modest provision of feedback.

General discussion

Introductory texts can make complex scientific topics seem easy to understand, and thus increase people's confidence with respect to understanding the topic. They can also foreground a topic as complex or encourage people to carefully consider their own prior knowledge, supporting increased calibration. The current project was intended to examine these possibilities. It was also intended to enrich accounts of metacomprehension by testing whether short, actionable task interventions could influence reader confidence, knowledge, or confidence-knowledge calibration. In three experiments, we examined whether participants who read an introductory text on the topic of natural selection would show patterns reflecting accurate or inaccurate calibration. This included testing the effects of three different text and task features regularly associated with learning, and relevant to text comprehension, on calibration patterns.

In Experiment 1, participants judged confidence in their knowledge of natural selection, read a text on natural selection or an unrelated control text, and judged confidence in their knowledge of natural selection a second time. Half of the participants who read the natural selection text also received a warning indicating that the topic is complex and difficult to understand. Participants who read the natural selection text showed greater subsequent confidence in their knowledge of natural selection than did participants who read the control text. The warning had little effect on confidence ratings, as participants showed similar increases after reading the natural selection text whether they were warned about the topic's complexity or not.

In Experiment 2, participants judged confidence in their knowledge of natural selection prior to reading a natural selection or control text, and then judged their confidence a second time. They also completed a test assessing their knowledge of natural selection (i.e., the CANS; Kalinowski et al., 2016) after reading their assigned text, with half of the participants also taking the test prior to reading. The test administered prior to reading could, in principle, encourage participants to reflect on their knowledge (Glenberg et al., 1987; Maki & Serra, 1992), which we predicted could support confidence-knowledge

calibration. Participants again showed increases in confidence after reading the natural selection text as compared to the control text, and that confidence was well-calibrated with the knowledge they gained from reading the natural selection text. The pre-reading CANS however did not affect confidence judgments, even though participants on average only answered the questions correctly around 50% of the time. Instead, the pre-reading CANS increased participants' actual knowledge about the topic. This resulted in participants being better calibrated if they received the pre-reading CANS than if they did not.

The fact that the warning and pre-reading CANS failed to reduce confidence judgments in Experiments 1 and 2 could reflect that, despite their inclusions, participants did not adjust their monitoring after receiving them. We proposed that the lack of CANS performance feedback might specifically have downplayed the potential influence of the pretest, as people might not have considered the test experience informative (e.g., Karpicke & Roediger, 2008). Thus, Experiment 3 used the same methods as Experiment 2, but required all participants to complete the CANS prior to reading their assigned text. Half of the participants received explicit feedback on their test performance. The CANS feedback specifically attenuated participants' confidence, which helpfully supported calibration between confidence and knowledge. Participants showed increases in confidence after reading the natural selection text as compared to the control text, and their confidence was well-calibrated with the knowledge they had acquired.

In sum, across the three experiments, participants generally showed confidence increases after reading the introductory natural selection text. Calibration between confidence and knowledge was supported by tests that enhanced participants' knowledge and by test feedback that informed participants' confidence considerations. Warnings, however, did little to influence participants' confidence decisions. The findings thus suggest that text and task features can affect knowledge or confidence differentially, with potential consequences for calibration profiles.

Which text and task features support calibration?

The results of Experiment 1 indicated that a warning was ineffective at reducing confidence ratings. This runs counter to previous work showing that warnings can help people attend to text contents and support comprehension (e.g., Imundo & Rapp, 2022). The warning we used focused specifically on the introductory nature of the text content, highlighting its complexity and that most people were unlikely to have expert understandings after reading it. Because we did not assess participant's prior knowledge in Experiment 1, it is possible that participants accurately estimated how much they learned after reading the introductory text. If that confidence was appropriately low given it reflected novice understanding, the warning might not have been able to substantially decrease confidence. Additionally, warnings can be ineffective if participants ignore or disregard them as relevant to their performance, or if the warning is seen as less than credible or as highlighting a possible but less than critical concern. Other projects have argued that text content can also influence whether people will recognize and heed admonitions (e.g., Schall et al., 2017). To date, relatively few projects have focused on warnings as specifically applied to people's confidence decisions rather than knowledge acquisition. Thus, the range of circumstances to which and individuals to whom warnings might prove useful in confidence contemplations represents an interesting area for future work.

Any such intended analyses can be informed by the findings of Experiments 2 and 3. Pre-reading tests, as examined in Experiment 2, helped support calibration by increasing participants' knowledge. Testing benefits are well documented in the extant literature, providing participants the opportunity to encode ideas prior to reading, and to practice retrieving their relevant knowledge (James & Storm, 2019; Karpicke, 2009; Richland et al., 2009; Roediger & Karpicke, 2006b). These benefits support knowledge acquisition, and such benefits were observed in Experiment 2. Participants' prior knowledge, as assessed by the CANS after reading, demonstrated the largest overall increases when the CANS was administered prior to reading. However, the pre-reading CANS did not seem to encourage participants to better appreciate their own knowledge (i.e., adjust confidence ratings), as confidence was the same regardless of whether or not participants took the CANS. This suggests that any efforts to encourage metacognitive considerations of knowledge may need to be more directive to influence confidence contemplations. One possibility is to

explicitly prompt people to consider how much they know with respect to a topic, or with respect to their general processing tendencies and mindsets (Salovich & Rapp, 2021). This might encourage self-regulated learning strategies, and as some work suggests, support knowledge acquisition over time (e.g., Fernandez & Jamet, 2017; Stone, 2000). Administering the pre-reading CANS on its own did not explicitly require those contemplations. Future work might examine how to encourage people to learn from pretests, either by motivating increased monitoring or educating people on the utility of pretest activity.

The findings of Experiment 3 demonstrated that providing participants with feedback on their pre-reading CANS performance can usefully support calibration. Participants who received feedback were explicitly told both how they had performed on the test and the average CANS score of other people. This feedback provided participants with information relevant for contemplating their confidence judgments and encouraged confidence changes. These confidence changes were reflected in more accurate calibration, despite participants not showing increased knowledge on the topic. Prior work suggests that feedback interventions delivered in classroom settings influence both confidence and learning outcomes (e.g., Hattie & Gan, 2010), while the present work indicated that only confidence was affected. This is likely due to the short length of the present intervention, since other interventions are typically longer and provide numerous types of feedback (e.g., Hattie & Timperley, 2007). The present work does however suggest that even short feedback interventions can be utilized to enhance calibration by affecting confidence judgments.

It is important to note though that feedback, while beneficial for confidence here, can also have potentially negative consequences. For example, people may exhibit "feedback snowballing," in which feedback encourages confidence adjustments not just to specifically relevant tasks but to other potentially unrelated tasks downstream (e.g., Murad & Starmer, 2021). Thus, confidence attenuation could be observed on subsequent tasks for which new confidence contemplations need to be applied but people instead base their considerations on potentially outdated and irrelevant feedback. Demonstrations of "feedback snowballing" suggest that people do not always reflect on what they know in ways that are most useful for supporting decisions about knowledge and self-regulated learning. While the current results are promising, further testing will be necessary to identify the scope and range of benefits derived from feedback as a tool for influencing confidence.

Limitations and future directions

The findings from the three experiments reported here suggest that calibrations after reading introductory texts can be supported. The specific text we used was an introductory description of natural selection that was designed to convey a basic understanding of the topic without including distracting or irrelevant information. Given that this carefully designed text was intended to offer a useful summary, it intentionally omitted features that might encourage inaccurate calibration. Thus, while the specific text used in this study did not affect calibration, other texts could offer various inclusions and descriptions that might exert different effects. For example, an informal blog post could claim that its introductory contents will help make the reader an expert when such a claim is unwarranted. Documenting the ways in which texts (and authors) intentionally or unintentionally convey that a reader will benefit from reading their contributions, and the consequences of reading those materials for calibration, can inform models of learning. We note that these models have focused most prominently on knowledge acquisition, and less on beliefs about what has been learned from a reading experience. The latter can be just as influential for comprehension as the former.

The project here also focused on a single topic, presenting a scientific theory that participants likely have heard of but do not fully understand. Other topics with varying levels of familiarity and relevance to one's concerns could differentially afford understandings amenable to calibration, or that perhaps even thwart it. Theories of text processing regularly foreground the need to account for how individual differences, text contents, and task factors impact and interactively influence comprehension processes and products (Gernsbacher et al., 1990; Kintsch, 1998; McCarthy & McNamara, 2021; McNamara & Magliano, 2009;

Rapp & van den Broek, 2005; Schober et al., 2018), including beliefs about what one knows and actual comprehension performance (Salovich & Rapp, 2022). The current project helps inform these models by examining text and task factors relevant to metacognitive decisions and performance derived from any acquired knowledge. Examinations of other text topics, structures, and rhetorical approaches will be needed to test the generalizability of these findings.

Finally, future work should examine whether the present interventions impact other types of metacognitive accuracy. We used calibration bias scores to discern whether participants were underconfident, overconfident, or calibrated. These scores are useful when determining whether and how participants are metacognitively monitoring their knowledge overall, but do not communicate the precision of participants' confidence judgment(s) (i.e., relative accuracy; Schraw, 2009). Studies of metacomprehension have often used precision rather than bias scores to measure metacognitive monitoring (Thiede et al., 2010; for a meta-analysis, see Prinz et al., 2020). To measure precision, participants need to rate their confidence in each question on a performance test or on multiple sets of questions. When summing over more fine-grained confidence judgments, the precision with which participants make a judgment may be masked by the calibration bias score (i.e., it is possible to appear well-calibrated in bias scores but actually have low precision). Future work could usefully examine whether warnings, pretesting, and feedback inclusions also exert an influence on relative accuracy, in accord with or different from our observed calibration patterns. This would help in the development of accounts intended to describe people's engagements with, and the consequences of, reading introductory materials.

Conclusion

People can form metacognitive judgments about what they know as they read new ideas. Previous work has often highlighted how metacognitive judgments can reflect a lack of calibration between what people think they know and their actual knowledge, usually framed with respect to overconfidence. In the current study, after reading an introductory text on the complicated topic of natural selection, participants were able to accurately estimate how much they learned from the experience. Specific task features helped support their calibration by fostering knowledge acquisition and/or confidence. These results suggest that readers may be able to engage in successful calibration when texts and tasks carefully introduce them to new topics, representing beneficial early reflections as people attempt to advance from novice to more-than-novice.

Notes

1. We opted not to present a warning prior to reading, as we conjectured doing so would influence participants' reading strategies for the task before they were even presented the text materials. The placement of warnings is an intriguing manipulation to consider in future work on people's strategic reading decisions and confidence contemplations.

Acknowledgement

A special thanks to Andrew Shtulman at Occidental College for feedback on the materials used in this project. We also thank the anonymous reviewers for their comments and advice on earlier versions of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Mandy M. Withall http://orcid.org/0000-0001-5364-6890

Michael C. Mensink http://orcid.org/0000-0001-9019-5195
David N. Rapp http://orcid.org/0000-0003-4515-5295

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research*, *27*(2), 123–156. https://doi.org/10.1086/314317

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 3. https://doi.org/10.3389/fpsyg.2012.00229

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51–64. https://doi.org/10.1037/xge0000505

Carpenter, S. K., Northern, P. E., Tauber, S. U., & Toftness, A. R. (2020). Effects of lecture fluency and instructor experience on student's judgments of learning, test scores, and evaluations of instructors. *Journal of Experimental Psychology: Applied*, *26*(1), 26–39. https://doi.org/10.1037/xap0000234

Clark, M. A., Choi, J. H., & Douglas, M. (2018). *Biology 2e* (2nd ed.). OpenStax College, Rice University. https://openstax.org/books/biology-2e/pages/1-introduction

Cromley, J. G. (2023). Metacognition, cognitive strategy instruction, and reading in adult literacy. In J. Comings, B. Garner, & C. Smith (Eds.) *Review of Adult Learning and Literacy* (Vol. 5, pp. 187–204). Routledge. https://doi.org/10.4324/9781003417958-7

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Donovan, A. M., & Rapp, D. N. (2020). Look it up: Online search reduces the problematic effects of exposures to inaccuracies. *Memory & Cognition*, *48*(7), 1128–1145. https://doi.org/10.3758/s13421-020-01047-z

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Applied Cognitive Psychology*, *35*(4), 1082–1089. https://doi.org/10.1002/acp.3814

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*(2), 131–156. https://doi.org/10.1007/s11409-016-9163-9

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology*, *3*, 552–564. https://doi.org/10.1037//0096-1523.3.4.552

Fisher, M., & Oppenheimer, D. M. (2021). Harder than you think: How outside assistance leads to overconfidence. *Psychological Science*, *32*(4), 598–610. https://doi.org/10.1177/0956797620975779
Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906
Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*(1), 241–268. https://doi.org/10.1146/annurev-psych-022423-032425
Fraknoi, A., Morrison, D., & Wolff, S. C. (2022). *Astronomy 2e* (2e ed.). OpenStax CNX. https://openstax.org/books/astronomy-2e/pages/1-introduction
Gan, M. J. S., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science*, *42*(6), 861–878. https://doi.org/10.1007/s11251-014-9319-4
Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430–445. https://doi.org/10.1037/0278-7393.16.3.430
Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 702–718. https://doi.org/10.1037/0278-7393.11.1-4.702
Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*(2), 119–136. https://doi.org/10.1037/0096-3445.116.2.119
Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach*, *2*(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1
Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1066–1092. https://doi.org/10.1037/xlm0000634
Gutierrez, A. P., & Price, A. F. (2017). Calibration Between undergraduate student's prediction of and actual performance: The role of gender and performance attributions. *The Journal of Experimental Education*, *85*(3), 486–500. https://doi.org/10.1080/00220973.2016.1180278
Gutierrez, A. P., & Schraw, G. (2015). Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education*, *83*(3), 386–404. https://doi.org/10.1080/00220973.2014.907230
Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (2009). *Handbook of metacognition in education*. Routledge.
Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*(2), 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003
Hattie, J., & Gan, M. (2010). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of Research on Learning and Instruction* (1st ed., pp. 249–271). Routledge. https://doi.org/10.4324/9780203839089
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487
Heubusch, J. D., & Lloyd, J. W. (1998). Corrective feedback in oral reading. *Journal of Behavioral Education*, *8*(1), 63–79. https://doi.org/10.1023/A:1022864707734
Hildenbrand, L., Sarmento, D., Griffin, T. D., & Wiley, J. (2024). Conceptual overlap among texts impedes comprehension monitoring. *Psychonomic Bulletin and Review*, *31*(2), 750–760. https://doi.org/10.3758/s13423-023-02349-4
Imundo, M. N., & Rapp, D. N. (2022). When fairness is flawed: Effects of false balance reporting and weight-of-evidence statements on beliefs and perceptions of climate change. *Journal of Applied Research in Memory and Cognition*, *11*(2), 258–271. https://doi.org/10.1016/j.jarmac.2021.10.002
James, K. K., & Storm, B. C. (2019). Beyond the pretesting effect: What happens to the information that is not pretested? *Journal of Experimental Psychology: Applied*, *25*(4), 576–587. https://doi.org/10.1037/xap0000231
Kalinowski, S. T., Leonard, M. J., & Taper, M. L. (2016). Development and validation of the conceptual assessment of natural selection (CANS). *CBE—Life Sciences Education*, *15*(4), 1–11. https://doi.org/10.1187/cbe.15-06-0134
Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469–486. https://doi.org/10.1037/a0017341
Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163. https://doi.org/10.1177/0963721412443552
Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408
Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

León, J. A., Martínez-huertas, J. Á., Olmos, R., Moreno, J. D., & Escudero, I. (2019). Metacomprehension skills depend on the type of text: an analysis from differential item functioning. *Psicothema*, *1*(31), 66–72. https://doi.org/10.7334/psicothema2018.163

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, *43*(1), 85–98. https://doi.org/10.3758/s13421-014-0453-7

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology*, *84*(2), 200–210. https://doi.org/10.1037/0022-0663.84.2.200

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist*, *56*(3), 196–214. https://doi.org/10.1080/00461520.2021.1872379

McCrudden, M., Schraw, G., & Kambe, G. (2005). The effect of relevance instructions on reading time and learning. *Journal of Educational Psychology*, *97*(1), 88–102. https://doi.org/10.1037/0022-0663.97.1.88

McCrudden, M. T. (2019). The effect of task relevance instructions on memory for text with seductive details. *Applied Cognitive Psychology*, *33*(1), 31–37. https://doi.org/10.1002/acp.3455

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology*, *35*(4), 229–241. https://doi.org/10.1016/j.cedpsych.2009.12.001

McCrudden, M. T., Schraw, G., & Hartley, K. (2006). The effect of general relevance instructions on shallow and deeper learning and reading time. *The Journal of Experimental Education*, *74*(4), 291–310. https://doi.org/10.3200/JEXE.74.4.291-310

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension In B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 51, pp. 297–384). Academic Press. https://doi.org/10.1016/S0079-7421(09)51009-2

Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology*, *114*(8), 1743–1772. https://doi.org/10.1037/edu0000764

Monteiro, V., Mata, L., Santos, N., Sanches, C., & Gomes, M. (2019). Classroom talk: The ubiquity of feedback. *Frontiers in Education*, *4*, 140. https://doi.org/10.3389/feduc.2019.00140

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Murad, Z., & Starmer, C. (2021). Confidence snowballing and relative performance feedback. *Journal of Economic Behavior & Organization*, *190*, 550–572. https://doi.org/10.1016/j.jebo.2021.08.006

Narvaez, D., Van Den Broek, P., & Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology*, *91*(3), 488–496. https://doi.org/10.1037/0022-0663.91.3.488

Nehm, R. H., Finch, S. J., & Sbeglia, G. C. (2022). Is active learning enough? The contributions of misconception-focused instruction and active-learning dosage on student learning of evolution. *BioScience*, *72*(11), 1105–1117. https://doi.org/10.1093/biosci/biac073

Nelson, T. O., & Narens, L. (1990). *Psychology of learning and motivation*. Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Peshkam, A., Mensink, M. C., Putnam, A. L., & Rapp, D. N. (2011). Warning readers to avoid irrelevant information: When being vague might be valuable. *Contemporary Educational Psychology*, *36*(3), 219–231. https://doi.org/10.1016/j.cedpsych.2010.10.006

Plohl, N., & Musil, B. (2018). Do I know as much as I think I do? The Dunning-Kruger effect, overclaiming, and the illusion of knowledge. *Psihološka Obzorja/Horizons of Psychology*, *27*(1), 20–30. https://doi.org/10.20419/2018.27.481

Prinz, A., Golke, S., & Wittwer, J. (2020). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review*, *31*, 100358. https://doi.org/10.1016/j.edurev.2020.100358

R Core Team. (2023). *A language and environment for statistical computing (Version 4.2.2) [R]*. R Foundation for Statistical Computing. https://www.R-project.org/

Rapp, D. N., & van den Broek, P. (2005). Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science*, *14*(5), 276–279. https://doi.org/10.1111/j.0963-7214.2005.00380.x

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research (Version 2.4.1) [R]*. Northwestern University. https://CRAN.R-project.org/package=psych

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, *33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x
Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x
Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology*, *109*(1), 99–117. https://doi.org/10.1037/edu0000132
Salovich, N. A., & Rapp, D. N. (2021). Misinformed and unaware? Metacognition and the influence of inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 608–624. https://doi.org/10.1037/xlm0000977
Salovich, N. A., & Rapp, D. N. (2022). How susceptible are you? Using feedback and monitoring to reduce the influence of false information. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1037/mac0000074
Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. https://doi.org/10.1037/pspa0000102
Sanchez, C., & Dunning, D. (2020). Decision fluency and overconfidence among beginners. *Decision*, *7*(3), 225–237. https://doi.org/10.1037/dec0000122
Schall, D. L., Doll, D., & Mohnen, A. (2017). Caution! Warnings as a useless countermeasure to reduce overconfidence? An experimental evaluation in light of enhanced and dynamic warning designs. *Journal of Behavioral Decision Making*, *30*(2), 347–358. https://doi.org/10.1002/bdm.1946
Schober, M. F., Rapp, D., & Britt, M. A. (Eds.). (2018). *The Routledge handbook of discourse processes* (2nd ed.). Routledge.
Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3
Senko, C., Perry, A. H., & Greiser, M. (2022). Does triggering learners' interest make them overconfident? *Journal of Educational Psychology*, *114*(3), 482–497. https://doi.org/10.1037/edu0000649
Serra, M. J., Metcalfe, J. (2010). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (1st ed., pp. 278–298). Routledge. https://doi.org/10.4324/9780203876428
Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, *52*(2), 170–194. https://doi.org/10.1016/j.cogpsych.2005.10.001
Stoen, S. M., McDaniel, M. A., Frey, R. F., Hynes, K. M., & Cahill, M. J. (2020). Force concept inventory: More than just conceptual understanding. *Physical Review Physics Education Research*, *16*(1), 010105. https://doi.org/10.1103/PhysRevPhysEducRes.16.010105
Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437–475. https://doi.org/10.1023/A:1009084430926
Tauber, S. K., & Rhodes, M. G. (2010). Does the amount of material to be remembered influence judgements of learning (JOLs)? *Memory*, *18*(3), 351–362. https://doi.org/10.1080/09658211003662755
Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5
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*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024
Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331–362. https://doi.org/10.1080/01638530902959927
Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264–273. https://doi.org/10.1348/135910710X510494
van den Bergh, L., Ros, A., & Beijaard, D. (2013). Teacher feedback during active learning: Current practices in primary schools. *British Journal of Educational Psychology*, *83*(2), 341–362. https://doi.org/10.1111/j.2044-8279.2012.02073.x
Vössing, J., Stamov-Roßnagel, C., & Heinitz, K. (2017). Text difficulty affects metacomprehension accuracy and knowledge test performance in text learning. *Journal of Computer Assisted Learning*, *33*(3), 282–291. https://doi.org/10.1111/jcal.12179
Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied*, *22*(4), 393–405. https://doi.org/10.1037/xap0000096
Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology*, *132*(4), 408–428. https://doi.org/10.3200/GENP.132.4.408-428
Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*(10). https://doi.org/10.18637/jss.v011.i10
Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.
Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*(1). https://doi.org/10.18637/jss.v095.i01

Appendices

Appendix A

Text materials

Text 1: natural selection

Natural selection, or "survival of the fittest," is when individuals with traits that help them survive environmental changes reproduce and create more individuals with those same traits. Individuals without those helpful traits will eventually die off.

An example of natural selection is the population of giant tortoises in the Galápagos Archipelago. The tortoises in the archipelago have longer necks than their ancestors did. This is because the long-neck Galapagos tortoises were "selected" over time due to their ability to reach more food than tortoises with short necks. In times of drought, when fewer leaves were available for the tortoises to eat, the ones that could reach more had a better chance to eat and survive than those that couldn't reach as much. Consequently, long-necked tortoises became more likely to be reproductively successful and pass the long-necked trait to their offspring while the short-necked tortoises slowly died off. Over time, only the long-necked tortoises survived.

Natural selection, Charles Darwin argued, was an inevitable outcome of three principles that operated in nature. First, most characteristics of organisms are inherited, or passed from parent to offspring. Second, more offspring are produced than can survive because there are limited resources in any given environment. Therefore, there is competition for those resources in each generation. Third, offspring vary in their characteristics. Darwin and others reasoned that offspring with inherited characteristics that allow them to best compete for limited resources will survive and have more offspring than individuals with variations that make them less able to compete. Because characteristics are inherited, the best traits are passed down to the next generation, while the worst traits are not. This leads to overall change in populations as generations go on.

Natural selection can only take place if there is variation, or differences, among individuals in a population. Importantly, these differences must have some genetic basis, otherwise selection will not lead to change in the next generation. This is a critical concept because variation can also arise for nongenetic reasons. For example, one individual may be larger than another simply because they had better nutrition and not because their genes are better suited for the environment.

These variations come from two main mechanisms: mutation and sexual reproduction. Mutation, a change in DNA, is the ultimate source of new alleles, which are new variations in genes. An example of a mutation is the first tortoise who had a longer neck: the trait spontaneously appeared due to a genetic mutation. The genetic changes that mutation causes can have a few different outcomes on the phenotype, or the characteristics that can be observed in an individual. First, a mutation can affect the organism's phenotype in a way that gives it reduced fitness—a lower likelihood of survival or fewer offspring. An example of this would be a tortoise with a mutation that made their neck even shorter. Second, a mutation may produce a phenotype that increases fitness, creating a higher likelihood of survival and more chances for offspring, like the tortoises with long necks. Lastly, there are many neutral mutations that will have no effect on the phenotype's fitness. Therefore, mutations can affect fitness a lot, a little bit, or not at all.

Sexual reproduction is the second mechanism that leads to genetic diversity. When two parents reproduce, unique combinations of alleles come together to produce unique mixtures of genes and unique phenotypes in each offspring.

An inherited trait is considered to be an adaptation if it helps the organism survive and reproduce in its current environment. They can be passed down from parents to children. Adaptations are consequence of natural selection. Scientists will say that a group of organisms is adapting to its environment when a genetic variation occurs over time that increases or maintains the population's "fit" to its environment. The webbed feet of a platypus are an adaptation for swimming. A snow leopard's thick fur is an adaptation for living in the cold. A cheetah's fast speed is an adaptation for catching prey.

Whether or not a trait is favorable depends on current environmental conditions. The same traits are not always selected because environmental conditions can change. For example, consider a plant species that grew in a moist climate and did not need to conserve water. Large leaves were naturally selected because they allowed the plant to obtain more energy from the sun. However, large leaves require more water to maintain than small leaves. The moist environment provided favorable conditions to support large leaves, so they were the most "fit" for that environment. If the plant's climate were to become drier, than the direction of natural selection would shift so that plants with small leaves were now the "fit" ones, since those populations were able to conserve water and survive the new environmental conditions.

Natural selection acts on individual organisms, and if this happens to enough individuals, it shapes their entire species. The gradual changes of a species over generations, driven by natural selection, is called evolution. Sometimes evolution gives rise to groups of organisms that become tremendously different from each other. One example is the Galapagos finches, who have different beaks depending on which island they live on. In other cases, similar observable traits evolve independently in distantly related species. For example, the ability to fly has evolved in both bats and insects, as they both

have structures we refer to as wings and that are used to fly. However, bat and insect wings have evolved from very different original structures.

Natural selection can generate populations that are better adapted to survive and successfully reproduce in their environments. However, natural selection cannot produce the perfect organism. Natural selection can only prune existing variation in the population. It does not create anything from scratch. Thus, it is limited by a population's existing genetic variance and whatever new alleles arise through mutation and gene flow.

To summarize: Natural selection is the process where individuals with helpful traits are more likely to survive and reproduce than individuals without those traits. A trait is helpful if it allows an organism to take advantage of the limited resources in its environment. The genes underlying this trait are passed down from parent to child because the "fit" parents can reproduce more. For natural selection to take place, there must be variation in the genes of the species. Natural selection will choose the helpful traits that are present in the population and remove the unhelpful ones. These variations happen through mutation or sexual reproduction. Over time, the process of natural selection leads to evolution.

The warning or filler section appeared here at the end of the text

Warning. Although these ideas may seem somewhat easy to understand, most non-experts do not have a good grasp of what natural selection is or how it works. It is a complex phenomenon that takes years of study to fully comprehend. This article has only given a casual, introductory summary to the core ideas.

Filler. The theory of natural selection was proposed by Charles Darwin, but many scientists, philosophers, and researchers had debated the topic for generations. His observations on the Galapagos Islands to the west of Ecuador helped refine his theory. He detailed the different aspects of this theory in "The Origin of Species" from 1859.

Text 2: control (comets)

A comet is a relatively small chunk of icy material, a few kilometers across, that develops an atmosphere as it approaches the Sun. It can develop a tail that extends several million kilometers away from the body of the comet. While we customarily think of comets as the head and tail, the word comet refers only to the chunk of material that composes its head.

Comets have been observed from the earliest times: Accounts of comets are found in the histories of virtually all ancient civilizations. These "hairy stars" were sometimes considered transmitters of disaster or as bad omens. However, we no longer fear comets. This may be because the typical comet is not as spectacular in our skies since there is now a lot of light pollution, or because we understand more about comets than our ancestors did. Here we will review more of what comets are, how their appearance forms, how they manifest in the night sky, and how new comets can be discovered.

Like the moon and planets, comets appear to gradually shift their location in the sky. Unlike the planets, most comets appear at unsystematic times and typically remain visible for periods that vary from a couple of weeks to several months. Still images of comets give the impression that they are moving rapidly across the sky, like a shooting star. Looking only at images, it is easy to confuse comets and meteors. But they are very different when seen in person: the meteor combusts in our atmosphere and is gone in a few seconds, whereas the comet may be visible for weeks in mostly the same spot.

Only a few comets, called short-period comets, return in a time measurable in human terms. Many short-period comets have had their orbits distorted by coming too close to one of the giant planets—most often Jupiter. Aside from the short-period comets, most comets will take thousands of years to return, if they return at all.

When we look at an active comet, all we normally see is its temporary atmosphere of gas and dust, called the comet's head or *coma*. Since the gravity of their small body mass is very weak, the atmosphere is constantly falling away from the comet and must be replenished by new material, which comes from the small, solid nucleus inside the comet. It is usually hidden by the refracted light coming from the massive atmosphere surrounding it.

The water vapor and other substances that escape from the nucleus when it is heated can be detected in the comet's head and tail. Therefore, we can use scientific instruments to analyze what atoms and molecules the nucleus ice is composed of. However, we are somewhat less certain of the non-ice components. We have never identified a fragment of solid material from a comet that has survived passage into Earth's atmosphere, so the main way that we acquire samples of the comet's components is from spacecrafts. They are equipped with dust detectors that, when the spacecraft approaches a comet, take samples of what is coming off the comet. Some of this comet dust has even been returned to Earth, where we were able to analyze its contents. It seems that much of the "dirt" in comets is dark, primitive hydrocarbons and silicates. It is similar to the material we think is in dark, "dirty" asteroids.

The miraculous activity that allows us to see comets is caused by the evaporation of cometary ices when they are heated by sunlight. Beyond the asteroid belt, where comets spend most of their time, these ices are completely frozen; but, as a comet approaches the Sun, it begins to warm up and melt away. This happens for the typical comet somewhere just beyond the orbit of Mars. The evaporating water in turn releases the dust that was mixed in with the ice. Because the

comet's nucleus is so small, its gravity cannot hold either the gas or the dust in its atmosphere, so both the gas and the dust flow away into space.

A comet continues to absorb energy as it approaches the Sun. An abundance of this energy goes into heating the surface of the comet and evaporating its ice. However, contemporary observations of numerous comets indicate that this evaporation is not a consistent process. Most of the gas is released in spontaneous emissions across scarce areas of the surface, like geysers that release water periodically. Expanding into space at a speed of about 1 kilometer per second, the comet's atmosphere can expand to an enormous size: the comet's head, made of both the nucleus and the atmosphere, is often as large or larger than Jupiter.

Most comets also develop tails as they approach the Sun. A comet's tail is an extension of its atmosphere, consisting of the same gas and dust that makes up its head. As early as the sixteenth century, observers realized that comet tails always point away from the Sun, and they hypothesized that this is generally because the force of sunlight drives particles away from the head.

All around the world, amateur sky observers spend countless nights scanning the sky for new comets because astronomy is one of the very few fields of science where amateurs can still make a meaningful contribution. The discovery of a new comet is one of the most exhilarating ways that amateurs can contribute. Typical comet-watchers spend an average of 420 hours before finding a new comet.

When hunters think they have found a new comet, they must first check the object's location in an atlas to make sure it really is a comet. Since the first sighting of a comet usually occurs when it is still far from the Sun and before it has an elongated tail, it will look like a small, ambiguous patch instead of a typical comet. Other objects in space also look like indistinct patches, so they must verify that it is actually a comet. Then, they must check that they have not come across a comet that has already been discovered. If everything checks out, they contact the Central Bureau for Astronomical Telegrams to have them validate the comet. If the discovery is confirmed, the bureau will send a notification out to astronomers and observatories around the world.

In summary: Comets are the icy masses inside the structure we think of as a comet. This center is surrounded by an atmosphere that is rapidly disintegrating away into space and is replenished by the disintegration of the actual comet as it moves closer to the Sun. The tail of the comet forms as it moves toward the Sun due to the force of sunlight pushing the disintegrating atmosphere back. When they are visible from Earth, they move slowly around the night sky for a while, where they can be observed and discovered by amateur comet hunters.

Appendix B

Confidence questions

(1) How knowledgeable would you say you are about natural selection? 0 indicates you know nothing about it and 100 indicates that you are an expert in it.
(2) If someone asked you to explain what natural selection is, how well do you think you could describe the theory? 0 indicates not at all, 100 indicates that you could describe it completely.
(3) If you were quizzed on the theory of natural selection, what do you think your grade would be on that quiz (as a percentage)?
(4) Please fill in the blank: I know more about natural selection than ___% of people.
(5) On a scale of 0–100, where 0 is easy and 100 is hard, how understandable is natural selection?
(6) How confident are you in the above answers? 0 indicates not at all, 100 indicates extreme confidence.

Appendix C

Table C1. Confidence question means, SDs, and correlations, with confidence intervals, as obtained prior to reading in Experiment 1.

Variable	*M*	*SD*	1	2	3	4	5
1. How knowledgeable	47.07	23.22					
2. Ability to explain	50.25	26.61	.83** [.79, .87]				
3. Quiz grade	61.30	22.02	.76** [.70, .81]	.78** [.72, .82]			
4. Percent of people	45.21	24.43	.78** [.72, .82]	.76** [.70, .81]	.74** [.67, .79]		
5. Easy/hard	56.82	26.15	.17* [.04, .29]	.17** [.04, .29]	.10 [−.03, .23]	.10 [−.03, .23]	
6. Confidence in answers	71.61	22.19	.33** [.21, .44]	.29** [.16, .40]	.22** [.10, .34]	.22** [.10, .34]	.15* [.02, .27]

M and *SD* represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). Variable numbers align with the question numbers appearing in Appendix A.
* indicates $p < .05$. ** indicates $p < .01$.

Table C2. Confidence question means, SDs, and correlations, with confidence intervals, as obtained after reading in Experiment 1.

Variable	*M*	*SD*	1	2	3	4	5
1. How knowledgeable	57.61	21.57					
2. Ability to explain	62.79	23.19	.81** [.76, .85]				
3. Quiz grade	70.08	19.03	.79** [.73, .83]	.80** [.74, .84]			
4. Percent of people	52.72	23.32	.78** [.73, .83]	.72** [.65, .78]	.73** [.67, .79]		
5. Easy/hard	57.84	28.66	.14* [.01, .26]	.10 [−.03, .23]	.18** [.05, .30]	.05 [−.08, .18]	
6. Confidence in answers	75.38	18.95	.38** [.26, .48]	.37** [.25, .47]	.39** [.27, .49]	.29** [.16, .40]	.13* [.01, .26]

M and *SD* represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). Variable numbers align with the question numbers appearing in Appendix A.
* indicates $p < .05$. ** indicates $p < .01$.