Refutation Text and Argumentation to Promote Conceptual Change

Christa Asterhan, Hebrew University Jerusalem, Israel, asterhan@huji.ac.il
Maya Resnick, Hebrew University Jerusalem, Israel, MayaResnick@mail.huji.ac.il

Abstract: In the present work, we examine the accumulative effect of two instructional methods for conceptual change, refutation text and argumentation, which are expected to support two complementary processes that, according to current models, underlie conceptual change: Promoting awareness to and reducing interference of irrelevant knowledge structures and sense-making of the counterintuitive, scientifically accepted notions. Hundred undergraduates were randomly assigned to either read a refutation text and then conduct a peer discussion (Ref+Arg), or to read a refutation (RefOnly) or an expository (Control) text, followed by a standard, individual problem solving task. Results showed strong effects for refutation texts on both immediate and delayed post-tests. Contrary to expectations, subsequent peer argumentation did not further improve learning gains. Dyadic dialogue protocols analyses showed that gaining dyads were overall more likely to be symmetrical and to include a discussion of the core principles of natural selection. Several directions for future research are discussed.

Introduction

For more than four decades, scholars from science education, developmental psychology, the learning sciences and cognitive science have documented how children's and adults' naïve theories about natural phenomena do not align with the scientifically accepted, but often counter-intuitive concepts that they are exposed to in science instruction. Coming to understand and being able to correctly use these canonical, scientific explanations is not a matter of "gap-filling", in that learners lack the necessary knowledge, but rather involves a substantive re-organization of existing knowledge structures, an outcome which is usually referred to as "conceptual change" (e.g., Chi, 2008; Thagard, 1992; Vosniadou & Brewer, 1994). Current cognitive accounts of conceptual change describe it in terms of a response competition at a deeper cognitive level. Accordingly, it constitutes an increase in the probability with which more advanced schema configurations are activated and used to construct temporary mental representations in working memory, when an individual is required to apply that knowledge to solve a problem (e.g., Potvin, Sauriol, & Riopel, 2015; Ramsburg & Ohlsson, 2016; Schnotz & Preuss, 1999). This response competition account is further supported by recent empirical evidence showing that conceptual change involves both an improved capability to construct the correct scientific explanation, as well as more efficient inhibition of automatically activated, but irrelevant schemas and propositions (e.g., Babai, Sekal & Stavy, 2010; Dunbar, Fugelsang & Stein, 2007; Masson, Potvin, Riopel, & Foisy, 2014; Shtulman & Valcarcel, 2012; Potvin, Masson, Lafortune & Cyr, 2015).

In order to be effective, instruction for conceptual change should then preferably support both these cognitive processes: to provide students with opportunities to become aware of and understand the errors in (their) naïve theories, as well to fully comprehend the scientifically accepted theory that is often counterintuitive to everyday experiences (see also Chan, Burtis & Bereiter, 1997). Not surprisingly, traditional tell-and-practice teaching approaches have not been found to be very effective for learning that requires conceptual change, especially in the case of robust misconceptions (Chi, 2008; Vosniadou & Mason, 2013). Researchers of instructional approaches for conceptual change have then studied the effectiveness of alternative instructional techniques, materials and activities, such as refutation texts, argumentation and modeling. Traditionally, each of these have been studied in isolation, however, which may perhaps explain the overall modest effect sizes in this field. In the present work, we examine the accumulative effect of two instructional methods for conceptual change, namely refutation text and argumentation, which are expected to complement each other they support the two aforementioned processes underlying conceptual change: Promoting awareness to and reducing interference of irrelevant knowledge structures (refutation text) and sense-making of the counterintuitive, scientifically accepted notions (collaborative argumentative dialogue).

Refutation texts and conceptual change

Much of what students learn in science classes still comes from textbooks, which are predominantly structured as expository texts of factual information on a scientific concept, without directly referring to common misconceptions (Osborne, 2010; Tippet, 2010). However, research on text reading has shown that expository science texts induce superficial processing and fail to support deep learning (Chambliss, 2002; Diakidoy, Kendeou & Ioannides, 2003). Refutation texts, on the other hand, provide an explicit statement of commonly held

misconceptions, directly refutes them, and then introduces scientific explanations as alternatives (Sinatra & Broughton, 2011; Tippett, 2010; Vosniadou & Mason, 2012). Studies have shown that when learning requires the restructuring of prior incorrect knowledge, refutation texts are generally more beneficial than standard expository science texts (e.g., Braasch, Goldman, & Wiley, 2013; Diakidoy et al., 2003; see Guzzetti et al, 1993 and Tippett, 2010 for reviews). Van den Broek and Kendeou (2008) have proposed that the presentation of the two side-by-side provides the reader with the opportunity to compare the two conceptions, detect inconsistencies and "revise" knowledge structures (the co-activation hypothesis). Findings from eye-tracking data while reading further supports this explanation (e.g., Ariasi & Mason, 2011; Kendeou, 2011).

However, not all studies have found an advantage of refutation over expository texts (see for example, Diakidoy et al., 2016; Hynd & Guzzetti, 1998; Mason, Gava, & Boldrin, 2008; Mason, Zaccoletti, Carretti, Scrimin Palmer, 2003). In addition, the strongest effects for refutation texts are typically reported on less complex topics, such as whether ostriches bury their heads in the sand, or not (see Tippet, 2010; Van Loon et al., 2015). These types of erroneous ideas can be relatively easily refuted by correcting a mistaken belief and do not require revisions or restructuring of complex knowledge systems. Based on the available research, it is not clear to what extent reading a refutation text can in by itself induce conceptual change of robust misconceptions on particularly complex science concepts, as considerably fewer studies have focused on them (but see for example Diakidoy et al., 2003 for an exception). In the present study, we compare the effect of reading a refutation *vs*. an expository text on a complex science topic for which students are known to have particular robust misconceptions (i.e., natural selection).

Finally, findings from two recent studies (Diakidoy et al., 2016; Van Loon et al., 2015) seem to indicate that refutation texts may be particularly effective in neutralizing the interfering influence of misconceptions, what Van Loon et al. (2015) have termed 'outdating'. In both studies, findings suggest that refutation texts may not be very effective in helping students to make sense of and construct scientifically correct explanations on their own ('updating'). Indeed, Vosniadou and Mason (2012) argue that "(r)efutation texts (…) must be used together with other instructional interventions in the context of a rich learning environment that fosters and sustains conceptual change" (p. 35). Accordingly, in order to ensure deep and long-lasting change on complex science topics, students should be given opportunities to make sense of, explain and practice their newly acquired understanding, after reading the refutation text. This proposition is tested in the current study, in which we compare the effects of reading a refutation text on a complex scientific topic (i.e., natural selection) with the effects of refutation text reading followed by subsequent dyadic argumentation.

Argumentation and conceptual change

Research has shown that peer argumentation can be an effective means for conceptual change type of learning (see review by Asterhan & Schwarz, 2016). However, productive peer argumentation, when two (or more) discussants compare and weigh different explanations through reasoned argument in a constructive atmosphere (Asterhan & Babichenko, 2015), is not easily elicited nor sustained. Close inspections of dialogue protocols reveal that often times participants simply do not detect when explanations are conceptually different (Sfard, 2009) and therefore may not feel the need to further explore them. The explanations students propose during discussions are often times short, shallow and/or partial. Even when they are using identical terminology, they may implicitly be attaching different meanings to it (Sfard, 2009). This lack of detailed information and ambiguous language use may create an 'illusion of consensus' between dialogue participants, believing that they are in fact proposing similar solutions and are in agreement.

Task design should then make sure that the differences between misconceptions and scientifically accepted theories are presented in a more salient and comprehensible manner so that student have access to them during argumentation. In the present study, we support student argumentation on scientific concepts by presenting them with erroneous worked-out examples (Asterhan et al., 2015; Durkin & Rittle-Johnsson, 2011) that are based on common misconceptions. Student dyads are asked to solve and correct through argumentation, which is expected to further help them in inhibiting irrelevant knowledge structures, as well as with understanding, consolidating, and strengthening the correct concept ('updating').

Hypotheses

Based on the aforementioned rationale the following hypotheses are formulated:

H1: Students who read a refutation text on natural selection will show larger learning gains than students who read an expository text on the same topic.

H2: Students who read a refutation text and then participate in peer argumentation on erroneous worked-out examples will show greater conceptual gains than students who will read refutation texts and then complete a standard individual problem-solving activity.

H3: Reading refutation texts will improve students' capability to identify and correct errors ('outdating'), whereas peer argumentation will further improve students' capability to identify and construct correct explanations ('updating').

H4: Dyads who conduct a critical discussion on the core principles of natural selection will show greater gains from the peer collaboration session than dyads who do not.

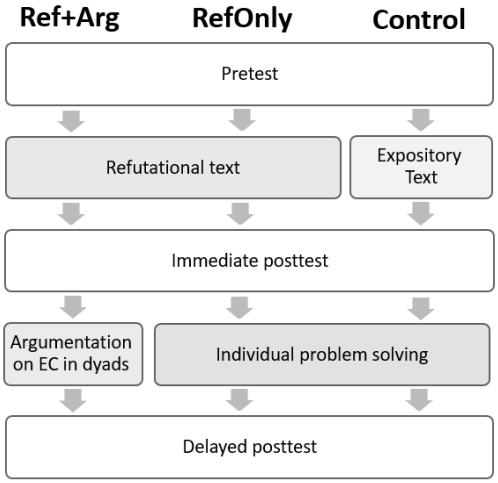
Method

Participants

One hundred undergraduate students (73 F, *M*age = 24.5) from a large university in Israel participated in the study. Requirements for participation were a lack of formal background in the Life and Exact Sciences and proficiency in the Hebrew language. Participants could choose between receiving course credit (31%) or financial reimbursement (approximately $15) for participation. Four participants failed to appear for the delayed posttest (2 from the Ref+Arg and 2 from the RefOnly condition). The pretest and immediate posttest scores of these participants were not found to be different from the remainder and their data was therefore omitted only from analyses that include the delayed posttest scores.

Design

A 1X3 between-subject experimental design was used. Participants were randomly assigned to one of three conditions (see Figure 1): (1) Refutation text + dyadic argumentation on erroneous solutions (Ref+Arg; N = 50); (2) Refutation text + individual problem solving (RefOnly; N = 26); (3) Expository text + individual problem solving (Control; N = 24). Individual conceptual understanding of natural selection was assessed on pre-test, immediate post-test (following text reading), and delayed post-test (a week later).



Tools

Demographics and background. The background questionnaire targeted the following demographic information: gender, age, degree, field of major, religious affiliation, degree of religiosity, and background in high school biology education and evolution. In addition, students' attitudes and beliefs regarding the theory of natural selection were assessed with 4 Likert scale items, ranging from 1 ("*I do not agree at all*") to 5 ("*I completely agree*"), which were translated to Hebrew from Shtulman (2006). Examples are "*Natural selection is the best explanation for the creation of species*" and "*Species change over time*". Internal reliability for the attitude measure was high, Cronbach's α = .84.

Conceptual understanding of natural selection. Individual conceptual understanding of natural selection at pre-test, immediate post-test and delayed post-test was assessed with open and closed items that targeted the evolution of selected animal traits. Test items were adapted from Asterhan et al. (2015). Eight different animals and traits were distributed over the three tests. The pretest and the delayed posttest each included questions about 3 different animals: Six different false/correct statements about the first animal, five about a second animal and one open-ended question about the third. The immediate post-test included questions about 2 animals: Six true/false statements about the first animal and one open-ended question about a second animal. For each animal, students were presented with a short text about a specific animal species, a physiological change in a specific trait over time and a short description about the importance of that trait. In the false/correct items, this text was followed by five (or six) statements, which each addressed a different principle of natural selection (e.g., intra-species variability, proportional change). Students were required to indicate whether the statement was false or correct and then explain and justify their choice in their own words. In the pretest and delayed posttest, which included sets of false/correct statements on 2 different animals, when a given principle was presented as a correct statement for the first animal, the principle was presented as an incorrect statement for the second animal. For a given animal, approximately half of the statements presented were correct and half incorrect. In the open item questions, students

were required to give a full explanation of how they think the given trait had evolved, according to the theory of natural selection.

Instructional texts. Two instructional texts on the topic of natural selection were created: an expository text (453 words in Hebrew) and a refutation text (525 words in Hebrew). Both were identical with regard to the background information and the correct explanations of natural selection, which related to the existence of differences within species and to the evolution of traits over time. The relevant terminology was explained in plain language and they included an elaborated description of a well-known example of change in a specific feature of a specific species, namely the increasing length of the giraffe's neck. The only difference between the texts was that while the expository text only presented the scientifically accepted theory, explanation and example, the refutation text in addition included (1) explicit references to common misconceptions (intentional change, intra-species diversity and the source of diversity and change); and (2) explicit statements that those are "wrong" (the refutation cue). The remainder of the text was verbatim identical.

Worksheets. The worksheet booklet students used in stage 4 of the experiment was based on materials used in a previous study (Asterhan et al., 2015), and consisted of three open-ended questions about two novel evolutionary phenomena, namely the webbed feet of ducks and the wing coloring of the peppered moth. Each question appeared on a separate page. 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 natural selection. In the third task, which was adapted from Shtulmann (2006), students were asked to depict the gradual change in wing coloring of the peppered moth population in a graphical manner. This item is used to distinguish between typological and selection-based representation of change.

Students in the Ref+Arg condition received the same booklet, with two changes: First, the space for writing the solutions was already filled with a solution provided by a (fictitious) peer student. The textual solutions were handwritten (each in a different handwriting) or hand-coloured. 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 misconceptions that individual animals intentionally changed a trait during their lifetime and that acquired changes in traits are passed on (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 coloured by a (fictitious) peer, to depict a classic typological model of change. Second, a separate space was reserved at the bottom of each page for the participants to fill in the corrected solution in their own handwriting and their own colouring.

Procedure

Except for stage 4 (problem solving stage), all stages of the experimental sessions were conducted in individual, separate rooms. Following a brief verbal explanation about the experiment, students filled in the background information survey, followed by the pretest. They were instructed to answer in full and elaborate their answers as much as possible, even when they were uncertain of their answers. Following (stage 2), each participant received either the refutation text (Ref+Arg and RefOnly conditions) or the expository text (control condition) about natural selection. They were told that the text presents the scientifically accepted explanation and recieved 10 *min* to study the text in detail. Following the reading phase, participants completed the immediate posttest individually (stage 3). They were then moved to a different room shared with another participant (stage 4). Assignment to peer participant was random within condition. Participants in the RefOnly and the control condition were told that the reasons for this move were logistic as the room was needed for another experiment. They 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 each other. Participants in the Ref+Arg condition were instructed to work in pairs to critically, yet constructively, discuss the erroneous student solutions in the filled-in worksheets. The dyadic argumentation sessions were audio-recorded. Upon completion of the discussion, each participant received a copy of the worksheets they previously discussed with room to correct the presented mistakes individually. A week later, participants returned for the final stage to complete the delayed posttest questionnaire and received a debriefing and reimbursement for their participation.

Coding

Coding understanding of evolution. Coding of students' written solutions to the test items was based on existing coding procedures developed in previous studies (Asterhan et al., 2015; Asterhan & Dotan, in press). Each written solution was graded according to accuracy and compliance with the main principles of natural selection: 0 for omissions, misconceptions or other crucial errors, .5 for partially correct solutions, or full credit (1) for solutions that contained no misconceptions and addressed the main tenets of natural selection correctly.

Each false/correct item targeted one of the six predefined principles of natural selection (Intra-species diversity, Source of diversity and change, Inheritance of traits, Learnt behaviours, Survival, Proportional vs typological change). When coding for the closed questions, two factors were considered in one single score: the indicated choice of right or wrong and the accompanying textual explanation. A correct choice between right and wrong together with a correct and sufficient explanation resulted in full credit. An incorrect right/wrong choice with an incorrect explanation, resulted in zero points. When these two components were not synchronized, more weight was given to the verbal explanation. Most of these cases indeed showed partial understanding and were given a score of 0.5, but there were several cases that showed clear misconceptions in the written explanations in spite of choosing the correct answer, and received 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, in press), instead of a particular principle. 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 (Asterhan & Schwarz, 2007), such as solutions that refer to existing intra-species variability on an "ability to change" together with a selection mechanism of those members who had managed to change themselves. After a period of training and discussion, three human coders scored the same 248 randomly chosen item responses. Interrater reliability was good, .72 < Cohen's Kappa < .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).

Coding of dialogues. Discussions of the 25 dyads in the Ref+Arg condition were audio recorded. Twenty-three discussions could be transcribed (1 was incomprehensible and 1 was never recorded due to technical failure). The mean length of these audio recorded discussions was 8 *min* and 16 *sec.* Transcriptions included all verbal content, as well as significantly long pauses, and laughing, but not intonation and other auditory features. Discussion protocols were coded as a whole. Following findings from previous work on the association between dyadic argumentation and conceptual change (e.g., Asterhan & Schwarz, 2009), initial coding efforts focused on two discussion characteristics: whether the discussion could be characterized as critical, dialectical overall and whether the interaction was symmetric. A third characteristic, rhetoric style (disputative or deliberative argumentation, Asterhan & Babichenko, 2015), proved to be irrelevant: Perhaps due to the fact that the argumentation instructions were explicitly modelled on and directed toward deliberation, clear cases of disputative argumentation were near non-existent. Following these top-down coding efforts with existing tools, all the discussion transcripts were read once more, so as to search the data set for additional features that were salient. This procedure yielded and additional coding category, namely the extent to which student dyads discussed the (six) core conceptual principles of natural selection or not.

In sum, three dialogue characteristics were coded: (1) Critical discourse (0, 1) – when the students overtly confronted 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 each of the two discussants exceeds 35 % of the total discussion word count (excl. repetitions) the discussion was deemed symmetrical; and (3) Discussion of core principles (0, 1) – when at least 5 of the 6 principles of natural selection came up during the discussion the grade 1 was assigned, when 4 or less core principles were mentioned it received the grade 0 (N. B. the erroneous examples referred to three core principles altogether). Two raters scored all the dialogue protocols independently, Cohen's Kappa = .75.

Results

Normalized (Hake) 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. No differences were found on pretest scores between the different conditions, F (2, 97) = .60, p = .560. Neither of the three control variables (i.e., attitudes toward natural selection, religiosity, and perceived understanding of evolution) yielded differences across conditions, nor did they correlate with normalized pre-to delayed posttest gain scores (r = .13, r = -.17 and r = -.09, *ns*, respectively).

Effects of condition on conceptual understanding

All univariate analyses of variance presented here were conducted on the normalized gain scores of the full data set. Levene's tests for equality of error variance across compared conditions showed that this assumption was not violated in any of the reported statistical analyses ($p > .80$). To test the stability of the results and the reliability of the chosen method for analyses, we reran each comparison with two additional statistical models (ANOVA on raw delta scores and ANCOVA with pretest or immediate posttest as covariates), as well as on the data set while excluding (eight) participants with high pretest scores (> .85; two in the control, two in the RefOnly and four in the Ref+Arg condition). These yielded identical results and are therefore not reported on further.

The effect of refutation text on conceptual gains. In order to examine the effect of text type (refutation vs. expository) on conceptual knowledge, 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 ($M = -1.94$, $SD = 52.77$, $N = 26$), $F(1, 98) = 17.75$, $p < .001$, with a large effect size of $\eta_p = .15$. 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 ($M = -1.94$, $SD = 52.77$, $N = 26$), $F(1, 94) = 18.10$, $p < .001$, with a large effect size of $\eta_p = .16$. These findings corroborate the first hypothesis (H1), according to which refutation texts are more effective than expository texts for both short-term and long-term conceptual gains.

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 $\eta_p = .16$. Post-hoc tests with Tukey corrections showed that the gains in the control condition ($M = 9.32$, $SD = 40.60$) was significantly lower than gains in the RefOnly condition ($M = 48.37$, $SD = 35.55$, $p = .002$) and gains in the Ref+Arg condition ($M = 45.06$, $SD = 37.29$, $p = .001$). However, no differences on normalized gain scores were found between the latter two, $p = .938$. 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, according to which peer argumentation on erroneous solutions would further increase learning gains compared to individual problem solving activities and result in a significant difference between the RefOnly and the Ref+Arg conditions. Even though immediate posttest scores in these two conditions were overall fairlyhigh ($M = 65.63$ and $M = 73.14$, respectively), there was definitely room for more improvement. It therefore does not seem that the lack of effect could be attributed to a ceiling effect.

Table 1. Mean normalized gain scores (and SD) for conceptual understanding, per experimental condition

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

Outdating and updating of conceptual knowledge.

In order to test the hypotheses regarding effects of condition on out-dating of misconceived knowledge and on up-dating of correct, scientific knowledge (H3), the following outcome variables were considered: Improvement on the mean normalized gain score of the false statement items only was considered as an indication of outdating, whereas improvement on the mean normalized gain score of the true statement items only was considered as an indication updating and (van Loon et al, 2015). In addition, improvement on the open question item was also considered as an indication of updating (Kendeou & van den Broek, 2005).

Performance on true vs. false statement items. The normalized gain score from pretest to delayed test was calculated separately for the true statement items only and for the false statement items only. As the immediate posttest included only 6 (instead of 11) forced choice questions, analyses were restricted to the pre-test and the delayed post-test scores on this measure. A main effect of condition was found on the normalized false statement gain score, $F(2, 93) = 8.36$, $p < .001$, with a large effect size of $\eta_p = .15$. In alignment with hypothesis 3a, learners who had read an expository text showed significantly less improvement on the recognition and correction of false statements ($M = 17.49$, $SD = 30.03$), when compared to both the RefOnly condition ($M = 51.21$, $SD = 29.76$, $p =$

.002) and to the Ref+Arg condition (M = 17.49, SD = 30.03, p = .001). No differences were found between the two refutation text conditions, p = .922.

A main effect was also found on the normalized true statement gain score, F (2, 93) = 7.27, p = .001, with a large effect size of η_p = .14. The pattern of differences across conditions mirrored the false statement gain score: Learners who had read a refutation text improved significantly more on true statement test items (M = 47.38, SD = 51.29 for the Ref+Arg condition and M = 56.21, SD = 45.00 for the RefOnly condition) than learners in the control condition (M = 6.02, SD = 55.70), p = .004 and p = .003, respectively. Contrary to expectations, however, the addition of a dyadic argumentation activity did not result in larger gains, as no differences were found between the gains in the Ref+Arg and the RefOnly condition, p = .781.

Open Question Items. 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 (see Asterhan & Dotan, in press). Only students who had a nominal, unweighted pretest open item score of .5 or lower were included in the analyses (N = 73). Seventy-five per cent of students in the refutation text conditions showed substantive conceptual gains from pretest to delayed posttest, whereas only 45% of students in the control condition did. A Chi square test showed this difference to be significant, χ^2 (1, 69) = 5.84, p = .016, suggesting that the experimental intervention also resulted in greater improvement on the open items. A Chi square test for the differences in conceptual gains from pretest to immediate posttest, following only the reading intervention, also showed a strong significant difference between refutation text (75% improved) and expository text readers (27% improved), χ^2 (1, 72) = 14.32, p < .001. A comparison between the Ref+Arg and RefOnly conditions only showed no significant differences in improvement from pretest to delayed posttest, χ^2 (1, 47) = 1.21, p = .271, nor from immediate to delayed posttest, χ^2 (1, 30) = 2.44, p = .118.

Taken together, this set of analyses suggests that refutation texts improved not only outdating, but also updating processes, which was evident on both immediate as well as delayed tests. Argumentation with a peer did not further improve students' capability of recognizing and formulating correct scientific explanations of natural selection processes further, compared to an additional individual problem solving activity.

Dialogue protocol analyses

In order to obtain further insight in the discussion features that were associated with learning gains, the 23 available dyadic dialogue protocols in the Ref+Arg condition were analysed according to the following three criteria: Contribution symmetricity, critical discourse, and reference to core conceptual principles (see Coding section). The discussion characteristics of dyads in which none of the dyad partners showed a substantive gain from the immediate to delayed post-test were compared to dyadic discussions in which at least one dyad partners showed such gains. In three dyads, both partners had near perfect scores (> 91) and these were not included in the discussion feature analyses. Substantive gains were defined as a further increase of 30% from the immediate to the delayed post-test (i.e., normalized gain score > 30). This definition resulted in 10 “gaining” and 10 “non-gaining” dyads. Table 2 presents the cross-tabulation of dialogue features and dyad gains. Dialogues of gaining dyads were more likely to include references to core conceptual principles of natural selection, 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 gaining dyads’ discussions seemed to be more often characterized as critical (8 out of 10), but this was not significant, χ (1, 20) = 1.98, p = .160. Even though 8 out of 10 gaining dyads conducted a critical discussion, 5 of these did not result in further gains.

Table 2. Dialogue features of gaining and non-gaining dyads

		Neither partner gains (N = 10)	*At least one partner gains* (N = 10)	
Critical discussion	Yes	50%	80%	χ (1, 20) = 1.98, p = .160
Symmetry	Yes	20%	60%	χ (1, 20) = 3.33, p = .068
Nr. of core principles	>5	20%	80%	χ (1, 20) = 7.20, p = .007

Discussion

The results presented here show, first and foremost, that refutation texts are an effective means to improve conceptual understanding, not only in the case of simple, unidimensional beliefs, but also for multi-dimensional, complex science topics, such as natural selection. This advantage was evidenced immediately following the text reading, but also on delayed post-tests, a week later, suggesting that its effect was not superficial, but caused a

substantive improvement. Moreover, and in contrast to previous findings (Diakidoy et al., 2016; van Loon et al., 2015) we found that refutation texts not only improved students' capability of detecting and correcting misconceived ideas (outdating), but also their capability to construct and identify correct explanations (updating).

The expectation that subsequent sense-making activities, in the form of dyadic argumentation on erroneous examples, would further improve conceptual understanding over and above the effect of refutation text was not confirmed by the findings. As the mean scores on the immediate test did not exceed 75%, the lack of effect cannot be attributed to a ceiling effect. Further analyses of the dyadic dialogue protocols revealed that in ten of the total of 20 dyads in which at least one student could improve substantially, this in fact happened. Their dialogues were overall more likely to be symmetrical and include a discussion of the core principles of natural selection than the dialogues of those in which neither partner showed substantive gains.

Limitations and future directions

Future research should extend this research to a population of younger, school-aged children and outside of university settings in order to investigate whether refutation text without further collaborative sense-making processes could be expected to be equally effective in these age groups. It is likely that the effects of refutation texts would be weaker and that the added sense-making activities would have a larger impact in those age groups. Second, the present study did not include a condition in which argumentation preceded refutation text reading. It is therefore not possible to draw any conclusions concerning the relative effectiveness of either instructional activity in isolation (i.e., whether refutation texts are more effective than argumentation). This could be explored in future studies.

Finally, it could be argued that the strong, positive effects of refutation texts in this study could be attributed to the testing format, as the tests included a large number of true/false statements, which were presented on the same test page. This is in some ways similar to the refutation text, in which correct and misconceived knowledge structures are directly compared. Findings from previous research have also suggested that refutation texts promote performance on closed test items, but not on open items (Kendeou & van den Broek, 2005). However, in the present study student were required to elaborate their choice and these elaborations were used as the main source for grading (as opposed to the actual false/true choice). Moreover, we also included open test items in our assessment tools. Findings from separate analyses of the open test performance mirrored those of the other measures: Refutation text reading resulted in substantively more improvement than expository text, whereas subsequent argumentation did not further add to these gains. We are therefore fairly confident that the effect of refutation texts on conceptual understanding should not be attributed to test format.

Selected references

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. & 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. (in press). Feedback that corrects and contrasts students' erroneous solutions with expert ones improves expository instruction for conceptual change. *Instructional Science.*,

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

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.

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

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

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.

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

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.