# Learning Evolution by Collaboration

ANDREW SHTULMAN AND ANDREW G. YOUNG

*Collaboration can be an effective means of learning, but is it effective in domains where collaborators rely on conceptually distinct forms of reasoning? We explored this question in the domain of evolution, where many students construe evolution as the uniform transformation of all members of a population rather than the selective survival and reproduction of a subset. College undergraduates (n = 174) completed an assessment of their evolutionary reasoning by themselves (pretest) and with a partner (dyad test); some (n = 44) also completed an assessment several months later (posttest). Higher-scoring partners pulled up lower-scoring partners to achieve a dyad score equivalent to the higher-scoring partner's pretest score. Lower-scoring partners retained a score boost when working alone at posttest. These findings indicate that students who hold different views of evolution are able to collaborate effectively, and such collaboration yields long-term learning gains for partners with lower levels of understanding.*

*Keywords: collaboration, science learning, intuitive theories, evolutionary reasoning*

**Evolution by natural selection is one of the most** difficult topics to teach and to learn, not just in biology but in science as a whole (Shtulman 2017). It is widely mischaracterized (Gould 1996), widely mistrusted (Miller et al. 2006), and widely misunderstood (Gregory 2009, Weisberg et al. 2018), and many strategies for teaching evolution, from case studies to hands-on activities, are met with limited success at improving conceptual understanding (Legare et al. 2018). In the present article, we explore the possibility of teaching evolution through peer collaboration.

Psychological research has found that people solve problems more successfully if they work with a partner (Gauvain and Rogoff 1989, Okada and Simon 1997), and science education research has found that students learn course material more successfully if they discuss it with their peers (Crouch and Mazur 2001, Smith et al. 2009). It remains an open question, however, whether peer collaboration is useful when students hold deep-seated misconceptions about the learning domain—misconceptions that conflict with scientific ideas and interfere with instruction (Gregory 2009, Pobiner 2016). The goal of the current study is to determine whether students who hold a scientific understanding of evolution can collaborate effectively with those who hold an alternative, nonscientific understanding (Shtulman 2006, Shtulman and Calabi 2012) and whether such collaboration yields learning. Addressing these questions can inform our understanding of collaborative learning, as well as best practices in biology education.

Learning about scientific topics such as evolution and natural selection requires conceptual change, or knowledge restructuring at the level of individual concepts (Chi 1992, Carey 2009). Science involves entities, properties, and mechanisms that defy basic intuitions about how the world works, and these intuitions must be reorganized and restructured in the course of science education (Nersessian 1989, Vosniadou 1994). Students who fail to undergo conceptual change develop systematic misconceptions about the domain at hand. These misconceptions are robust in the face of counterevidence and counterinstruction, and they impede communication between those who have achieved conceptual change and those who have not (for reviews, see Carey 2009, Shtulman 2017).

Achieving an accurate understanding of evolution is impeded by several factors, including a poor understanding of randomness (Fiedler et al. 2017), teleological beliefs (Barnes et al. 2017), and belief in creationism (Weisberg et al. 2018). In the present article, we focus on obstacles to understanding the population-based logic of natural selection. From a scientific point of view, evolution is the outcome of differential survival and differential reproduction within a population; traits possessed by the most reproductively successful individuals spread through the population over time (albeit under the influence of additional factors, such as genetic drift). Most people do not view evolution in these terms. Instead, they view evolution as the uniform transformation of an entire population, where every organism is guaranteed to have offspring more adapted to the environment than it was at birth (Bishop and Anderson 1990, Shtulman 2006, Kampourakis 2014). This view is grounded in the commonsense assumption that all members of a species share the same inner nature, or *essence*, which determines their outward appearance and behavior (Gelman 2003, Shtulman and Schulz 2008). Evolution is incorrectly

*BioScience* 71: 1091–1102. 
https://doi.org/10.1093/biosci/biab089 Advance Access publication 1 September 2021

viewed as the cross-generational metamorphosis of the species (and its essence), with all organisms acquiring the traits they need to acquire in order to survive; selection plays no role in this process.

Essentialist misconceptions about evolution have been documented in people of varying ages and educational backgrounds, including children (Berti et al. 2010, Shtulman et al. 2016), adolescents (Donnelly et al. 2016, Wyner and Doherty 2017), college biology majors (Nehm and Reilly 2007, Coley and Tanner 2015), science graduate students (Brumby 1984, Gregory and Ellis 2009), preservice science teachers (Nadelson 2009, Rice and Kaya 2012), and high school biology teachers (Nehm et al. 2009, Yates and Marek 2014). These misconceptions characterize how people reason about a variety of evolutionary phenomena, from the microevolutionary phenomena of variation, inheritance, and adaptation to the macroevolutionary phenomena of domestication, speciation, and extinction (Shtulman and Calabi 2012, Shtulman 2006). Collectively, they constitute an alternative way to understand evolution, yielding coherent inferences despite their inconsistency with natural selection.

Conceptual change requires abandoning our intuitive conceptions of a domain in favor of more accurate, scientific conceptions, where ideas entailed by the former no longer make sense on the latter. Consider the phenomena of speciation and extinction. Students who hold an essentialist view of evolution reject the idea that species share common ancestors, particularly species from different phyla or kingdoms, because each species is believed to possess a distinct essence, whereas students who hold a selection-based view endorse not only the idea of common ancestry but also the idea that all species share a common ancestor (Poling and Evans 2004b, Shtulman 2006, Horn et al. 2016). Students who hold an essentialist view also reject the idea that species frequently go extinct, because species are believed to develop the traits they need in order to survive, whereas students who hold a selection-based view endorse the opposite idea—namely, that species are more likely to go extinct than to adapt to unpredictable environmental changes (Poling and Evans 2004a, Shtulman 2006).

Because intuitive conceptions of a domain support different ideas and beliefs than scientific conceptions, people who have achieved conceptual change can have difficulty conversing with people who have not. Such difficulties have been observed in conversations between children and adults (Carey 1985, Vosniadou and Brewer 1992), conversations between science students and science teachers (Reiner et al. 2000, Wiser and Amin 2001), and conversations between scientists working with different conceptual models (Dunbar 1997, Paletz et al. 2016). Impasses in communication regularly occur in the context of learning, as when preschoolers learn the properties of living things from a parent (Carey 1985) or when middle schoolers learn the properties of thermal systems from a teacher (Wiser and Amin 2001), but it is unclear how they affect learning. Conceptual change requires overcoming the conceptual gap responsible for the impasse, but how do learners navigate the impasse? How do they communicate about scientific phenomena if they understand those phenomena differently?

These questions are particularly important in light of the finding that collaboration facilitates learning. For many inductive problems, individuals are more likely to solve them—and learn from them—if they collaborate with a partner (Gauvain and Rogoff 1989, Laughlin et al. 1991, Leman et al. 2016). Collaboration, or the construction of a shared understanding through active communication, can be effective for several reasons. It opens partners' eyes to ideas they would not have generated on their own, revealing alternative approaches to the same problem (Schwarz et al. 2000, Young et al. 2012) or alternative explanations for the same phenomenon (Ames and Murray 1982, Howe 2009). It forces collaborators to articulate their reasons for endorsing a particular hypothesis or favoring a particular solution and defend those reasons with evidence (Teasley 1995, Okada and Simon 1997). And it introduces social incentives for completing the task at hand, increasing partners' persistence in the face of obstacles (Butler and Walton 2013).

Given the pedagogical benefits of collaboration, we sought to determine whether collaboration is useful in a domain requiring conceptual change (evolution). This question has both practical and theoretical significance. From a practical point of view, educators who instruct students on complex scientific topics—including not just evolution but also other complex topics such as microbiology (Au et al. 2008) and genetics (Duncan et al. 2009)—would benefit from knowing whether peer collaboration is likely to be productive or counterproductive. From a theoretical point of view, models of conceptual change would be informed by clarifying how this process is best achieved. Some forms of collaboration, such as parent–child conversation, have proven successful at improving conceptual understanding (Jipson and Callanan 2003, Gunderson and Levine 2011), but so have other activities, including refutation-based instruction (Asterhan and Resnick 2020), inquiry-based instruction (Sandoval and Reiser 2004), and extended case studies (Kelemen et al. 2014). Is collaboration a reliable means of facilitating conceptual change?

Previous studies suggest yes. Asterhan and Schwarz (2007, 2009) found that undergraduates who collaborated on explaining two instances of adaptation (mosquitos developing resistance to insecticide and cheetahs acquiring the ability to outrun other mammals) provided more selection-based explanations following collaboration. Loyens and colleagues (2015) found that undergraduates who collaborated on determining the paths of three projectiles (a child jumping from a swing, an object falling on someone's head, and a coyote falling from a cliff) drew more accurate paths following collaboration. These studies involved domains in which learning a correct, scientific understanding is difficult to achieve, but they are limited in that they explore a single facet of the domain (adaptation and free fall, respectively) and they provided explicit guidance on how participants

should collaborate. In the present article, we explore the effects of collaboration without guidance and across several facets of the target domain: variation, inheritance, adaptation, domestication, speciation, and extinction. This breadth of topics provides a more comprehensive assessment of how collaboration might facilitate conceptual understanding, as well as the frequency and scope of those benefits.

We assessed the effects of collaboration both in the short term and the long term. Our short-term assessment was a comparison of how participants reasoned about evolution when working alone and when working with a partner, as a dyad. We expected dyads to perform more accurately than individuals, but how much more accurately was an open question. Although dyad members with higher levels of understanding could demonstrate accurate reasoning, their partners might not be persuaded by that reasoning or might explicitly reject it. Partners could also help generate new ideas—ideas that neither dyad member had generated on their own (Smith et al. 2009). Alternatively, dyad members with higher levels of understanding could be swayed by their partner to accept inaccurate reasoning. We explored these possibilities by comparing participants' responses to a comprehension assessment before and during collaboration, as well as by analyzing what dyad members said to each other when collaborating.

Our long-term assessment of the effects of collaboration was a comparison of how participants reasoned about evolution prior to collaboration and then several months later. We expected participants to retain some of the conceptual gains they made when working with a partner, but it was an open question how much they would retain and whether both partners would retain similar amounts. Although partners with lower levels of understanding might demonstrate reliable gains, partners with higher levels of understanding could demonstrate reliable losses, if they abandoned their reasoning strategies in favor of their partners'.

## Method

Our study was conducted at Occidental College and was approved by the Occidental Institutional Review Board, as proposal Shtu-D12069.

**Participants.** The participants were 174 college undergraduates, recruited from introductory psychology and cognitive science courses and compensated with extra credit or a small stipend, depending on their preference. The participants had most likely taken biology in high school, and some may have taken biology in college, but we did not ask them to report the number or content of those classes. The participants completed the study in pairs, forming a total of 87 dyads.

The dyads were created by convenience, not pretesting; participants who signed up to complete the study at the same time were partnered. Some participants knew their partner prior to participation, and some did not, but familiarity with one's partner did not influence participants' performance on the task. Participants reported their familiarity on a scale from 1 (not at all familiar) to 5 (very familiar), and these ratings did not correlate with their scores on the evolution comprehension assessment, either when working alone or when working together. Familiarity ratings did not correlate with changes in scores either, indicating that familiar partners were no more likely to benefit from collaboration than unfamiliar partners.

All participants were invited to complete a posttest (for an Amazon gift card), but only 44 did so. Those 44 came from 36 different dyads. Approximately half were the higher-scoring partner in their dyad ($n$ = 25) and half were the lower-scoring partner ($n$ = 19). Details about the posttest sample are presented below, in our analysis of long-term effects of collaboration.

**Materials.** Participants were assessed on their understanding of evolution using an instrument developed by Shtulman (2006). We chose this assessment because each question is designed to differentiate correct, selection-based reasoning from incorrect, essentialist reasoning. Other assessments, such as those developed by Anderson and colleagues (2002) or Rutledge and Warden (2000), measure scientific knowledge of evolution but do not diagnose alternative views of evolution. The questions developed by Shtulman (2006) have proven successful at distinguishing essentialist views from selection-based ones in middle school students (Coley et al. 2017), college biology majors (Shtulman and Calabi 2013), college students from other majors (Nettle 2010, Sota 2012, Heddy and Sinatra 2013, Asterhan and Resnick 2020), noncollege adults (Shtulman and Schulz 2008), and high school biology teachers (Furtak 2012). Its validity has been confirmed with professional biologists (Shtulman 2006), and its reliability has been confirmed with studies showing that students tend to achieve the same score across multiple administrations (Heddy and Sinatra 2013), even after having taken a full semester of biology (Shtulman and Calabi 2013).

The assessment consisted of six sections, each devoted to a different biological phenomenon: inheritance, variation, adaptation, domestication, speciation, or extinction. Participants' understanding of the phenomenon was assessed with five questions designed to distinguish essentialist and selection-based interpretations, as is described in table 1. All questions required a closed-ended (multiple-choice) response, but many required an open-ended (self-generated) response as well. These responses were used to determine whether participants chose the right multiple-choice option for the wrong reason (scored as incorrect) or the wrong option for an acceptable reason (scored as correct). The full battery can be found in the appendix of Shtulman (2006), along with a scoring rubric.

As an illustration, consider this question about parent–offspring inheritance: "Imagine that biologists discover a new species of woodpecker that lives in isolation on a secluded island. These woodpeckers have, on average, a

***Table 1. Correct, selection-based interpretations of the topics on the evolution comprehension assessment and incorrect, essentialist ones.***

| Topic | Score | Interpretation |
|---|---|---|
| Variation | + | Individual differences are fodder for selection. |
| | – | Individual differences are minor and nonadaptive. |
| Inheritance | + | Differences between parents and offspring are random and unpredictable. |
| | – | Differences between parents and offspring are adaptive and purposeful. |
| Adaptation | + | Adaptation results from differential survival and reproduction. |
| | – | Adaptation results from widespread mutations in response to need. |
| Domestication | + | Species are domesticated by selective breeding. |
| | – | Species are domesticated by changing individual organisms. |
| Speciation | + | New species emerge when two populations diverge. |
| | – | New species emerge when one population transforms into another. |
| Extinction | + | Extinction is more common than adaptation. |
| | – | Adaptation is more common than extinction. |

*Note:* Correct responses received positive scores (+), and incorrect responses received negative scores (–).

one-inch beak and their only food source is a tree-dwelling insect that lives, on average, one-and-a-half inches under the tree bark. Compared to its parents, the offspring of any two woodpeckers should develop: (a) a longer beak, (b) a shorter beak, or (c) either a longer beak or a shorter beak; neither is more likely." The correct response is (c), because offspring vary randomly from their parents, but the lure response is (a), based on the idea that offspring inherit the traits they need to inherit to survive—traits the species, as a whole, acquires as its essence adapts to the changing environment.

Or consider this task designed to probe participants' understanding of within-species variation: "During the 19th century, England's native moth species, *Biston betularia*, evolved darker coloration in response to the pollution produced by the Industrial Revolution. Imagine that biologists gathered a random sample of *Biston betularia* once every 25 years from 1800 to 1900. What range of coloration would you expect to find at each point in time?" Participants were given a five-by-five matrix of moth outlines and instructed to shade the moths by selecting a color ranging from white to dark gray to reflect how the moths might look at 1800, 1825, 1850, 1875, and 1900.

The two most common response patterns are depicted in figure 1. The pattern on the left depicts a mutation for darker coloration spreading through the population over time and is consistent with a selection-based view of evolution. The pattern on the right depicts a holistic transformation of the population, such that variation occurs between generations but not within generations, and is consistent with an essentialist view. The shading patterns were coded by quantifying the amount of variation depicted within and across generations. Patterns that depicted more variation across generations than within were scored as essentialist, whereas patterns that depicted similar amounts of variation were scored as selection-based.

**Coding.** Participants provided a total of 9150 responses: 30 for each of 174 pretests, 87 dyad tests, and 44 posttests. Responses that revealed correct, selection-based reasoning were assigned 1 point; responses that revealed incorrect, essentialist reasoning were assigned –1 point; and responses too vague to be coded one way or the other were assigned 0 points. Participants answered five questions for each of six sections, so their composite scores could range from –30 to 30. In actuality, they ranged from –20 to 27.

Assessments were scored on a three-point scale to reflect the conceptual distinctions between essentialist and selection-based responses. Coding essentialist responses as the opposite of selection-based ones (as opposed to their absence) yields a scoring continuum from pure essentialist reasoning to mixed reasoning to pure selection-based reasoning and allows us to assess how strongly a participant relies on one form of reasoning or the other. This benefit comes with a cost, however, in that it obscures the total number of correct responses provided. A participant could answer half the assessment correctly but still earn a score of 0 if they provided essentialist responses on the other half. If what matters for successful collaboration is that at least one collaborator answered the question correctly, then our coding strategy may underestimate a participant's collaborative potential. For this reason, we coded the data twice: once using the trichotomous scheme described above and once using the dichotomous scheme of assigning 1 point to every selection-based answer and 0 points to all other answers, essentialist or ambiguous. Below, we report analyses using both coding schemes and found no differences between them.

In addition to coding assessment responses, we coded conversations that occurred between dyad members as they completed the dyad test. Their conversations were audio recorded and later transcribed. We reviewed the transcripts for evidence that dyad members forged a shared understanding of the task through six communicative activities: proposing an idea, asking a question, offering an explanation, expressing agreement with one's partner, expressing disagreement, and expressing uncertainty. This range of activities allowed us to assess whether partners were exchanging (and constructing) ideas communally or whether one partner was dominating or directing the conversation. The number of ideas proposed was the clearest indicator of whether one partner dominated the conversation, but we also coded for expressions of agreement, disagreement, and uncertainty

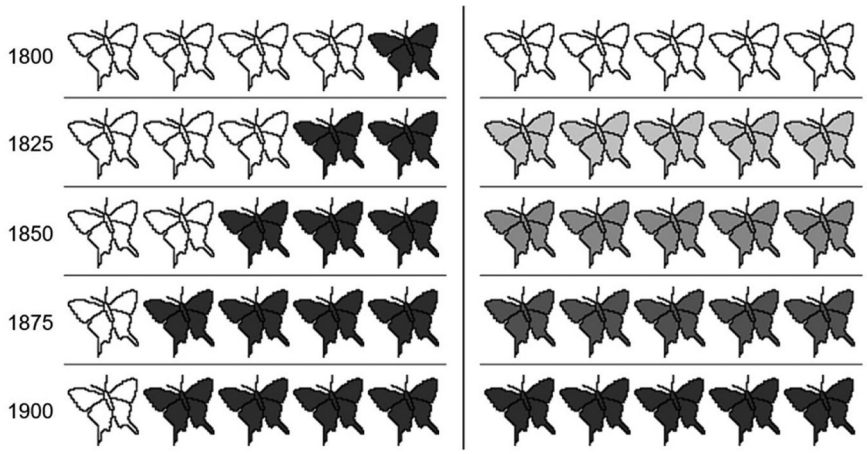


***Figure 1. A selection-based response pattern (left) and an essentialist response pattern (right) on the moth-shading task of the evolution comprehension assessment.***

to assess how partners reacted to these proposals—whether one partner uniformly accepted the other's proposals or whether both partners accepted (and challenged) proposals in equal measure. We further coded for questions and explanations to gauge whether partners merely proposed ideas or also evaluated and elaborated on those ideas (Callanan et al. 1995).

Four coders reviewed the transcripts in pairs, reading them from the perspective of a particular dyad member and assigning each utterance a code (if applicable). Coders were blind to the scoring status of each dyad member, as well as the score they achieved together. Although 87 dyads were tested, only 73 conversations were recorded because of experimenter error or equipment failure. These 73 conversations averaged 31 minutes in length and contained an average of 3338 words.

**Procedure.** The evolution comprehension assessment was administered on a computer and took between 30 and 45 minutes to complete. Participants were tested in pairs in a room in the Psychology Department. They completed the pretest by themselves, and they completed the dyad test together immediately following the pretest. Participants were given no instruction on how to coordinate their responses; they were simply asked to complete the survey as a pair, on a single computer. The posttest was administered one semester (i.e., half a year) after the collaboration session. Participants completed the same assessment at pretest, dyad test, and posttest. Although administering the same assessment introduces the risk of practice effects, we did not provide participants with feedback, so changes in performance from one administration to the next reflect the perception that an earlier answer was incorrect rather than externally verified knowledge of the correct answer. The misconceptions tapped by the assessment are also difficult to correct, even with direct instruction (Shtulman and Calabi 2013).

## Results

We analyze the effects of collaboration at two timescales: short-term effects, or how assessment scores changed from pretest to dyad test (completed during collaboration), and long-term effects, or how assessment scores changed from pretest to posttest (completed several months after collaboration). With respect to the latter, we also explore how participants' scores were influenced by their partners' scores.

**Short-term effects of collaboration.** Comprehension assessment scores prior to collaboration (pretest) and during collaboration (dyad test) are displayed in figure 2. Pretest scores are displayed with respect to whether participants were the lower-scoring partner in their dyad or the higher-scoring partner. Paired samples *t*-tests reveal that lower-scoring partners improved their score from pretest to dyad test ($t(86) = 11.36$, $p < .001$), whereas higher-scoring partners maintained similar scores ($t(86) = 1.74$, $p = .085$). The same finding was obtained when assessments were scored dichotomously (lower scorers, $t(86) = 11.08$, $p < .001$; higher scorers, $t(86) = 1.82$, $p = .073$). These results indicate that higher-scoring partners pulled up lower-scoring ones to the level of performance the former had demonstrated prior to collaboration. This pattern of dyads outscoring the lower-scoring partner was observed for 75 of the 87 dyads, or 86% of the sample.

Scores alone cannot reveal whether partners were truly collaborating or whether higher scorers were dictating correct answers to their lower-scoring partners. We addressed this concern in two ways. First, we explored the source of correct responses on the dyad test. If higher scorers dominated the conversation or if lower scorers regularly deferred to higher scorers, then all questions answered correctly on the dyad test should have been answered correctly by the higher scorer on the pretest (or by both partners). In reality, nearly a quarter of these questions were *not* answered correctly by the higher scorer; 12% were answered correctly only by the lower scorer on the pretest, and 12% were answered correctly by neither partner on the pretest. Both frequencies are significantly greater than 0 (correct by lower scorer, $t(86) = 11.10$, $p < .001$; correct by neither, $t(86) = 10.80$, $p < .001$), indicating that dyads' responses were not a mere copy of the higher scorers' responses. Some correct ideas were contributed by the lower scorer, and some were constructed during the act of collaboration (similar to what was observed by Smith et al. 2009).

Second, we explored the dynamics of the collaboration in terms of how often dyad members proposed ideas, asked questions, provided explanations, and expressed agreement, disagreement, or uncertainty toward their partner.

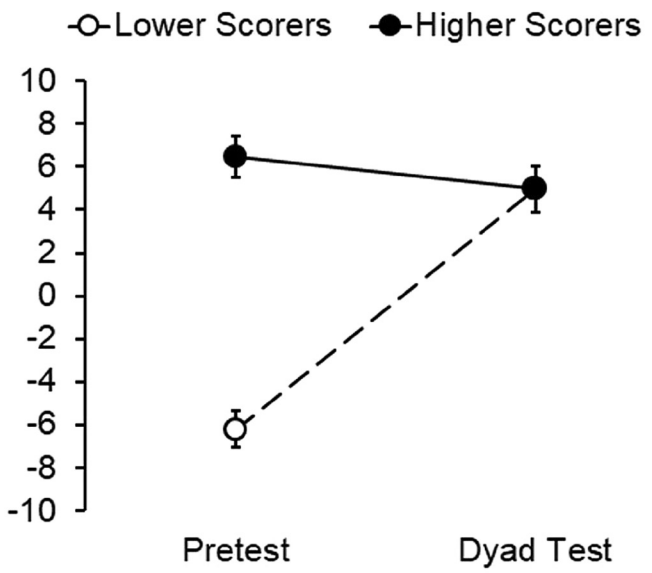


***Figure 2. Mean evolution scores on the pretest and the dyad test by whether the participants were the lower scorer (n = 87) or the higher scorer (n = 87) in their dyad. Error bars represent the standard error.***

If higher scorers dominated the collaboration, we would expect them to propose more ideas, provide more explanations, and express more disagreement with their partner. Lower scorers, on the other hand, should ask more questions and express more agreement or uncertainty. None of these differences were observed (see table 2). Instead, higher scorers and lower scorers contributed equally to the conversation. Partners not only uttered a similar number of words ($t(72) = 0.78$, $p = .451$) but also articulated a similar number of ideas ($t(72) = 0.90$, $p = .370$), questions ($t(72) = 1.03$, $p = .305$), and explanations ($t(72) = 0.49$, $p = .628$), and expressed a similar amount of agreement ($t(72) = 0.60$, $p = .548$), disagreement ($t(72) = 0.64$, $p = .526$), and uncertainty ($t(72) = 0.57$, $p = .572$).

There were, however, conversational asymmetries within the dyads, and these asymmetries were predicted by scoring differences between partners on the pretest. The greater the scoring difference, the more lower scorers asked questions relative to higher scorers ($r = .26$, $p = .028$) and the more higher scorers provided explanations relative to lower scorers ($r = .25$, $p = .033$). All together, these findings confirm that dyad conversations were two-sided, while also suggesting that differences in understanding between dyad members changed the dynamic of their explanatory activities. We shall return to the latter finding, in light of how scoring differences on the pretest affected learning gains from pretest to posttest.

**Long-term effects of collaboration.** A subset of participants ($n = 44$) completed a posttest several months later. Although there may have been motivational differences between those who opted to complete a posttest and those who did not, their pretests revealed no differences in understanding. The 19 lower-scoring partners who completed a posttest scored similarly to the 68 who did not (mean [$M$] = –7.3 versus $M = –6.2$, $t(85) = 0.54$, $p = .590$), and the 25 higher-scoring partners who completed a posttest scored similarly to the 62 who did not ($M = 7.1$ versus $M = 6.3$, $t(85) = 0.38$, $p = .705$). The average delay was 7.6 months (standard deviation [SD] = 4.3 months), and the delay for higher-scoring partners was equivalent to the delay for lower-scoring partners ($M = 6.8$ versus $M = 8.8$, $t(42) = 1.13$, $p = .14$).

Comprehension assessment scores for the posttest sample are displayed in figure 3. Mean dyad scores for lower-scoring partners were not equivalent to those for higher-scoring partners, as they are in figure 2, because participants in this smaller sample generally came from different dyads. Still, the same patterns are apparent: dyads scored higher than lower-scoring partners did on their own but no higher than higher-scoring partners did.

We analyzed assessment scores for effects of scoring status (higher versus lower within a dyad) and assessment period (pretest versus dyad test versus posttest). A repeated-measures ANOVA revealed a significant effect of assessment period ($F(2,84) = 5.53$, $p = .006$, $\eta_p^2 = .12$) qualified by a strong interaction with scoring status ($F(2,84) = 11.08$, $p < .001$, $\eta_p^2 = .21$). Follow-up analyses revealed that lower-scoring partners achieved higher scores on the posttest than on the pretest ($t(18) = 4.15$, $p < .001$), but higher-scoring partners achieved similar scores ($t(24) = 1.45$, $p = .159$). Neither group exhibited a change from dyad test to posttest (lower scorers, $t(18) = 1.30$, $p = .209$; higher scorers, $t(24) = 0.27$, $p = .792$). The interaction between scoring status and assessment period held when the assessments were scored dichotomously ($F(2,84) = 10.80$, $p < .001$, $\eta_p^2 = .21$), with lower-scoring partners demonstrating significant pre-post gains ($t(18) = 4.08$, $p < .001$) and higher-scoring partners demonstrating neither gains nor losses ($t(24) = 1.75$, $p = .094$). These results suggest that lower-scoring partners increased their understanding of evolution as a result of collaboration, whereas higher-scoring partners retained the same level of understanding.

Might lower scorers have performed better on the posttest because they simply memorized the answers provided by their higher-scoring partners? Two additional analyses suggest not. First, we observed no correlation between how much time had passed from pretest to posttest (in days) and how many points participants gained from pretest to posttest ($r = .15$, $p = .331$), whereas a negative correlation would be expected if correct responses were recalled verbatim from memory, which would have faded with time. Memory for episodic details of the collaboration would have been sparse for *all* participants, given that posttests were administered several months later.

Second, we observed the same pre-post gains for questions that required self-generated descriptions or justifications, as is shown in figure 4. These 14 questions provide a more stringent test of whether lower

*Table 2. Mean conversational patterns of higher and lower scorers, along with correlations between conversational asymmetries and pretest asymmetries between higher and lower scorers.*

| Measure | Higher scorer | Lower scorer | Difference | Pretest correlation |
|---|---|---|---|---|
| Words uttered | 1635.8 | 1702.3 | –66.5 | .16 |
| Ideas proposed | 42.9 | 41.1 | 1.8 | .12 |
| Questions asked | 21.5 | 23.3 | –1.8 | –.26* |
| Explanations offered | 38.0 | 39.0 | –1.0 | .25* |
| Expressions of agreement | 32.4 | 31.3 | 1.2 | –.11 |
| Expressions of disagreement | 4.0 | 3.6 | 0.5 | .01 |
| Expressions of uncertainty | 14.1 | 14.6 | –0.5 | –.15 |

*$p < .05$.

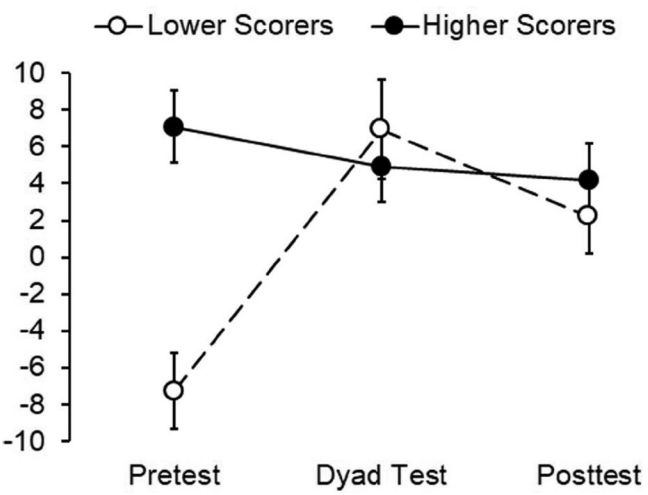


*Figure 3. Mean evolution scores on the pretest, the dyad test, and the posttest for lower scorers (n = 19) and higher scorers (n = 25) who completed the posttest. Error bars represent the standard error.*

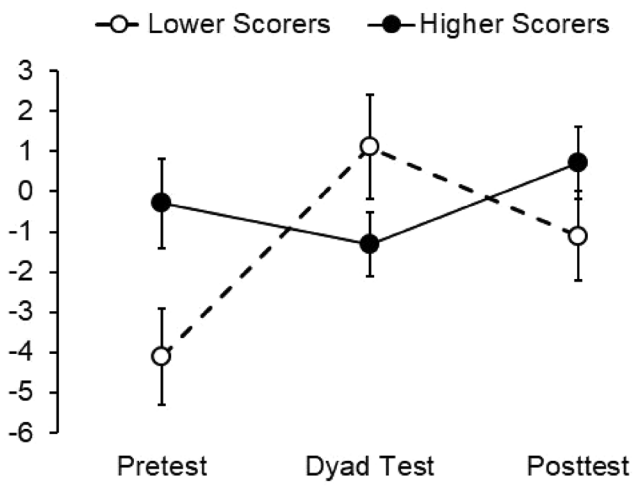


*Figure 4. Mean scores on just the open-ended questions on the pretest, the dyad test, and the posttest for lower scorers (n = 19) and higher scorers (n = 25) who completed the posttest. Error bars represent the standard error.*

scorers learned from collaboration, because they had to articulate selection-based reasoning to be scored as correct. A repeated-measures ANOVA for scores on the open-ended questions alone revealed a significant effect of assessment period ($F(2,84) = 5.65$, $p = .005$, $\eta_p^2 = .12$) qualified by a strong interaction with scoring status ($F(2,84) = 13.20$, $p < .001$, $\eta_p^2 = .24$). The interaction reflects a significant pre-post gain for the lower scorers ($t(18) = 2.29$, $p = .034$) but no concomitant gain for the higher scorers ($t(24) = 1.51$, $p = .144$). The same results were found when responses were scored dichotomously: a marginal effect of assessment period ($F(2,84) = 2.76$, $p = .069$, $\eta_p^2 = .06$) qualified by an interaction with scoring status ($F(2,84) = 8.91$, $p < .001$, $\eta_p^2 = .18$), with lower scorers increasing their score from pretest to posttest ($t(18) = 2.30$, $p = .034$) but higher scorers earning similar scores ($t(24) = 0.57$, $p = .577$). These findings indicate that lower scorers improved their performance on questions where they could not simply have memorized the correct response option; those options had to be justified with appropriate, selection-based reasoning.

The results presented thus far indicate that lower-scoring partners benefited from collaborating with a higher-scoring partner, but does the nature of the pairing matter? Do lower scorers learn more from partners with higher levels of understanding? We explored this possibility by comparing participants' pre-post gains to their partner's pretest score. For higher-scoring partners, the two measures were uncorrelated ($r = –.10$, $p = .526$), but for lower-scoring partners, the two measures were negatively correlated ($r = –.66$, $p < .001$), indicating that lower-scoring partners made *fewer* gains when collaborating with partners with much higher scores. These patterns remained the same when assessments were scored dichotomously (lower scorers, $r = –.55$, $p < .001$; higher scorers, $r = –.09$, $p = .573$). It would appear that lower scorers learned more from partners with moderately higher levels of understanding than from partners with substantially higher levels.

## Discussion

Collaboration can be an effective and efficient means of devising new hypotheses (Okada and Simon 1997) and learning new problem-solving strategies (Schwarz et al. 2000), but how effective is collaboration when individuals apply distinct reasoning strategies to the task at hand? In the present article, we observed students collaborate on tasks within the domain of evolutionary biology—a domain in which misconceptions are as common as correct conceptions, if not more common (Shtulman 2006)—and we found that collaborators reasoned more accurately together than alone. Partners who employed different reasoning strategies on their own were able to discern whose reasoning was more accurate when working together. On occasion, partners even generated accurate ideas that neither had generated by themselves.

Collaboration not only facilitated more accurate responding, but it also facilitated learning. Individuals who entered the collaboration with lower levels of understanding demonstrated increased understanding at posttest several months later. From working alone to working with a partner, lower-scoring individuals gained an average of 11.2 points on the comprehension assessment (*SD* = 9.2). Those who took the posttest retained the majority of those points, scoring an average of 9.5 points higher than they had on the pretest (*SD* = 9.9). Pre-post gains of this magnitude (Cohen's *d* = .97) are unusually strong for studies on evolution education, which typically document smaller gains (see Legare et al. 2018 for a review).

As a point of comparison, consider the findings from Shtulman and Calabi (2013), who administered the same comprehension assessment to 291 students enrolled in one of six university courses on evolutionary biology, ranging from introductory courses, such as BIO 102: Evolutionary Biology, to advanced courses, such as BIO 352: Evolution. Prior to instruction, the students earned an average score of –8.4 (*SD* = 11.3). Following instruction, they earned an average score of –5.6 (*SD* = 11.6). A full semester of college-level biology proved largely ineffective at helping the students understand evolution as a selection-based process (Cohen's *d* = 0.24), which suggests not only that collaboration may be a particularly effective learning experience but also that pre-post gains in the present study are unlikely to be an artifact of multiple testing.

One important difference between our study and the study by Shtulman and Calabi is that the students in our study received feedback on the assessment by way of collaborating with a partner—a partner who occasionally provided different responses to the same questions. Still, none of these responses were labeled as correct, and students had to decide for themselves which response to endorse in cases of disagreement. They did so not by fiat, with the higher-scoring partner dictating answers to the lower-scoring one, but by engaging in activities that facilitated a mutual recognition of accurate reasoning, with both partners proposing ideas, asking questions, and providing explanations.

These findings complement findings from other areas of science education, where students learned more from peer collaboration than standard, lecture-based instruction (Crouch and Mazur 2001, Knight and Wood 2005, Smith et al. 2009). When peers are asked to "think, pair, and share," they consistently outperform students who tackle the material on their own, both at the moment of collaboration (Smith et al. 2009) and over the full course of instruction (Crouch and Mazur 2001). Our results confirm this pattern and also extend it, by showing that collaboration is fruitful even when partners approach the same question from conceptually distinct viewpoints.

Two trivial explanations for why the lower-scoring partners benefited from collaboration can be ruled out by the data. First, it was not the case that higher-scoring partners simply told lower-scoring partners the correct answer. Analyses of the partners' conversations show that they were two sided, with both partners contributing a similar number of ideas, explanations, and questions. Furthermore, nearly a quarter of the questions answered correctly on the dyad test were not answered correctly by the higher-scoring partner on the pretest, indicating that the correct answer was either proposed by the lower-scoring partner or was discovered through the act of collaboration.

Second, it was not the case that lower scorers improved their score from pretest to posttest by simply memorizing the responses provided by their partner. Pre-post gains were uncorrelated with how much time had passed between collaboration and posttest. If lower scorers improved their score because they were able to recall their partners' answers, then their performance should have worsened with longer delays. More importantly, pre-post gains for lower-scoring partners were observed for questions that required a self-generated justification or description, where their responses would have been scored as incorrect if they were unable to articulate a selection-based rationale for making them. A purely memory-based explanation is also implausible in light of research on refutation-based instruction (Kendeou and Van Den Broek 2005, 2007, Lewandowsky et al. 2012). Refutation of incorrect ideas succeeds only if students understand the refutation; students who do not will revert back to their original ideas later on. Because lower scorers provided correct responses several months after collaboration, it is likely they understood why those responses were correct and were not simply recalling them verbatim.

Might the pre-post gains be due to information learned outside the collaboration? None of the participants who completed the posttest were biology majors, and, therefore, none were likely to have taken biology courses in the intervening months. Posttests also happen to have been administered at the beginning of the fall semester, so most of the time between pretest and posttest fell over the summer. Although participants may have encountered information that improved their understanding of evolution independent

of the collaboration, there is no reason to think that only lower scorers would encounter such information, but only lower scorers exhibited pre-post gains. This asymmetry implies that participants' status within the collaboration is what determined learning rather than some factor external to it.

The finding that higher scorers demonstrated no reliable gains in learning confirms previous research indicating that partners with lower levels of understanding benefit more from collaboration than those with higher levels (Murray 1972, Miller and Brownell 1975, Radziszewska and Rogoff 1991, Neugebauer et al. 2016). Dyads did sometimes generate more accurate responses than higher-scoring partners had generated on their own, but higher-scoring partners showed no evidence of learning from this interaction. It's possible that higher-scoring partners learned from the interaction in ways that were not detectable by the assessment, such as consolidating ideas that were previously fragmented or connecting ideas that were previously isolated. It's also possible that many higher-scoring partners would have benefited from collaboration if they had been partnered with someone who had an even higher score than their own. The mean pretest score for higher-scoring partners was 6.5 (out of a possible 30), and only a quarter answered more than half of the pretest questions correctly, suggesting that most could still have benefited from collaboration. Whether the benefits of collaboration diminish as partners' collective understanding increases is an open question, particularly when both partners appear to have crossed the threshold from relying on essentialist reasoning to relying on selection-based reasoning.

One of the more provocative findings was that lower scorers learned more from partners with moderately higher scores than from partners with substantially higher scores. Those who collaborated with partners near the extreme side of the scoring spectrum (30) benefited the least, at least on the posttest. This finding, however tentative, may have resulted from communication errors; the greater the discrepancy between partners' understanding of the domain, the more likely they encountered impasses in communication, and the more strained their collaboration may have become. Consider the following conversation between a participant who earned a pretest score of 19 (P1) and a participant who earned a pretest score of 2 (P2) about the woodpecker question presented above:

**P1.** "Alright, for the first one, I put either a shorter or longer beak, because it says 'compared to its parents,' and compared to its parents, it pretty much has the same beak because it has the same genes."

**P2.** "Okay. Hmm. I put longer beak… because… Yeah, they have to eventually evolve into the thing, but I can see what you are saying about, like, it wouldn't take one generation."

**P1.** "Well… the next generation would end up with a longer beak, but this one particular woodpecker would have the same [beak] as its parents, if you understand what I'm saying. The generations would get longer beaks because the ones with the shorter beaks will be killed off. [But] no matter what, the offspring are gonna have beaks pretty much the same as [their] parents."

**P2.** "Okay, I see what you're saying. Yeah, I guess I just assumed that they would interbreed or they would have a woodpecker from a different… Okay, I see what you're saying."

P2 claims to understand what P1 is saying, but P2's attempts to resolve the discrepancy, by acknowledging that "it wouldn't take one generation" and that birds with different beak lengths did not "interbreed," do not actually address P2's argument that evolutionary change occurs at the level of the population, not the individual. This type of impasse may be more common between partners with discrepant levels of understanding than between partners with similar levels of understanding, because the latter may be better able to recognize the source of their disagreement and then use that recognition as a stepping stone for improving their collective understanding.

In support of this possibility, we found that partners with greater score differences on the pretest exhibited greater asymmetries in how often the lower scorer asked questions (relative to the higher scorer) and how often the higher scorer provided explanations (relative to the lower scorer). Collaboration between partners with grossly different levels (or kinds) of understanding may involve too many instances in which higher-scoring partners propose ideas that lower-scoring partners do not understand. Lower-scoring partners then ask for clarification, and higher-scoring partners provide explanations, but those explanations may not be adequate to bridge the gap in understanding that initiated the exchange. Indeed, wide gaps in understanding may be the reason lecture-based instruction often proves inferior to peer collaboration (Crouch and Mazur 2001, Knight and Wood 2005, Smith et al. 2009), if the gap between instructors and students yields more communication errors than the more modest gaps among students.

Previous research on how domain experts converse with domain novices suggests that experts supply novices with specialized knowledge, in the moment, by adjusting how they label or describe objects of shared attention (Isaacs and Clark 1987, Clark and Schaefer 1989). Such studies have involved domains in which the difference between novices and experts is more quantitative than qualitative, such as differences in how much they know about New York City landmarks, and it remains an open question how experts adjust their discourse patterns when the domain of expertise entails conceptual change. Because science learning typically involves some form of conceptual change (Vosniadou 1994, Shtulman 2017), science instructors could maximize the effects of collaboration by pairing students with partners

who understand the topic better than they do but not too much better. Collaborating with the person who happens to be sitting next to you may be helpful, but collaborating with a peer matched for moderate differences in understanding may be optimal.

There is, however, another interpretation of the finding that lower scorers benefit more from collaborating with moderate-knowledge peers, which is that moderate-knowledge peers may be better at explaining evolutionary concepts than high-knowledge peers. In the current study, these two factors were confounded: The higher the pretest score of the higher-scoring partner, the larger the scoring difference between the partners. A more controlled design is needed to tease apart the effects of collaborating with a high-knowledge partner and a partner with higher knowledge. Because participants were paired by convenience, as would typically happen in a classroom, we did not control for absolute levels of understanding, but future research could partner participants to create dyads that vary in absolute but not relative understanding (by partnering participants with the same score difference regardless of where they fall on the scoring continuum) and dyads that vary in relative but not absolute understanding (by partnering participants on opposite sides of the scoring continuum regardless of their particular scores). Such research could shed light on whether partners' absolute levels of understanding matter more or less than the difference between them.

In conclusion, we found that students who reasoned about evolution in conceptually distinct ways were able to communicate across this divide and determine which partners' ideas were more accurate. Partners' made this determination without feedback, without instruction, and without reference to additional materials. Sharing and evaluating different reasoning strategies appeared to be sufficient. Partners who held more accurate views of evolution prior to collaborating did not benefit from the collaboration, but neither were they harmed by it. Partners who held less accurate views, on the other hand, showed robust improvements in understanding, both in the short term and in the long term. For these partners, an hour of collaboration yielded greater gains in conceptual understanding than typically achieved by a semester (or more) of lecture-based instruction (Bishop and Anderson 1990, Demastes et al. 1995, Jensen and Finley 1996, Shtulman and Calabi 2013), implying that collaboration may be a particularly useful tool for facilitating conceptual change. Collaboration forces individuals to confront and address their misconceptions in ways that direct instruction may not, although future research is needed to determine the conditions under which collaboration is most productive and whether that productivity extends to other conceptually complex domains.

## Acknowledgments

This research was supported by National Science Foundation grant DRL-0953384 and by an Understanding Human Cognition Scholars Award from the James S. McDonnell Foundation to Andrew Shtulman. We would like to thank Valerie Bourassa, Cameron Gillis, Tiffany Kim, Jai Levin, Lisa Matsukata, Evan Thomas, Debra Skinner, Maahir Uttam, and Shannon Xu for their assistance with data collection and data analysis.

## References cited

Ames GJ, Murray FB. 1982. When two wrongs make a right. Developmental Psychology 18: 894–897.

Anderson DL, Fisher KM, Norman GJ. 2002. Development and evaluation of the conceptual inventory of natural selection. Journal of Research in Science Teaching 39: 952–978.

Asterhan CS, Resnick MS. 2020. Refutation texts and argumentation for conceptual change: A winning or a redundant combination? Learning and Instruction 65: 101265.

Asterhan CS, Schwarz BB. 2007. The effects of monological and dialogical argumentation on concept learning in evolutionary theory. Journal of Educational Psychology 99: 626–639.

Asterhan CS, Schwarz BB. 2009. Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. Cognitive Science 33: 374–400.

Au TKF, Chan CK, Chan TK, Cheung MW, Ho JY, Ip GW. 2008. Folkbiology meets microbiology: A study of conceptual and behavioral change. Cognitive Psychology 57: 1–19.

Barnes ME, Evans EM, Hazel A, Brownell SE, Nesse RM. 2017. Teleological reasoning, not acceptance of evolution, impacts students' ability to learn natural selection. Evolution: Education and Outreach 10: 1–12.

Berti AE, Toneatti L, Rosati V. 2010. Children's conceptions about the origin of species: A study of Italian children's conceptions with and without instruction. Journal of the Learning Sciences 19: 506–538.

Bishop B, Anderson CA. 1990. Student conceptions of natural selection and its role in evolution. Journal of Research in Science Teaching 27: 415–427.

Brumby MN. 1984. Misconceptions about the concept of natural selection by medical biology students. Science Education 68: 493–503.

Butler LP, Walton GM. 2013. The opportunity to collaborate increases preschoolers' motivation for challenging tasks. Journal of Experimental Child Psychology 116: 953–961.

Callanan MA, Shrager J, Moore JL. 1995. Parent–child collaborative explanations: Methods of identification and analysis. Journal of the Learning Sciences 4: 105–129.

Carey S. 1985. Conceptual Change in Childhood. MIT Press.

Carey S. 2009. The Origin of Concepts. Oxford University Press.

Chi M. 1992. Conceptual change within and across ontological categories. Pages 129–186 in Giere R, ed. Cognitive Models of Science. University of Minnesota Press.

Clark HH, Schaefer EF. 1989. Contributing to discourse. Cognitive Science 13: 259–294.

Coley JD, Arenson M, Xu Y, Tanner KD. 2017. Intuitive biological thought: Developmental changes and effects of biology education in late adolescence. Cognitive Psychology 92: 1–21.

Coley JD, Tanner K. 2015. Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. CBE: Life Sciences Education 14: 1–19.

Crouch CH, Mazur E. 2001. Peer instruction: Ten years of experience and results. American Journal of Physics 69: 970–977.

Demastes SS, Settlage Jr J, Good R. 1995. Students' conceptions of natural selection and its role in evolution: Cases of replication and comparison. Journal of Research in Science Teaching 32: 535–550.

Donnelly DF, Namdar B, Vitale JM, Lai K, Linn MC. 2016. Enhancing student explanations of evolution: Comparing elaborating and competing theory prompts. Journal of Research in Science Teaching 53: 1341–1363.

Dunbar K. 1997. How scientists think: On-line creativity and conceptual change in science. Pages 461–493 in Ward TB, Smith SM, eds. Creative Thought: An Investigation of Conceptual Structures and Processes. American Psychological Association.

Duncan RG, Rogat AD, Yarden A. 2009. A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. Journal of Research in Science Teaching 46: 655–674.
Fiedler D, Tröbst S, Harms U. 2017. University students' conceptual knowledge of randomness and probability in the contexts of evolution and mathematics. CBE—Life Sciences Education 16: 38.
Furtak EM. 2012. Linking a learning progression for natural selection to teachers' enactment of formative assessment. Journal of Research in Science Teaching 49: 1181–1210.
Gauvain M, Rogoff B. 1989. Collaborative problem solving and children's planning skills. Developmental Psychology 25: 139–151.
Gelman SA. 2003. The Essential Child. Oxford University Press.
Gould SJ. 1996. Full House: The Spread of Excellence from Plato to Darwin. Harmony Books.
Gregory TR. 2009. Understanding natural selection: Essential concepts and common misconceptions. Evolution: Education and Outreach 2: 156–175.
Gregory TR, Ellis CA. 2009. Conceptions of evolution among science graduate students. BioScience 59: 792–799.
Gunderson EA, Levine SC. 2011. Some types of parent number talk count more than others: Relations between parents' input and children's cardinal-number knowledge. Developmental Science 14: 1021–1032.
Heddy BC, Sinatra GM. 2013. Transforming misconceptions: Using transformative experience to promote positive affect and conceptual change in students learning about biological evolution. Science Education 97: 723–744.
Horn MS, Phillips BC, Evans EM, Block F, Diamond J, Shen C. 2016. Visualizing biological data in museums: Visitor learning with an interactive tree of life exhibit. Journal of Research in Science Teaching 53: 895–918.
Howe C. 2009. Collaborative group work in middle childhood. Human Development 52: 215–239.
Isaacs EA, Clark HH. 1987. References in conversation between experts and novices. Journal of Experimental Psychology 116: 26–37.
Jensen MS, Finley FN. 1996. Changes in students' understanding of evolution resulting from different curricular and instructional strategies. Journal of Research in Science Teaching 33: 879–900.
Jipson JL, Callanan MA. 2003. Mother–child conversation and children's understanding of biological and nonbiological changes in size. Child Development 74: 629–644.
Kampourakis K. 2014. Understanding Evolution. Cambridge University Press.
Kelemen D, Emmons NA, Seston Schillaci R, Ganea PA. 2014. Young children can be taught basic natural selection using a picture-storybook intervention. Psychological Science 25: 893–902.
Kendeou PA, Van Den Broek P. 2005. The effects of readers' misconceptions on comprehension of scientific text. Journal of Educational Psychology 97: 235–245.
Kendeou P, Van Den Broek P. 2007. The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. Memory and Cognition 35: 1567–1577.
Knight JK, Wood WB. 2005. Teaching more by lecturing less. Cell Biology Education 4: 298–310.
Laughlin PR, Vanderstoep SW, Hollingshead AB. 1991. Collective versus individual induction. Journal of Personality and Social Psychology 61: 50–67.
Legare CH, Opfer J, Busch JTA, Shtulman A. 2018. A field guide for teaching evolution in the social sciences. Evolution and Human Behavior 39: 257–268.
Leman PJ, Skipper Y, Watling D, Rutland A. 2016. Conceptual change in science is facilitated through peer collaboration for boys but not for girls. Child Development 87: 176–183.
Lewandowsky S, Ecker UK, Seifert CM, Schwarz N, Cook J. 2012. Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest 13: 106–131.
Loyens SM, Jones SH, Mikkers J, van Gog T. 2015. Problem-based learning as a facilitator of conceptual change. Learning and Instruction 38: 34–42.
Miller JD, Scott E, Okamoto S. 2006. Public acceptance of evolution. Science 313: 765–766.
Miller SA, Brownell CA. 1975. Peers, persuasion, and Piaget: Dyadic interaction between conservers and nonconservers. Child Development 46: 992–997.
Murray F. 1972. The acquisition of conservation through social interaction. Developmental Psychology 6: 1–6.
Nadelson LS. 2009. Preservice teacher understanding and vision of how to teach biological evolution. Evolution: Education and Outreach 2: 490–504.
Nehm RH, Kim SY, Sheppard K. 2009. Academic preparation in biology and advocacy for teaching evolution: Biology versus non-biology teachers. Science Education 93: 1122–1146.
Nehm RH, Reilly L. 2007. Biology majors' knowledge and misconceptions of natural selection. BioScience 57: 263–272.
Nersessian NJ. 1989. Conceptual change in science and in science education. Synthese 80: 163–183.
Nettle D. 2010. Understanding of evolution may be improved by thinking about people. Evolutionary Psychology 8: 205–228.
Neugebauer J, Ray DG, Sassenberg K. 2016. When being worse helps: The influence of upward social comparisons and knowledge awareness on learner engagement and learning in peer-to-peer knowledge exchange. Learning and Instruction 44: 41–52.
Okada T, Simon HA. 1997. Collaborative discovery in a scientific domain. Cognitive Science 21: 109–146.
Paletz SB, Chan J, Schunn CD. 2016. Uncovering uncertainty through disagreement. Applied Cognitive Psychology 30: 387–400.
Pobiner B. 2016. Accepting, understanding, teaching, and learning (human) evolution: Obstacles and opportunities. American Journal of Physical Anthropology 159: 232–274.
Poling DA, Evans EM. 2004a. Are dinosaurs the rule or the exception? Developing concepts of death and extinction. Cognitive Development 19: 363–383.
Poling DA, Evans EM. 2004b. Religious belief, scientific expertise, and folk ecology. Journal of Cognition and Culture 4: 485–524.
Radziszewska B, Rogoff B. 1991. Children's guided participation in planning imaginary errands with skilled adult or peer partners. Developmental Psychology 27: 381–389.
Reiner M, Slotta JD, Chi MT, Resnick LB. 2000. Naive physics reasoning: A commitment to substance-based conceptions. Cognition and Instruction 18: 1–34.
Rice DC, Kaya S. 2012. Exploring relations among preservice elementary teachers' ideas about evolution, understanding of relevant science concepts, and college science coursework. Research in Science Education42: 165–179.
Rutledge ML, Warden MA. 2000. Evolutionary theory, the nature of science and high school biology teachers: Critical relationships. American Biology Teacher 62: 23–31.
Sandoval WA, Reiser BJ. 2004. Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. Science Education 88: 345–372.
Schwarz BB, Neuman Y, Biezuner S. 2000. Two wrongs may make a right… If they argue together! Cognition and Instruction 18: 461–494.
Shtulman A. 2006. Qualitative differences between naive and scientific theories of evolution. Cognitive Psychology 52: 170–194.
Shtulman A. 2017. Scienceblind: Why Our Intuitive Theories about the World Are So Often Wrong. Basic Books.
Shtulman A, Calabi P. 2012. Cognitive constraints on the understanding and acceptance of evolution. Pages 47–65 in Rosengren KS, Brem S, Evans EM, eds. Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution. Oxford University Press.
Shtulman A, Calabi P. 2013. Tuition versus intuition: Effects of instruction on naïve theories of evolution. Merrill-Palmer Quarterly 59: 141–167.

Shtulman A, Neal C, Lindquist G. 2016. Children's ability to learn evolutionary explanations for biological adaptation. Early Education and Development 27: 1222–1236.

Shtulman A, Schulz L. 2008. The relationship between essentialist beliefs and evolutionary reasoning. Cognitive Science 32: 1049–1062.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT. 2009. Why peer discussion improves student performance on in-class concept questions. Science 323: 122–124.

Sota M. 2012. Effect of contrasting analogies on understanding of and reasoning about natural selection. Doctoral dissertation, Florida State University.

Teasley SD. 1995. The role of talk in children's peer collaborations. Developmental Psychology 31: 207–220.

Vosniadou S. 1994. Capturing and modeling the process of conceptual change. Learning and Instruction 4: 45–69.

Vosniadou S, Brewer WF. 1992. Mental models of the earth: A study of conceptual change in childhood. Cognitive Psychology 24: 535–585.

Weisberg DS, Landrum AR, Metz SE, Weisberg M. 2018. No missing link: Knowledge predicts acceptance of evolution in the United States. BioScience 68: 212–222.

Wiser M, Amin T. 2001. "Is heat hot?" Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. Learning and Instruction 11: 331–355.

Wyner Y, Doherty JH. 2017. Developing a learning progression for three-dimensional learning of the patterns of evolution. Science Education 101: 787–817.

Yates TB, Marek EA. 2014. Teachers teaching misconceptions: A study of factors contributing to high school biology students' acquisition of biological evolution-related misconceptions. Evolution: Education and Outreach 7: 7.

Young AG, Alibali MW, Kalish CW. 2012. Disagreement and causal learning: Others' hypotheses affect children's evaluations of evidence. Developmental Psychology 48: 1242–1253.

*Andrew Shtulman (shtulman@oxy.edu) is affiliated with the Department of Psychology at Occidental College, in Los Angeles, California, in the United States. Andrew G. Young (ayoung20@neiu.edu) is affiliated with the Department of Psychology at Northeastern Illinois University, in Chicago, Illinois, in the United States.*