Refutation texts and argumentation for conceptual change: A winning or a redundant combination?

Christa S. C. Asterhan & Maya Resnick

The Seymour (Shlomo) Fox School of Education
The Hebrew University of Jerusalem
asterhan@huji.ac.il

In press for *Learning & Instruction*
September, 2019

Cite as: Asterhan, C. S. C. & Resnick, M. (in press). Refutation texts and argumentation for conceptual change:A winning or a redundant combination? *Learning & Instruction*

Abstract

Effective instruction for conceptual change should aim to reduce the interference of irrelevant knowledge structures, as well as to improve sense-making of counterintuitive scientific notions. Refutation texts are designed to support such processes, yet evidence for its effect on individual conceptual change of robust, complex misconceptions has not been equivocal. In the present work, we examine whether effects of refutation text reading on conceptual change in biological evolution can be augmented with subsequent peer argumentation activities. Hundred undergraduates read a refutation text followed by either peer argumentation on erroneous worked-out solutions or by standard, individual problem solving. Control group subjects read an expository text followed by individual problem solving. Results showed strong effects for the refutation text. Surprisingly, subsequent peer argumentation did not further improve learning gains after refutation text reading. Dialogue protocols analyses showed that gaining dyads were more likely to be symmetrical and to discuss core conceptual principles.

1. Introduction

For more than four decades, scholars have documented how student explanations about natural phenomena may not align, and are at times even incommensurate, with the scientific yet often counterintuitive concepts that they are taught in science classes. Coming to understand and being able to correctly use these scientific explanations is not a matter of "gap-filling", in that learners merely lack the necessary knowledge, but rather involves a substantive reorganization of existing knowledge structures, an outcome that is usually referred to as "conceptual change" (e.g., Chi, 2008; Vosniadou & Brewer, 1994).

Traditionally, conceptual change has often been described as a correction or replacement of misconceived conceptions that reside in the mind (e.g., reviews in Özdemir & Clark, 2007; Vosniadou, 2009). Yet, current accounts describe it in terms of a response competition between commonly used and rapidly activated knowledge components that lay at the basis of erroneous scientific explanations, on the one hand, and rarely used and/or weakly connected knowledge components that are needed to construct a scientifically accurate and full explanation, on the other (e.g., Kendeou & O'Brien, 2014; Potvin, Sauriol, & Riopel, 2015; Ramsburg & Ohlsson, 2016; Schnotz & Preuß, 1999). Improved inhibition of these irrelevant, yet easily activated knowledge components reduces their interfering influence on the construction of an accurate representation in working memory. This response competition account is further supported by recent empirical evidence showing that conceptual change involves both an improved capability to construct and identify the correct scientific explanation, as well as increased inhibitory activities, even among experts (e.g., Babai, Sekal & Stavy, 2010; Dunbar, Fugelsang & Stein, 2007; Masson, Potvin, Riopel, & Foisy, 2014; Potvin, Masson, Lafortune & Cyr, 2015; Shtulman & Valcarcel, 2012).

Extrapolating from this response competition account, effective instructional activities that aim for conceptual change should support both these cognitive processes. That is, they should provide learners with opportunities to comprehend the scientifically accepted, yet often times counterintuitive explanations, but also provide them with opportunities to become aware of and understand the errors in (their) lay theories (Chan, Burtis & Bereiter, 1997). Not surprisingly, traditional tell-and-practice teaching approaches in which students are only exposed to the full, correct

explanations have not been found to be very effective for learning that requires knowledge revision, especially in the case of robust misconceptions (Chi, 2008; Vosniadou & Mason, 2012). Researchers of instructional approaches for conceptual change have then studied the effectiveness of alternative instructional techniques, materials and activities. One of these has been to replace expository with refutation texts.

1.1 Refutation texts and conceptual change

Science textbooks traditionally contain expository texts in which scientific concepts are explained in detail, without directly referring to common misconceptions (Osborne, 2010; Tippet, 2010). In refutation texts, on the other hand, the commonly held misconception is explicitly stated upfront and then refuted, after which the reader is introduced to the established, correct scientific explanation (Sinatra & Broughton, 2011; Tippett, 2010; Vosniadou & Mason, 2012). The rationale behind the advantage of refutation texts for knowledge revision processes is rooted in both conceptual change as well as reading comprehension theories, and has been summarized by Kendeou and O'Brien (2014) as follows: For knowledge revision to occur, the correct and incorrect knowledge components have to be co-activated in working memory. This supports their comparison and contrast. Readers are more likely to notice the discrepancy between their own intuitive understanding (as presented in the common misconception) and the scientific one, and to encode the newly presented information correctly. In contrast, readers with misconceptions make more invalid inferences while reading expository texts, as the newly presented information is assimilated into the incorrect mental representations constructed in a person's working memory based on his/her pre-existing knowledge (Kendeou & van den Broek, 2007; Van den Broek & Kendeou, 2008). Explicit references and statements about the incorrectness of misconceptions (refutation cues) play an important role in refutation texts. Self-directed comparisons between presentations of the correct and the erroneous conceptions without these explicit refutation cues have not been as successful, particularly among learners with misconceptions (e.g., Asterhan & Dotan, 2018; Braasch, Wiley & Goldman, 2013; Weingartner & Masnick, 2019). Refutation texts are also more effective when the text includes and interconnects evidence to support the scientific concept (Kendeou & O'Brien, 2014).

The use of refutation texts has been investigated rather intensively over the years (see Guzzetti et al, 1993; Tippett, 2010; Vosniadou & Mason, 2012; for

reviews). Particularly, considerable progress has been made in identifying the online cognitive and affective processes during reading (e.g., Diakidoy et al., 2016; Kendeou & van den Broek, 2007; Muis, Sinatra, Pekrun, Winne, Trevors, Lossenno & Munzar, 2018; Trevors & Muis, 2015), the effectiveness of different refutation text design features (e.g., Braasch et al. 2013; Danielson, Sinatra & Kendeou, 2016; Franco, Muis, Kendeou, Ranellucci, Sampasivam, & Wang, 2012; Mason, Baldi, Di Ronco, Scrimin, Danielson, & Sinatra, 2017), the effects on reading comprehension and text recall measures (e.g., e.g., Diakidoy, Mouskounti & Ionnides, 2011; Diakidoy et al., 2016; Kendeou, Walsh, Smith, & O'Brien, 2014) and the interaction with individual characteristics (e.g., Cordova, Sinatra, Broughton, Taasoobshirazi, & Lombardi, 2014; Mason, Gava, & Boldrin, 2008; Trevor & Muis, 2015).

In the present work, we focus on the effects of refutation texts on conceptual change outcomes, that is: changes from students' pretest understanding of a complex scientific topic toward a more advanced, scientifically correct understanding at posttest. In spite of the strong rationale favoring refutation texts for conceptual change learning and the substantive amount of research in this field, results from research *directly* comparing the effects of refutation and expository texts on students' conceptual learning outcomes have not been unequivocal. Alongside studies reporting strong positive effects (e.g., Ariasi & Mason, 2011; Muis et al., 2018; Van Loon et al., 2015), reports on no, partial or small effects on conceptual learning outcomes are not uncommon (e.g., Alverman & Hague, 1989; Braasch et al., 2013; Broughton & Sinatra, 2010; Diakidoy et al., 2011, 2016; Lombardi, Danielson & Young, 2016; Mason, Zaccoletti, Carretti, Scrimin & Diakidoy, in press; Hart & Nisbet, 2012; Palmer, 2003). The reasons behind these mixed findings are not known at this point and, at the time of writing this report, meta analyses are not available.

The present study aims to contribute to this body of work in two ways: (1) We focus on a particular type of robust misconceptions that has been less frequently studied in the literature on refutation text effects; and (2) we explore whether refutation text effects may be augmented with subsequent peer argumentation activities.

1.2 Levels of knowledge revision

Some of the work reporting (strong) positive effects of refutation texts focused on knowledge revision at the level of discrete beliefs, such as whether ostriches bury their heads in the sand or whether people only use 10% of their brains (e.g., Beker,

Kim, van Boekel, van den Broek, & Kendeou, 2019; Donnovan et al., in press; Kendeou et al., 2014; Tippet, 2010; Van Loon et al., 2015). These types of erroneous ideas can be refuted by correcting a single mistaken belief, one of the lower levels of knowledge revision (Chi, 2013). In contrast, the robust misconceptions that have traditionally been the focus of conceptual change research in science education typically require a substantive restructuring of complex knowledge systems. This involves the revision of multiple misconceived knowledge components, as well as the way in which these are interconnected, (e.g., Chi, Roscoe, Slotta, Roy, & Chase, 2012).

For example, biological evolution is a complex, multi-faceted concept whose understanding requires the integration of several, often counterintuitive notions, such as intra-population variance, proportional change, randomness and basic genetics (Ferrari & Chi, 1998; Shtulman, 2006). Moreover, intuitive theories about biological evolution are usually based on explanatory schemata that are incommensurate with the scientifically accepted account (e.g., Shtulman, 2006). Chi and colleagues identified that these misconceptions often times show features of direct and sequential, instead of emergent process concepts (Chi et al., 2012). A sequential process is, among others, characterized by having a clear beginning and end, a sequence of distinct actions that are contingent and causal, and an identifiable, explicit goal. Emergent processes, such as for example diffusion and biological evolution, on the other hand, are uniform, simultaneous and ongoing, and emerge from random actions and interactions between actors on a micro level. Previous research has shown that standard tell-and-practice forms of instruction are insufficient to induce a substantive and lasting change in students' understanding of biological evolution (e.g., Astrehan & Schwarz, 2007, 2018; Jensen & Finley, 1996; Jimenez-Aleixandre, 1992).

When refutation text research focuses on robust misconceptions of complex scientific ideas, such as for example natural selection, energy and photosynthesis, empirical evidence of positive effects of refutation texts for conceptual change outcomes is considerably less frequent (but see Ariasi & Mason, 2011; Diakidoy et al., 2003; Mason et al., 2008; Mikkilä-Erdmann, 2001; Muis et al., 2018). More research that compares the effects of refutation *vs*. expository texts on such conceptual change outcomes would then be welcome. In the present work, we contribute to this line of research by comparing the effects of reading a refutation *vs*.

an expository text on conceptual understanding of a complex science topic for which students are known to have particularly robust misconceptions (biological evolution).

1.3 Augmenting refutation text effects with subsequent peer argumentation activities

It has been suggested that, in order to foster and sustain long-lasting conceptual change on complex science topics, students should engage in subsequent learning activities to make sense of, explain and practice their newly acquired knowledge from refutation texts (Hynd & Guzetti, 1997; Guzzetti, Williams, Skeels, & Wu, 1997; Vosniadou & Mason, 2012). Previous research has shown that the effect of refutation texts can be ameliorated by integrating visualizations or analogies into the materials (e.g., Danielson et al., 2016; Mason et al., 2017).

Yet, few have explored the added benefits of social sense-making activities following refutation text reading. Guzzetti and colleagues (1997) conducted in-depth case study analyses of three secondary school Physics classes. They concluded that whereas refutation text reading was effective in drawing students' attention to the incongruence between intuitive and scientific explanations, it was overall more successful when supplemented with discussion. In the present study, we test these insights from classroom observations in a controlled experiment. We explore whether refutation text effects are augmented when reading is followed by a particular social sense-making activity, namely deliberative peer argumentation.

In its ideal form, peer argumentation involves two (or more) discussants who compare, critique and weigh different explanations through reasoned argument in a constructive and collaborative atmosphere (Asterhan, 2013 Felton et al., 2009). Previous research has shown that this type of peer dialogue, also termed deliberative argumentation, can support conceptual change (see review by Asterhan & Schwarz, 2016; Asterhan & Schwarz, 2007, 2009; Asterhan & Babichenko, 2015; Berland & Hammer, 2012; deVries, Lund & Baker, 2002; Nussbaum & Sinatra, 2003; Osborne, 2010). However, in spite of its promise and its proven effect in tightly controlled and scripted settings, empirical evidence on conceptual change through peer argumentation is not plentiful (Asterhan & Schwarz, 2016).

One reason may be that in standard argumentation tasks, key differences between misconceptions and scientifically accepted notions are not made salient or accessible enough for students (Asterhan & Dotan, 2018; Asterhan & Schwarz, 2016). Typically, they are given an expository knowledge source explaining the correct scientific account to prepare them for the peer discussion phase. The underlying

assumption is that students will notice the difference between the scientific and their own and/or their partner's intuitive notions by themselves, and then try to resolve these differences together. However, they often times do not detect when explanations are conceptually different and therefore fail to discuss differences (Sfard, 2009), foregoing the particular affordances of peer argumentation through the exploration and comparison of different perspectives. In contrast, one-sided argumentation, in which students only elaborate and develop one explanation type and fail to explore and juxtapose different explanations, has not been found to support conceptual change (e.g., Asterhan & Schwarz, 2007, 2009).

In a refutation text, common misconceptions and correct scientific explanations are juxtaposed explicitly. Yet, when used in isolation, without further instructional activities to support and consolidate the cognitive processes set in motion by text reading, the effects of refutation texts may be short-lived, especially in the case of robust misconceptions for complex scientific notions (Guzzetti et al., 1997). Refutation text reading and peer argumentation activities are then expected to complement each other by supporting the two aforementioned processes underlying conceptual change, that is: promoting awareness to and reducing interference of irrelevant knowledge structures and sense-making of the counterintuitive, scientifically accepted notions.

In the present work, we test the expectation that student learning outcomes are augmented when refutation text reading is followed by a structured social sense-making activity that requires learners to collaboratively correct and discuss differences between erroneous and correct explanations through peer argumentation.

1.5 The present study and hypotheses

The aforementioned expectations were tested in a randomized experiment in which undergraduate students participated in one of three different learning activity sequences on the topic of biological evolution. In the two experimental conditions, undergraduate students read a refutation text which was followed by either an argumentive peer discussion on erroneous solutions or by a standard individual problem solving activity. Students in the control condition read an expository text, followed by a standard individual problem solving activity.

Based on the aforementioned rationale the following hypotheses were formulated:

H1: Reading a refutation text on biological evolution results in larger learning gains than reading an expository text on the same topic.

H2: Reading a refutation text followed by dyadic peer argumentation results in larger conceptual gains than reading refutation texts followed by a standard, individual problem solving activity, which in turn leads to larger gains compared to reading an expository text followed by individual problem solving.

2. Method

2.1 Participants

One hundred undergraduate students (73 females, M_{age} = 24.5) from a large university in central Israel participated in the study. We targeted adult university students as they were assumed to have the argumentive skills and norms necessary to engage in deliberative peer argumentation on a scientific topic without having to undergo intensive training in argumentation prior to the experiment (Kuhn, 1991). We targeted only students without a higher education background in the exact or life sciences, as previous studies have shown that misconceptions about natural selection are abundant in this group (Asterhan & Dotan, 2018; Asterhan & Schwarz, 2007). Furthermore, as the texts were in Hebrew, only students with a Hebrew proficiency at the mother tongue level were eligible for participation. Recruitment was achieved through the locally designated online system for participation in experiments. Participants were offered a choice of course credit (31% of the sample) or financial reimbursement for participation ($15). Four participants failed to appear for the delayed posttest (2 from the Ref+Arg and 2 from the RefOnly condition). Their background information and pretest scores were similar to the sample mean, as well as within the relevant condition. Their data was omitted only from analyses that include the delayed posttest scores.

2.2 Design

A between-subject experimental design was employed. Participants were randomly assigned to one of three conditions (see Figure 1): (1) Refutation text + dyadic argumentation on erroneous worked-out examples (Ref+Arg; N = 50); (2) Refutation text + individually solving open questions (RefOnly; N = 26); (3) Expository text + individually solving open questions (Control; N = 24). Individual conceptual understanding of biological evolution was assessed on pre-test, immediate posttest (following text reading, prior to the second activity), and delayed posttest (a

week later). We used G*Power 3.1 software to conduct an *a priori* power analysis to determine the sample size required to detect a large effect size (η^2= .15) with a power of .80 and α error probability of .05 for the main analysis (one way ANOVA with three conditions). The recommended sample size was n = 22 per condition. As the Ref+ARg condition included a dyadic interaction, the sample size was doubled for that condition.

Insert Figure 1 Here

2.3 Instruments

2.3.1 Demographics and background. Gender, age, degree, major program, religious affiliation, degree of religiosity, and background in high school biology education and evolution were all collected. In addition, questions regarding students' attitudes and beliefs concerning the theory of biological evolution were translated to Hebrew from Shtulman (2006) (α = .84).

2.3.2 Conceptual understanding of biological evolution. Individual conceptual understanding of biological evolution at pre-test, immediate and delayed posttest was assessed with open and forced choice items that targeted the evolution of selected animal traits (adapted from As, 2015). Eight different animal species and traits were distributed over the three tests (see Table1). The pretest and the delayed posttest each included questions about 3 different animals: Six different false/correct statements about the first, five about a second and one open-ended construction item about the third. To avoid test fatigue, the immediate posttest was slightly shorter and included questions about 2 animals only (six true/false and one open-ended).

For each animal species, students were presented with a short text introducing the species, a physiological change in a specific trait over time and a short description about its importance for survival. Each of the false/true items addressed a different principle of biological evolution (see Table 2). Students were required to indicate whether the statement was false or correct and then explain their choice. For a given animal species, approximately half of the statements were incorrect, targeting common misconceptions (counterbalanced per principle). In the open items, students were required to give a full explanation of how the given trait had evolved, according to the theory of natural selection. Examples of the different test item formats are presented in Appendix A.

The reason for including forced choice test items was based on previous work showing that participants often do not refer to each and every principle of biological

evolution in open construction test items. Whereas the overall schema of change alluded to can be deduced from these responses, it is difficult and even impossible at times to assess whether they understood a particular principle or not. The combination of forced choice with open construction items then allowed for a balanced and comprehensive assessment of student conceptual understanding. Internal reliability for the pretest (Cronbach $\alpha = .78$), the immediate posttest (Cronbach $\alpha = .78$) and the delayed posttest (Cronbach $\alpha = .86$) was good.

Insert Tables 1 and 2 Here

2.3.3 Instructional texts. An expository text (453 words in Hebrew) and a refutation text (525 words in Hebrew) on natural selection were created, based on common middle school biology textbooks. They were verbatim identical with regard to the background information, the presentation of the scientifically accepted theory, the explanations and a well-known example of change in a specific animal species (i.e., changes in the giraffe's neck). In addition, the refutation text included (1) explicit references to common misconceptions, and (2) refutation cues stating that those are incorrect. Refutations specifically referred to misconceived notions about intra-species diversity, the source of diversity and the intentionality of change (see Table 2). An excerpt of the text is presented in Appendix B.

2.3.4 Problem solving tasks. The problem-solving tasks for stage 4 (see Figure 1) consisted of three open-ended questions (each on a separate page) about two novel evolutionary phenomena, namely the webbed feet of ducks and the wing coloring of the peppered moth (adapted from Asterhan & Schwarz, 2007). The first two questions were textual and similar in format to the conceptual knowledge test open items, which required students to explain the described change in terms of biological evolution. In the third task (adapted from Shtulmann, 2006), students were asked to depict graphically the gradual change in wing coloring over time. This item is used to distinguish between typological and selection-based representation of change (see Figure 2).

Students in the Ref+Arg condition received the same booklet, with two changes: First, the space for writing the solutions was already filled with handwritten solutions provided by three (fictitious) peer students. They were erroneous, targeting a particular common misconception in each of the three solutions and adapted from common student answers from previous data bases. The errors in the two textual solutions were highlighted with a yellow marker. For example, the solution to the

webbed duck feet question was designed to refer to the common misconception that individual animals intentionally change a trait during their lifetime (highlighted error in italics here):

> "The ducks needed webs to swim. They had to know how to swim in order to survive. *Some of the ducks managed to develop webbed feet for themselves*. They survived and managed to reproduce. Those ducks that *did not manage to develop* webbed feet did not survive"

The erroneous textual solution to the moths question alluded to misconceptions about existing intra-species variability ("before, *all the moths were white*") and a typological change ("*in each generation every moth became a bit darker*"). Finally, the graphic depiction was already colored by a (fictitious) peer, to depict a classic typological model of change (see Figure 2). A separate space was reserved below each erroneous answer for the participants to fill in their corrected solution. The worksheets in both scenarios (peer argumentation on common errors and individual problem solving) were used for the purpose of the experimental intervention and were not used for data analysis.

Insert Figure 2 Here

2.4 Coding

2.4.1 Coding understanding of evolution. Coding of written solutions was based on existing coding procedures developed in previous studies (Asterhan & Dotan, 2018). Each written solution was graded according to accuracy and compliance with the main principles of biological evolution: (0) for omissions, misconceptions or other crucial errors, (.5) for partially correct solutions, or full credit (1.0) for solutions that contained no misconceptions and addressed the main tenets of biological evolution correctly.

Each false/correct item targeted one of the six predefined principles of biological evolution (see Table 2). When coding for the true/false items, the indicated choice of right or wrong and the accompanying textual explanation were considered together. A correct choice together with a correct and sufficient explanation resulted in full credit. An incorrect true/false choice with an incorrect explanation, resulted in zero points. When these two components were not aligned, points were assigned based on the verbal explanation. Most of these cases indeed showed partial understanding (0.5 points), but there were several cases that showed clear

misconceptions in the written explanations with the correct choice of true/false (0 points).

Written solutions to the open items were coded in a similar manner, but regarded the overall model of evolutionary change represented in the student answers (see Asterhan & Dotan, 2018), instead of only the particular principle targeted. Solutions that contained no misconceptions and correctly explained change in terms of existing variability, selection and proportional change received full credit (1). Answers that were partially correct or presented both correct as well as incorrect aspects received .5 points. This is also the case for well-documented hybrid models to explain biological evolution (Asterhan & Schwarz, 2007).

Following a training period, three human coders scored 248 randomly chosen item responses (about 9% of the total data set). Interrater reliability was satisfactory, .72 < Cohen's κ <.79. Differences were resolved through discussion, after which the entire data set was coded. A total conceptual understanding score was compiled by adding the different scores for each test item on each test, while assigning the open test item score a weight of 5 points (instead of 1). They were then transformed into percentage scores that ranged from 0-100.

2.4.2 Coding of dialogues. Twenty-three audio-recorded discussions were transcribed (1 was incomprehensible and 1 was lost due to technical failure). The mean length of these audio-recorded discussions was 8:16 *min* (ranging from 4:15 *min* and 19:01 *min*). Transcriptions included all verbal content, but not intonation and other auditory features. Following previous work (e.g., Asterhan & Babichenko, 2015; Asterhan & Schwarz, 2009), initial coding efforts focused on three discussion characteristics: Whether the discussion could be characterized as critical-dialectical overall, interaction symmetry and rhetoric style (i.e., disputative or deliberative argumentation). As the argumentation instructions were explicitly modelled on and directed toward deliberation, clear cases of disputative argumentation were near non-existent and rhetorical style was therefore dropped from the analyses. Following these top-down efforts with existing schemes, we searched the data set for additional salient features (Chi, 2007). This procedure yielded an additional coding category, namely the extent to which student dyads discussed the core conceptual principles of biological evolution or not.

In sum, three dialogue characteristics were coded:

(1) Critical discourse (0, 1): When the students overtly disagreed about the different solutions and related to the differences by providing justifications, explanations and counterargument, it was considered a critical discussion.

(2) Symmetry (0,1): When the word count from all the conversational turns from one discussant partner was less than 35% of the total word count (excl. repetitions) the discussion was deemed asymmetrical.

(3) Discussion of core principles (0, 1): When at least 5 of the 6 principles of biological evolution (see Table 2) came up during the discussion, a score of 1 was assigned. When 4 or less core principles were mentioned it received score of 0. It should be noted that the erroneous worked-out examples referred to three different core principles altogether, but a fully detailed, correct explanation would require all 6 principles.

Two raters scored the 23 dialogue protocols independently. Interrater agreement was satisfactory, and ranged from Cohen's κ = .82 (core principles) to κ = .68 (critical dialogue).

2.5 Procedure

Except for stage 4 (see Figure 1), all experimental stages were conducted in individual, separate rooms. In stage 1, students completed the background information survey and the pretest. They then received either the refutation text (Ref+Arg and RefOnly conditions) or the expository text (control condition) (10 *min*). Following the immediate posttest (stage 3), individuals were moved to a different room shared with another participant (stage 4). Assignment to peer participant was random within condition and participants were allotted a maximum of 20 *min* to complete this stage. Students in the RefOnly and the Control conditions were seated with their backs to one another and received the standard worksheet with open questions, which they were instructed to solve individually and without talking to one another. Participants in the Ref+Arg condition, on the other hand, were instructed to work in pairs to critically, yet constructively, discuss the erroneous student solutions in the filled-in worksheets (Asterhan & Schwarz, 2016). Following are the exact instructions:

> "In this section, you will conduct a critical discussion with a fellow student. The discussion will revolve around erroneous and correct solutions to questions about natural selection. What do we mean by a "critical discussion"? The purpose of a discussion of this kind is to reach a better and deeper

understanding of the topic, in this case: biological evolution of animals. A productive discussion is one in which the discussants examine in depth every idea and explanation that comes up in the conversation, while giving arguments and reasons. It is important to emphasize that a critical discussion is supposed to help both of you reach a better understanding. In every step of the discussion, try to consider the weaknesses and strengths of each explanation offered, whether it was raised by you or by your partner. Try to think about the reasons why a certain idea or solution may or may not make sense. To what extent do the reasons, the evidence and the explanations support the presented explanation? Are there alternative explanations you have not considered? "

The argumentation sessions were audio-recorded. Each dyad member then received a clean work sheet copy, with a designated space to correct the erroneous explanations individually, to avoid that one student would dominate the writing and decision making in spite of the instructions. A week later, participants returned for stage 5 to complete the delayed posttest, to receive a debriefing and to be reimbursed (15 *min*).

3. Results

Table 3 presents the mean conceptual understanding scores per condition, for the pretest, immediate posttest and delayed posttest. Normalized gain scores were computed for the overall conceptual understanding score, as well as for the false statement items score and for the correct statement items score separately[1]. No differences were found on pretest scores between conditions, F (2, 97) = .60, p = .560.

None of the three control variables (i.e., attitudes, religiosity, and perceived understanding) yielded differences between conditions, nor did they correlate with normalized pre-to delayed posttest gain scores (r = .13, r = -.17 and r = -.09, *ns*, respectively).

3.1 Effects of condition on conceptual gains

The mean conceptual gain scores per condition are presented in Table 4. Distributions of the gain score were checked for outliers. Gain score residues were

[1] Normalized gain scores express a relative progress as a function of the total amount of potential progress, according to the following formula: gain = ((t_2) – (t_1) / (100- t_1)) * 100. Thus, when t_1 =34.38 and t_2 = 78.13 then the normalized gain score is 65.00. When the two scores were identical, a gain score of 0 was assigned. When t_2 < t_1, the formula was adjusted to ((t_2) – (t_1) / (t_1)) * 100.

checked for normality assumptions by inspecting skewness and kurtosis (<1), Q–Q plots, and Kolomogorov–Smirnov tests of normality. Comparisons were conducted with one-way ANOVA tests. To test the stability of the results and the reliability of the chosen method for analysis, all comparisons were checked with two additional statistical models, namely ANOVA on the difference between the raw pre- and posttest scores (without the normalized gain score transformation) and ANCOVA on the delayed post test scores with pretest or immediate posttest as covariates. As relative gain scores may favor high pretest scorers who make small raw gains, we also re-ran analyses while excluding (eight) participants with relatively high pretest scores (> .85). All models produced identical results, showing the robustness of the findings and proving that the findings cannot be attributed to the particular statistical method and method of gain score calculation chosen here. Levene's tests for equality of error variance across compared conditions showed that this assumption was not violated (p > .80).

Insert Tables 3 and 4 Here

3.1.2 The effect of refutation text on conceptual gains. In order to examine the effect of text type (refutation vs. expository) on conceptual knowledge (H1), mean normalized gain scores from pretest to immediate posttest were compared between students who had read a refutation vs. an expository text. Students in the refutation text condition showed larger conceptual gains on the immediate posttest (M = 43.21, SD = 44.86, N = 74) than students in the expository text condition (M = -1.94, SD = 52.77, N = 26), t (98) = 4.21, p < .001, with a large effect size of d = .92.

Further analyses showed that this advantage was also evident on the delayed posttest: Students who had read a refutation text showed larger conceptual gains (M = 46.10, SD = 36.53, N = 70) than students who had read an expository text (M = 9.32, SD = 40.57, N = 26), t (94) = 4.25, p < .001, with a large effect size of d = .95. These findings corroborate the first hypothesis (H1), according to which refutation texts are more effective than expository texts for both short-term, as well as long-term conceptual gains.

3.1.2 The additional effect of argumentation on conceptual understanding. A one-way ANOVA compared the mean normalized gain scores from pretest to delayed posttest, across the three conditions. A main effect of condition on normalized gains was found, F (2, 93) = 9.02, p < .001, with a large effect size of η^2_p = .16. Planned comparisons showed that the mean gain in the control condition (M = 9.32, SD =

40.60) was significantly lower than the mean gains in the RefOnly condition (M = 48.37, SD = 35.55, p < .001, d = 1.02) and the Ref+Arg condition (M = 45.06, SD = 37.29, p < .001, d = .92). However, no differences on mean gain scores were found between the latter two, p = .727.

A one-way ANOVA compared the mean normalized gain scores from immediate to delayed posttest, across the three conditions. Overall, gains from the additional activity were low with a large variance (M = 10.26, SD = 34.71) and no significant differences were found between the normalized gain scores of the Ref+Arg (M = 14.40, SD = 35.16) the RefOnly (M = 8.71, SD = 38.60), and the control condition (M = 3.94, SD = 30.41), F < 1. 01.

Taken together, these findings do not support the second hypothesis (H2), according to which peer argumentation on erroneous solutions would further increase learning gains compared to individual problem solving activities. Even though the immediate posttest scores were overall fairly high (M = 65.63 and M = 73.14 in RefOnly and the Ref+Arg, respectively), there was definitely room for more improvement and a ceiling effect is therefore unlikely.

3.1.3 Effects on learning outcomes per test item types. One could argue that the obtained pattern of results could be attributed to the particular test item format that was used predominantly in the assessments, namely forced choice true/false items. Having to make a decision about whether a given statement is true or false may come easier to students who had read the refutation text, not necessarily because they have a more advanced understanding, but because they were better prepared for this type of test item format. To control for this possibility, we compared student performance on the open test item that required them to autonomously construct a full explanation to a new phenomenon. Evidence of substantive conceptual gains (conceptual change) in student explanations to the open-ended item was defined as an improvement of .5 on the nominal, unweighted open item test score from one test occasion to the other (Asterhan & Dotan, 2018). Only students with pretest scores of .5 or less were included in the analyses, since a maximum score of 1 indicates that the student had already provided a natural selection-based explanation on the pretest.

A Chi square test showed that, compared to the control condition (55%, N= 10), students who had studied refutation texts showed substantive conceptual gains more often (75%, N = 36), χ^2 (1, 69) = 5.84, p = .016. A comparison between the Ref+Arg and RefOnly conditions showed no significant differences in improvement

from pretest to delayed posttest, χ^2 (1, 48) = 1.21, p = .271, nor from immediate to delayed posttest, χ^2 (1, 30) < 1.

Insert Table 5 Here

Finally, separate analyses on the gain scores for false statements and for the correct statements, respectively, mirrored the main findings on the total score (see Table 5): Main effects of condition were found on the normalized false statement gain score, F (2, 93) = 8.36, p < .001, η = .15, and the true statement gain score, F (2, 93) = 7.27, p = .001, η = .14. Learners who had read a refutation text improved significantly more on true as well as false statement test items than learners in the control condition (.002 < p < .004). No differences were found between the Ref+Arg and the RefOnly condition.

3.2 Dialogue protocol analyses

The discussion characteristics of dyads in which none of the dyad partners showed a substantive gain from the immediate to delayed posttest was compared to dyadic discussions in which at least one dyad partners showed such gains. In three dyads, both partners had near perfect scores (> 91) at the immediate posttest and were therefore not included in the discussion feature analyses. Substantive gains were defined as further increase of 30% from the immediate to the delayed posttest (normalized gain score > 30). This definition resulted in 10 “gaining” and 10 “non-gaining” dyads. Table 6 presents the cross-tabulation of dialogue features and dyad gains.

Insert Table 6 Here

Table 6 shows that the dialogues of gaining dyads were more likely to include references to core conceptual principles of biological evolution, χ (1, 20) = 7.20, p = .007. Their interactions also tended to be symmetrical more often, even though this difference was only marginally significant, χ (1, 20) = 3.33, p = .068. A certain trend could be observed by which the discussion of gaining dyads seemed to be characterized as critical more often, but this was not statistically significant, χ (1, 20) = 1.98, p = .160. Even though 8 out of 10 gaining dyads conducted a critical discussion, 5 out of 13 critical discussions were not followed by substantive gains by at least one of the pair.

4. Discussion

4.1 Refutation texts for conceptual change of robust science misconceptions

The findings presented here show, first and foremost, that replacing a standard expository text with a refutation text that explicitly mentions common misconceptions, refutes and contrasts them with the correct, scientific explanations significantly improves student conceptual understanding of biological evolution. This advantage was evidenced immediately following the text reading, but also on delayed posttests administered a week later, thus suggesting a stable knowledge revision. The present findings add to existing research showing strong and stable refutation text effects for, not only for incorrect beliefs, but also for complex scientific topics for which learners have robust misconceptions, such as energy and biological evolution (Ariasi & Mason, 2011; Diakidoy et al., 2003; Mason et al., 2008; Mikkilä-Erdmann, 2001; Muis et al., 2018).

Second, we explored whether dyadic argumentation on common misconceptions would further improve conceptual understanding over and above the effect of refutation text. Argumentation and refutation text are two well-known instructional interventions for conceptual change that have each been studied in isolation. Yet, the additive effect of their combination has hitherto not been studied. Surprisingly, our expectations were not confirmed, as conceptual understanding did not further improve over and above the immediate posttest, neither by subsequent peer argumentation, nor by a standard individual problem solving activity. We offer the following potential explanations:

One could argue that peer argumentation and reading refutation texts are both argumentive activities in essence and are therefore too similar for the second activity to have an additional effect. They both involve critique, contrasting perspectives, justifications and comparisons, albeit in different formats (oral *vs*. textual) and with distinct degrees of explicit, expert authority. However, students typically receive less support in peer argumentation activities. In our study, they had access to materials showing common misconceptions about biological evolution (i.e., the erroneous worked-out examples), but not to the correct explanations. Moreover, whereas in refutation texts the two are presented side-by-side and explicitly compared and contrasted, in peer argumentation activities students are expected to do this by themselves. Previous research has shown that students and even undergraduates find this difficult (e.g., Asterhan & Dotan, 2018; Durkin & Rittle-Johnsson, 2012). Perhaps the task that students were given in the argumentation phase directed their attention too much toward structured correction of erroneous examples, instead of

collaborative construction activities. Future research could explore whether a more open-ended collaborative argumentation task would be more productive.

A second explanation could be sought in the particular population we sampled from, namely undergraduate university students. The benefits of subsequent collaborative sense-making activities might be expected to be more prominent among younger learners or among learners from a more heterogeneous population with regard to prior academic achievement and competencies. Prior research has shown the advantage of refutation over expository text reading among school-aged children (e.g., Diakidoy et al., 2003; Mason et al., 2008; Mikkilä-Erdmann, 2001). Yet, these did not explore whether learning gains could be further enhanced with subsequent sense-making activities, such as collaborative argumentation. Future research should explore whether refutation text reading effects could be ameliorated with subsequent collaborative peer argumentation in a younger and more heterogeneous population of school-aged children and adolescents.

Finally, post hoc dialogue protocol analyses suggested that gains from the argumentive activity were contingent on the quality of the dyadic interaction: A comparison between gaining and non-gaining dyads showed that the dialogues of gaining dyads tended to be more symmetrical (i.e., more egalitarian distribution of verbal contributions) and included more references to the core principles of biological evolution. Thus, even though a main effect of task design was not found, associations between differences in dyadic argumentation quality and further learning gains were observed.

4.2 Limitations and future research directions

In addition to the aforementioned suggestions for future research, we highlight several additional limitations and directions for further investigations. First, the findings presented here were obtained on a specific topic (the biological evolution of animals), in a specific country (Israel) and with participants from a specific age group (undergraduates). We have already alluded to the need for more research on younger and more cognitively heterogeneous populations, especially regarding the potential of subsequent collaborative argumentation for learning. The combined effects of argumentation and refutation text should also be tested in natural classroom settings. Peer argumentation in a more familiar and social setting may prove to be more productive than in a controlled laboratory setting with an unknown peer. More research is also needed to explore whether the current findings about combining

refutation and peer argumentation can be generalized to different conceptual change topics. Specifically, even though biological evolution of animals is not considered a controversial topic in Israel, in many other countries it is. In those cases, conceptual change is not only a matter of purely cognitive complexity, but also of affect, values and motivation (e.g., Sinatra, Brem, & Evans, 2008).

Second, it could be argued that the strong, positive effects of refutation texts in this study could be attributed to the testing format, as the conceptual knowledge tests included true/false statement items, presented on the same test page. This is in some ways similar to the refutation text, in which correct and misconceived knowledge structures are presented next to each other and explicitly appraised. However, in the present study, students were not merely required to indicate their choice, but also to elaborate and explain it, which was used as the main source for grading decisions. Moreover, findings from separate analyses of the open test performance grades and the closed ones mirrored those of the overall measures. We are therefore fairly confident that the effect of refutation texts on conceptual understanding should not be attributed to test format.

Third, the majority of participants in this study were female students, which is typical of the population this sample was drawn from (i.e., majors in Humanities and Social Sciences). Research on engendered discourse styles has shown a female preference toward more consensual and less critical styles of discourse (e.g., Sullivan, Kapur, Madden & Shipe, 2015). Even though the evidence on this issue is very tentative at this point, instructions for deliberative argumentation, that emphasize both the critical and the collaborative aspects of argumentation for learning, have not been found to be as successful for facilitating critical discourse among female dyads (Asterhan & Schwarz, 2016). As gender was not controlled for in this study, its role during the argumentation phase could not be systematically explored.

Lastly, the design of the current study did not include a condition in which argumentation preceded refutation text reading. It is therefore not possible to draw any conclusions about the comparative effectiveness of refutation texts or argumentation in isolation, or whether refutation texts are more effective than dyadic peer argumentation. A future study that directly compares between argumentation with and without a refutation text would be needed to explore whether and how effects of argumentation may be augmented with refutation texts.

5. References

Alvermann, D. E., & Hague, S. A. (1989). Comprehension of counterintuitive science text: Effects of prior knowledge and text structure. *Journal of Educational Research, 82,* 197–202.

Ariasi, N., & Mason, L. (2011). Uncovering the effect of text structure in learning from a science text: An eye-tracking study. *Instructional science, 39*(5), 581-601.

Asterhan, C. S. C. (2013). Epistemic and interpersonal dimensions of peer argumentation: Conceptualization and quantitative assessment. In: M. Baker, J. Andriessen & S. Jarvela (Eds), *Affective learning together* (pp. 251-272). New York, NY: Routledge, Advances in Learning & Instruction series.

Asterhan, C. S. C. & Babichenko, M. (2015). The social dimension of learning through argumentation: Effects of human presence and discourse style. *Journal of Educational Psychology, 107*(3), 740-755.

Asterhan, C. S. C., & Dotan, A. (2018). Feedback that corrects and contrasts students' erroneous solutions with expert ones improves expository instruction for conceptual change. *Instructional Science, 46,* 337–355.

Asterhan, C. S. C. & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist, 51*(2), 164-187.

Asterhan, C. S. C. & Schwarz, B. B. (2009). The role of argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialogue. *Cognitive Science, 33*, 373-399.

Asterhan, C. S. C. & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology, 99,* 626-639.

Babai, R., Sekal, R., & Stavy, R. (2010). Persistence of the intuitive conception of living things in adolescence. *Journal of Science Education and Technology, 19,* 20-26.

Beker, K., Kim, J., Van Boekel, M., van den Broek, P., & Kendeou, P. (2019). Refutation texts enhance spontaneous transfer of knowledge. *Contemporary Educational Psychology, 56,* 67-78.

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching, 49,* 68-94.

Braasch, J. L., Goldman, S. R., & Wiley, J. (2013). The influences of text and reader characteristics on learning from refutations in science texts. *Journal of Educational Psychology*, *105*(3), 561.

Chambliss, M. J. (2002). The characteristics of well-designed science textbooks. In: Otero, J., & Graesser, A. C. (Eds.). (2014). *The psychology of science text comprehension* (pp. 51-72). Routledge.

Chan, C., Burtis, J., & Bereiter, C. (1997). Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction*, *15*(1), 1-40.

Chi, M. T. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In: S. Vosniadou, S. (Ed.), *Handbook of research on conceptual change* (pp 61-82). Hillsdale, NJ: Erlbaum.

Chi, M. T., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science*, *36*(1), 1-61.

Cordova, J. R., Sinatra, G. M., Jones, S. H., Taasoobshirazi, G., & Lombardi, D. (2014). Confidence in prior knowledge, self-efficacy, interest and prior knowledge: Influences on conceptual change. *Contemporary Educational Psychology*, *39*(2), 164-174.

De Vries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. *Journal of the Learning Sciences*, *11*, 63-103.

Diakidoy, I. A. N., Kendeou, P., & Ioannides, C. (2003). Reading about energy: The effects of text structure in science learning and conceptual change. *Contemporary Educational Psychology*, *28*(3), 335-356.

Diakidoy, I. A. N., Mouskounti, T., Fella, A., & Ioannides, C. (2016). Comprehension processes and outcomes with refutation and expository texts and their contribution to learning. *Learning and Instruction*, *41*, 60-69.

Diakidoy, I. A. N., Mouskounti, T., & Ioannides, C. (2011). Comprehension and learning from refutation and expository texts. *Reading Research Quarterly*, *46*(1), 22-38.

Donovan, A.M., Zhan, J., & Rapp, D.N. (in press) Supporting historical understandings with refutation texts. *Contemporary Educational Psychology*.

Durkin, K. & Rittle-Johnsson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning & Instruction, 22*, 206-214.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Felton, M., Garcia-Mila, M., & Gilabert, S. (2009). Deliberation versus dispute: The impact of argumentative discourse goals on learning and reasoning in the science classroom. *Informal Logic*, *29*(4), 417-446.

Ferrari, M. & Chi, M. T. H. (1998). The nature of naive explanations of natural selection. *International Journal of Science Education, 20*, 1231-1256.

Franco, G. M., Muis, K. R., Kendeou, P., Ranellucci, J., Sampasivam, L., & Wang, X. (2012). Examining the influences of epistemic beliefs and knowledge representations on cognitive processing and conceptual change when learning physics. *Learning and Instruction*, *22*(1), 62-77.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, 117-159.

Guzzetti, B. J., Williams, W. O., Skeels, S. A., & Wu, S. M. (1997). Influence of text structure on learning counterintuitive physics concepts. *Journal of Research in Science Teaching*, *34*(7), 701-719.Hynd, C., & Guzzetti, B. J. (1998). When knowledge contradicts intuition: Conceptual change. In C. Hynd (Ed.), *Learning from text across conceptual domains* (pp. 139-164). Mahwah, NJ: Lawrence Erlbaum.

Jensen, M. S., & Finley, F. N. (1996). Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *Journal of Research in Science Teaching*, *33*(8), 879-900.

Jiménez-Aleixandre, M. P. (1992). Thinking about theories or thinking with theories? A classroom study with natural selection. *International Journal of Science Education*, *14*(1), 51-61.

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) Framework: Processes and Mechanisms. In D.N. Rapp & J.L.G. Braasch (Eds.). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge, MA: MIT Press.

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. (2014). Knowledge revision processes in refutation texts. *Discourse Processes, 51*, 374-397.

Kendeou, P., & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. Memory & Cognition, 35, 1567-1577.

Lombardi, D., Danielson, R. W., & Young, N. (2016). A plausible connection: Models examining the relations between evaluation, plausibility, and the refutation text effect. *Learning and Instruction*, *44*, 74-86.

Mason, L., Baldi, R., Di Ronco, S., Scrimin, S., Danielson, R. W., & Sinatra, G. M. (2017). Textual and graph refutations: Effects on conceptual change learning. *Contemporary Educational Psychology*, *47*, 275-288.

Mason, L., Gava, M., & Boldrin, A. (2008). On warm conceptual change: The interplay of text, epistemological beliefs, and topic interest. *Journal of Educational Psychology*, *100*(2), 291.

Mason, L., Zaccoletti, S., Carretti, B., Scrimin, S., & Diakidoy, I. (in press). The role of inhibition in conceptual learning from refutation and standard expository texts. *International Journal of Science and Mathematics Education*.

Masson, S., Potvin, P., Riopel, M., & Foisy, L. M. B. (2014). Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity. *Mind, Brain, and Education*, *8*(1), 44-55.

Mikkilä-Erdmann, M. (2001). Improving conceptual change concerning photosynthesis through text design. *Learning and Instruction, 11*, 241-257.

Nussbaum, E. M., & Sinatra, G. M. (2003). Argument and conceptual engagement. *Contemporary Educational Psychology*, *28*(3), 384-395.

Ohlsson, S. & Bee, N. V. (1992). *The effect of expository text on students' explanations of biological evolution.* OERI Report. Learning Research and Development Center, University of Pittsburgh.

Osborne, J. (2010). Arguing to Learn in Science: The role of collaborative, critical discourse. *Science*, *328*, 463.

Palmer, D. H. (2003). Investigating the relationship between refutational texts and conceptual change. *Science Education, 87*, 663-684.

Potvin, P., Masson, S., Lafortune, S., & Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference. *International Journal of Science and Mathematics Education*, *13*(1), 21-43.

Potvin, P., Sauriol, É., & Riopel, M. (2015). Experimental evidence of the superiority of the prevalence model of conceptual change over the classical models and repetition. *Journal of Research in Science Teaching*, *52*(8), 1082-1108.

Ramsburg, J. T., & Ohlsson, S. (2016). Category change in the absence of cognitive conflict. *Journal of Educational Psychology*, *108*(1), 98.

Schnotz, W., & Preuß, A. (1999). Task-dependent construction of mental models as a basis for conceptual change. In: W. Schnotz, S. Vosniadou & M. Carretero (Eds), *New perspectives on conceptual change*. Amsterdam: Pergamon Press

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*(2), 209-215.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, *52*(2), 170-194.

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly*, *46*(4), 374-393.

Sinatra, G. M., Brem, S. K., & Evans, E. M. (2008). Changing minds? Implications of conceptual change for teaching and learning about biological evolution. *Evolution: Education and outreach*, *1*(2), 189-195.

Trevors, G., & Muis, K. R. (2015). Effects of text structure, reading goals and epistemic beliefs on conceptual change. *Journal of Research in Reading*, *38*(4), 361-386.

Van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: the role of co-activation in confronting misconceptions. *Applied Cognitive Psychology*, 22(3), 335-351.

Van Loon, M. H., Dunlosky, J., Van Gog, T., Van Merriënboer, J. J., & De Bruin, A. B. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology*, *42*, 39-48.

Vosniadou, S., & Mason, L. (2012). Conceptual change induced by instruction: a complex interplay of multiple factors. In S. Graham, J. Royer, & M. Zeidner (Eds.), *Individual differences and cultural and contextual factors*, Volume 2 (pp. 221-246). In K. Harris, S. Graham, & T. Urdan (Eds.), APA Educational Psychology Handbook Series. APA Publications.

Weingartner, K. M., & Masnick, A. M. (2019). Refutation Texts: Implying the Refutation of a Scientific Misconception Can Facilitate Knowledge Revision. *Contemporary Educational Psychology*, *58*, 138-148.

Table 1

The phenomena tested on each of the three test occasions as a function of test item type

Item type	*Pre-test*	*Immediate posttest*	*Delayed posttest*
True/False	Sea iguanas (swimming ability)	Polar bear (fur coloration)	Angler fish (esca)
True/False	Guppy fish (coloration)		Lizard (head shape)
Open	Cheetah (running speed)	Turtle (shell shape)	Armadillo (nasal features)

Table 2

*Six principles of biological evolution targeted in the assessment**

Principle	*Description*
Intra-species diversity	Individuals in one generation of a particular species differ from each other on many dimensions.
Source of diversity and change	This diversity is the result of random changes in genetic material (sexual recombination and random mutations) and not a result of a need or necessity that is intentionally addressed by the individual changing the trait.
Inheritance of traits	Genetic traits are passed on from parent to offspring, regardless of whether they are beneficial or not.
Learnt behaviors	Learnt behaviors and other changes in phenotype that are acquired during an individual's lifetime are not genetically passed on to offspring.
Survival	Individuals within a certain species which have advantageous traits survive longer and reproduce more.
Proportional vs typological change	When this selection process is repeated over many generations, the accumulated effect is an increase in the proportion of individuals carrying the advantageous traits, whereas the proportion of those without the traits decreases.

* Adapted from Asterhan & Dotan (2018), Ferrari & Chi (1998), Ohlsson & Bee (1992) and Shtulman (2006)

Table 3

Mean conceptual understanding scores (and SD), per experimental condition

	Ref+Arg	*RefOnly*	*Control*
	N = 50	*N* = 24	*N* = 26
Pre-test	44.14 (26.02)	42.61 (28.57)	37.26 (24.04)
Immediate posttest	65.63 (30.35)	73.14 (26.65)	42.13 (35.43)
Delayed posttest	69.14 (26.52)*	70.89 (24.44)**	45.55 (28.08)

* *N* = 48, ** *N* = 22

Table 4

Mean normalized gain scores (and SD) for conceptual understanding, per experimental condition

Normalized gain score	*Ref+Arg*	*RefOnly*	*Control*
Pre-test → immediate posttest	39.86 (48.62)	50.18 (35.73)	-1.94 (52.77)
Immediate → delayed posttest	14.40 (35.16)	8.71 (38.60)	3.94 (30.41)
Pre-test → delayed posttest	45.06 (37.29)	48.37 (35.55)	9.32 (40.60)

Table 5

Mean normalized gain scores (and SD) for false and for true statement test items separately, per experimental condition

Normalized gain score	*Ref+Arg*	*RefOnly*	*Control*
False statements	47.88 (36.90)	51.21 (29.76)	17.49 (30.04)
True statements	47.38 (51.29)	56.21 (45.00)	6.02 (51.29)

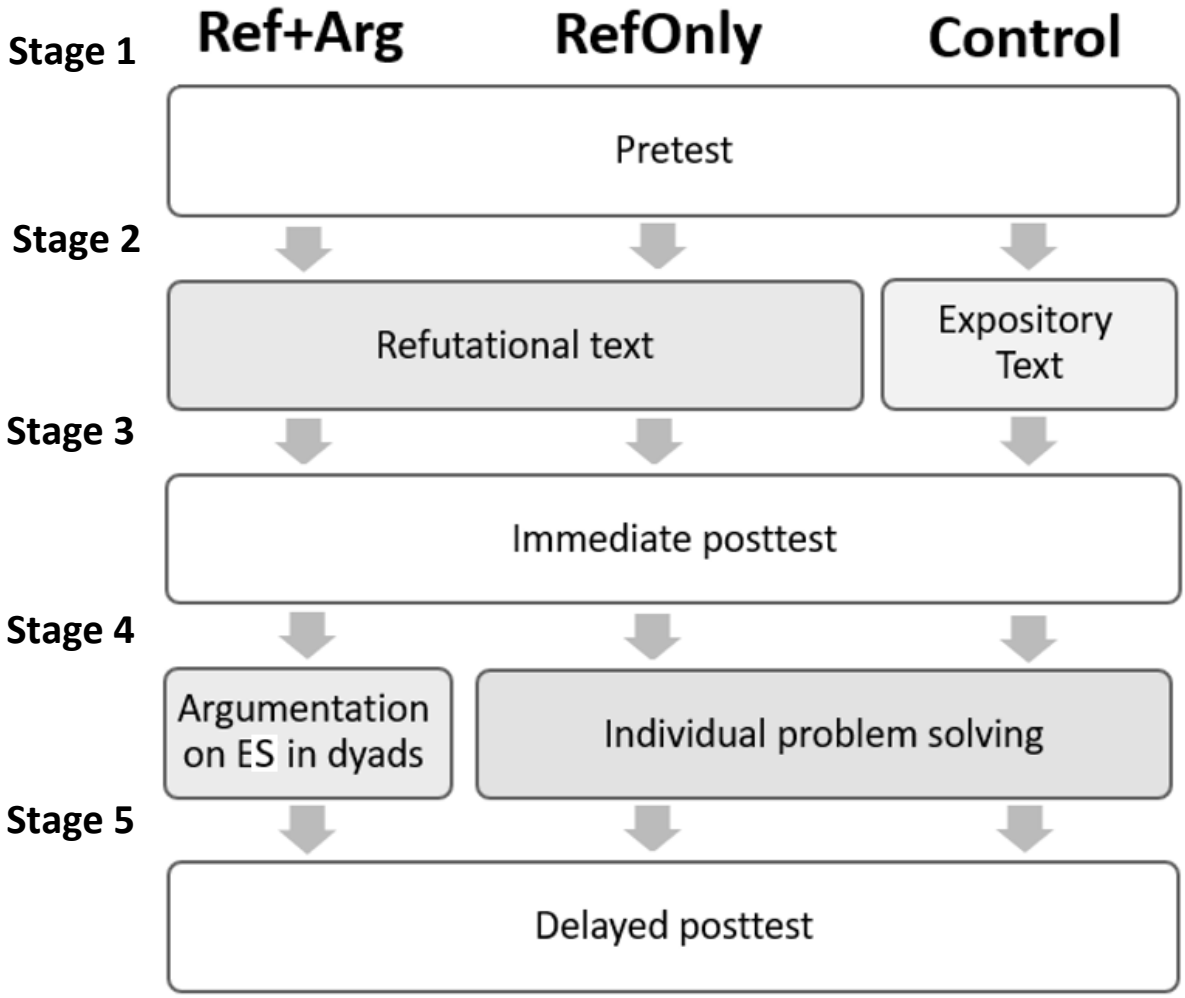
Table 6

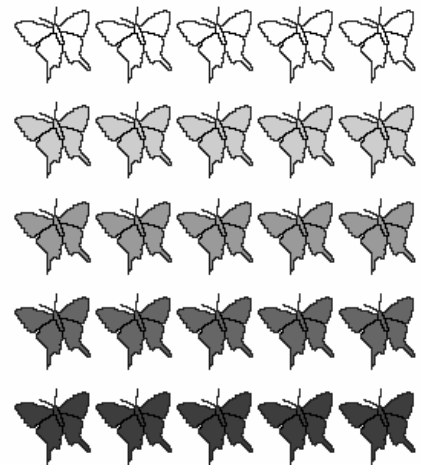
Dialogue features of gaining and non-gaining dyads

Dialogue features		*Non-gaining dyads* (*N* = 10)	*Gaining dyads* (*N* = 10)	
Critical discussion	No	5	2	χ (1, 20) = 1.98, *p* = .160
	Yes	5	8	
Symmetry	No	8	4	χ (1, 20) = 3.33, *p* = .068
	Yes	2	6	
Nr. of core principles	<4	8	2	χ (1, 20) = 7.20, *p* = .007
	>5	2	8	

Figure 1. Stages of the experimental design

Figure 2. Visualization of a common erroneous solution to the peppered moths question (adapted from Shtulman, 2006)


Stage 1
Ref+Arg
RefOnly
Control
Pretest
Stage 2
Refutational text
Expository Text
Stage 3
Immediate posttest
Stage 4
Argumentation on ES in dyads
Individual problem solving
Stage 5
Delayed posttest



Appendix A

Examples of assessment items

a) Example from the open question item type.

The cheetah. When chasing prey, the running speed of a cheetah can reach 95 km/h, or more. This makes it the fastest predator in its natural habitat. The cheetah's ancient predecessors, on the other hand, were significantly slower, with a maximum speed of 3B km/h at the most. How would natural selection account for and explain this change? Please explain and describe the process of change in your own words.

2) Example from the forced choice item (with explanations) type.

The sea iguana. The Galapagos Islands habit a special type of iguana, the sea iguana, which differs from the land iguana, even though they share common ancestors. In contrast to the land iguana, sea iguanas are excellent swimmers. They dive into the depths of the sea, are able to hold their breath for a very long time and feed on seaweed, which they are able to digest.

Following are five statements regarding sea iguanas. Please circle whether a statement is true or false and explain your choice in an elaborated manner.

(1) Individual sea iguanas in the Galapagos Islands share many central traits, but they are also different from each other.
Right / Wrong. Explain: _____

(2) Changes started to occur in the iguanas' physique because they needed them to be able to swim.
Right / Wrong. Explain: _____

(3) Iguanas that have excellent hearing, which does not give them a survival advantage, will still transmit this genetic information to their offspring.
Right / Wrong. Explain: _____

(4) An iguana that, during its lifetime, has learned to wag its tail in order to advance in water, will pass this trait on to their offspring.
Right / Wrong. Explain: _____

(5) Iguanas that were less successful in the competition for food were more likely to die of starvation or malnutrition at an earlier age.
Right / Wrong. Explain: _____

Appendix B

Excerpt from the refutation text showcasing the difference with the expository text

The following excerpt shows the illustrative example that was included in the text (the evolution of the giraffe's neck). The underlined text only appeared in the refutation text version, whereas the remainder appeared in both versions:

> "For example, many people think that in order to improve their chances of survival, individual giraffes exerted extra efforts to reach the highest branches for food. They believe that giraffes intentionally managed to extend their necks by exerting effort and will during their lifetime, and then passed this acquired trait to their offspring. This is a mistake.
> The correct explanation is that in each generation, giraffes with various neck lengths are born (intra-species variability in a population). Giraffes that were born with slightly longer necks had better access to food on the higher branches, which short-necked giraffes could not reach. Thus, this feature was advantageous in achieving food and allowed them to survive longer, mate more and have more offspring. Over the generations, the proportion of longer necked giraffes in the population increased, while the proportion of giraffes with shorter necks gradually diminished.
> In summary, natural selection explains the slow, gradual change of traits through generations. It is important to remember that an individual's chances of survival depend on the extent to which its traits are a good fit (adaptive) to the environment, as environmental conditions are subject to change."

Appendix C

Illustrating student references to core principles of biological evolution (Table 2) in the dialogue protocols

Speaker	*Verbal content within a speaker turn*	*Referring to conceptual principle*
A	Uhmmm… All in all it's pretty similar only that… like because of the soot of the industrial revolution walls, let say, became darker. The soot accumulated so to avoid predators the moths needed to be darker to disguise better, to assimilate better with the walls that had become darker, and again, they say here that from one generation to the next, they became slightly darker and slightly darker, so it's like... and it's not necessarily because...	
B	No it's definitely not	
A	It's definitely not because every white individual passes on in general to its offspring to also be white and it's not a continuous process and then again	3
B	It's a matter of eh… it's a matter of eh... battle for survival, like those that-	5
A	-Yes, that's the natural selection-	
B	- those that are more -	1
A	No, but when he says that each individual… I will pass on to my children that they will be slightly darker than me and my grand-children will be slightly darker than them. So no, because they pass on to their children more or less the same genes, and whenever there is an incidental genetic change, then -	3, 2
B	But the issue is that not all the moths go through this process together, meaning it's not that in every generation a moth is born that…	
A	In principle, in every generation a moth is born in the exact colour of his parents except for a one-in-a-million deviation	2

B	No, those that are darker then they will succeed in surviving better and therefore they will manage to thrive more	1, 5
A	Exactly	
B	It's a matter of competition between the moths, it's not that all the moths change together but just the strongest ones.	6
A	Yes, heredity doesn't work in a way that... that...	
B	That every time it changes a bit more	