International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Providing vertical coherence in explanations and promoting reasoning across levels of biological organization when teaching evolution

Janina Jördens, Roman Asshoff, Harald Kullmann & Marcus Hammann

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2016
VOL. 38, NO. 6, 960–992
http://dx.doi.org/10.1080/09500693.2016.1174790

Routledge
Taylor & Francis Group

Providing vertical coherence in explanations and promoting reasoning across levels of biological organization when teaching evolution

Janina Jördens, Roman Asshoff, Harald Kullmann and Marcus Hammann

Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Münster, Germany

ABSTRACT
Students' explanations of biological phenomena are frequently characterized by disconnects between levels and confusion of levels. The purpose of this research is to investigate the effects of a hands-on lab activity that aims at fostering the ability to reason across levels. A total of 197 students (18 years of age) participated in a randomized, pre–post-test design study. Students in the experimental group engaged in a lab activity focused on artificial selection and designed to demonstrate how selection affects both phenotypes and genotypes. In contrast, the lab activity in the comparison group focused on phenotype alone. Data sources for the study included pre-tests of basic concepts in genetics and evolution and two post-test items requiring the students to reproduce and apply their knowledge about artificial selection. The findings indicated that the lab activity which allowed students to explore the interplay between different levels, provided vertical coherence and enhanced students' ability to explain evolutionary change in both reproduction and transfer items. In contrast, the lab activity in the comparison group failed to do so, and most students did not improve their ability to explain evolutionary change. Implications for instruction and recommendations for further research are discussed in light of these findings.

ARTICLE HISTORY
Received 3 August 2015
Accepted 1 April 2016

KEYWORDS
Levels of biological organization; vertical coherence; evolution; thinking across levels; integrated knowledge

Introduction

Because biological systems are multi-leveled and hierarchically structured, the argument has been made that scientific phenomena 'can best be understood through a perspective of levels' (Wilensky & Resnick, 1999, p. 17) because 'in biological systems, the explanations for, or mechanisms of, phenomena apparent at one scale often lie at a different scale' (Parker et al., 2012, p. 49). For example, it is necessary to refer to photosynthesis in explaining the growth of a maple tree and thus to link the level of the organism (growth) with the subcellular level (photosynthesis). Similarly, when tracing matter in earth systems, 'sense can be made of the complexity of the biosphere by viewing it as a set of interrelated systems that can range in size from the subcellular to the ecosystems

CONTACT Marcus Hammann hammann.m@uni-muenster.de Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Schlossplatz 34, 48143 Münster, Germany

level' (Wilson et al., 2006, p. 324). As a consequence, various instructional strategies have been designed that focus on fostering students' thinking in levels, on the one hand, and on improving reasoning across levels, on the other. Chief among these strategies are design-based learning and teaching strategies that encourage thinking across levels in cell biology and genetics (Knippels, 2002; Knippels, Waarlo, & Boersma, 2005; Verhoeff, Waarlo, & Boersma, 2008), which consist of the following components:

1. distinguishing different levels of organization;
2. interrelating concepts at the same level of organization (horizontal coherence);
3. interrelating concepts at different levels of organization (vertical coherence);
4. thinking back and forth between levels (also called yo–yo learning); and
5. meta-reflection about the question which levels have been transected.

Other authors have described a similar instructional strategy of 'organizing systems and identifying scale' (Parker et al., 2012), which is also called 'keeping track of scale' and is expected to help students reason across different levels in plant physiology and ecology. In the context of learning with multiple external representations, the instructional strategy of 'vertical translation across levels of representations' has also been the subject of discussion (Tsui & Treagust, 2013). Generally, educators agree that biology instruction should help students move fluidly between levels and reason across them to contribute to biological literacy (Brown & Schwartz, 2009; Duncan & Reiser, 2007; Parker et al., 2012; Wilensky & Resnick, 1999; Wilson et al., 2006). Thus, in this paper, we describe the results of a study investigating the effects of providing vertical coherence and promoting reasoning across levels of biological organization on students' ability to explain evolutionary change.

How good are biology students at thinking across levels of biological organization?

When aiming to make sense of biological phenomena, students often find it difficult to adequately address the different levels involved. There is ample evidence for the so-called micro–macro problem (e.g. Knippels, 2002), and it comes from research fields as diverse as students' systems thinking (e.g. Penner, 2000; Resnick, 1996; Wilensky & Resnick, 1999), student conceptions of cell biology (e.g. Dreyfus & Jungwirth, 1989; Flores, Tovar, & Gallegos, 2003), genetics (e.g. Duncan & Reiser, 2007; Manokore & Williams, 2012; Marbach-Ad & Stavy, 2000), physiology (e.g. Anderson, Sheldon, & Dubay, 1990; Brown & Schwartz, 2009; Songer & Mintzes, 1994; Stavy, Eisen, & Yaakobi, 1987), ecology (e.g. Ebert-May, Bazli, & Lim, 2003), and evolution (e.g. Ferrari & Chi, 1998; Shtulman, 2006). Students' problems when thinking across levels entail, among others, confusion of levels – sometimes also called 'slippage between levels' (Wilensky & Resnick, 1999, p. 3) – and disconnects between levels. The latter is sometimes also associated with fragmented and compartmentalized knowledge (Brown & Schwartz, 2009). Examples of both aspects will be provided below. Particular emphasis will be placed on students' problems regarding thinking across levels in the context of evolutionary biology.

Confusing levels of biological organization

Confusion of levels is considered an obstacle to students' understanding of emergent phenomena, where 'macro-level properties emerge as the result of micro-level interaction between system components' (Penner, 2000, p. 784). Thus, when asked to explain emergent processes, such as the aggregation process of slime-mold cells into a single multicellular organism, students frequently assume centralized control and incorrectly explain the phenomenon at the level of the individual (e.g. one slime-mold cell gives orders), rather than focusing on the emergence of new qualities at the multicellular level through simple, decentralized interactions (Wilensky & Resnick, 1999). Further examples come from the literature on student conceptions, in which students must also address multi-leveled phenomena, and they confuse one level for another. For example, students have been shown to argue that one-celled organisms (cellular level) possess structures (e.g. lungs and intestines) and functions (e.g. digestion) that are observable only at the multicellular level (Dreyfus & Jungwirth, 1989; Flores et al., 2003). When observing diffusion or osmosis, such as the swelling of a cell, students frequently argue that molecules undertake directional movement toward lower concentrations, thus projecting characteristics at the macro-level (i.e. greater numbers of molecules move from areas of higher to lower concentration, rather than vice versa) onto the micro-level, where such characteristics simply do not exist (i.e. where each molecule moves randomly) (Chi, 2005; Meir, Perry, Stal, Maruca, & Klopfer, 2005). Additionally, students confuse levels when they posit that genes and traits are identical and when they consider the traits of offspring to be inherited directly (Manokore & Williams, 2012; Marbach-Ad & Stavy, 2000). Research into problem solving in genetics has further revealed that students confuse dominance with frequency (Slack & Stewart, 1990; Smith & Good, 1984). Accordingly, students confuse the molecular level with the population level, assuming that 'a gene is dominant if the phenotype it determines is the most frequent in the population' (Smith & Good, 1984, p. 906). When asked to define cell respiration in humans, to cite an example from human physiology, students have been shown to describe gas exchange instead, thus switching to the level of the organism and thereby confusing cellular respiration with human breathing (Anderson et al., 1990; Songer & Mintzes, 1994; Stavy et al., 1987).

Disconnects between levels of biological organization

In addition to confusion of levels, disconnects between levels are considered an obstacle to an integrated understanding of complex biological phenomena, which require reasoning across different levels. In this context, students' understanding of plant physiology and their understanding of genetics are well studied examples. In plant physiology, students and pre-service teachers rarely see the ecological implications of photosynthesis and cell respiration, thus displaying a lack of systems awareness (Brown & Schwartz, 2009; Canal, 1999; Waheed & Lucas, 1992). In genetics, students frequently fail to reason across two ontologically distinct levels, 'an information level containing the genetic information, and a physical level containing hierarchically organized biophysical entities such as proteins, cells, tissues, etc.' (Duncan & Reiser, 2007, p. 938). These problems are aggravated by the invisibility and inaccessibility of genetic phenomena. Truncated explanations are symptomatic, that is, explanations that directly link the genotype to the phenotype. As a consequence, it has been argued that biology instruction needs to help students construct 'causal mechanistic explanations of how the genetic information brings about physical effects (features or traits)' and thus to establish links between levels that appear disconnected to many students (Duncan & Reiser, 2007, p. 947).

Thinking across levels of biological organization in evolution instruction

Thinking across levels has also been identified as a challenge to understanding evolutionary biology. Ferrari and Chi (1998) argued that evolutionary principles are difficult to understand because they require 'reconciling different levels of organization for such concepts as genes, individuals, populations and species' (Ferrari & Chi, 1998, p. 1234). Similarly, other researchers in biology education have emphasized the importance of intra-species genetic variation (i.e. the genetic level) in understanding evolutionary phenomena (Brumby, 1984; Deadman & Kelly, 1978; Ferrari & Chi, 1998; Halldén, 1988; Shtulman, 2006). However, it is difficult to arrive at an integrated understanding of genetics and evolution. For example, one study hypothesized that teaching genetics prior to evolution 'would help pupils to understand the role that intra-species variation plays in evolution' and would provide them with arguments against the Lamarckian view of evolutionary change (Halldén, 1988, pp. 542–543). After instruction, the students were asked 'how hereditary characteristics undergo change over time' (Halldén, 1988, p. 543). However, the students could not answer this question conclusively, which was understood as indicating that students' had problems 'relating these facts to one another in coherent descriptions and explanations' (Halldén, 1988, p. 541). In terms of fragmented knowledge, the students' inabilities to explain how hereditary characteristics undergo change over time can be traced back to disconnects between levels, that is, not relating the genetic level (i.e. intra-species genetic variation as a prerequisite for change in hereditary characteristics over time) to the level of the individual (i.e. selection leading to differential survival and differential reproduction among individuals) and of the population (i.e. differential survival and reproduction leading to changes in allele frequencies in the population).

Further evidence for difficulties concerning thinking across levels comes from studies investigating students' teleological thinking, which is widespread and involves the belief that certain traits evolved because of their function or because of some type of intentional process on the part of the individual or of the species as a whole (Kelemen, 2012). Typically, teleological thinking does not consider the genetic level (i.e. disconnect between levels) when explanations are 'basic function-based' – for example, 'giraffes have long necks so that they can reach high food' – or when they are 'basic need-based' – for example, 'giraffes got long necks because they needed them to reach high food' (Kelemen, 2012, pp. 67–68). The third type of teleological explanation – 'elaborate need-based' – is characterized by more detailed accounts of causal mechanisms responsible for evolutionary change, for example, when organisms undertake efforts to adapt in a goal-directed fashion or when 'mother nature' responds to the needs of an organism to ensure survival. Additionally, elaborate need-based explanations include the naive Lamarckian conception that organisms acquire new traits during their lifetimes and pass them on to the next generation (Kelemen, 2012, pp. 68–69). However, Lamarckian student conceptions must be distinguished from teleological student conceptions (see Kampourakis & Zogza, 2007). Whereas the former denote the 'effect of use or disuse that would produce changes on body structures' and the inheritance of acquired characteristics, the latter imply goal-directed changes (Kampourakis & Zogza, 2007, p. 393). Confusion of levels can be observed in elaborate need-based explanations, in which students argue that organisms can modify their genetic make-up to adapt better to a changing environment. Thus, students confuse levels in making evolutionary explanations, when they refer to goal-directed behavior (at the genetic level), which is intended to ensure survival, rather than to changes in allele frequencies through variation and selection.

As another prominent example of slippage of levels, students have reportedly conflated the level of the individual and the level of the population, seeing traits as gradually changing in all members of a population ('transformational view'), rather than focusing on how new traits come into being and how their frequencies change in the population ('variational view') (Shtulman, 2006; see also Bishop & Anderson, 1990). Thus, the argument has been made, that whereas evolutionary biologists 'would explain this change in terms of two processes (mutation and selection) operating on a population of individuals, transformationists would explain this change in terms of a single process operating on the species' "essence"' (Shtulman, 2006, pp. 172–173). Essentially, students frequently explain evolution without reference to selection, as the following quote reveals: 'They (cheetahs) might have had to run fast to escape predators and gradually their muscles and bones changed to adapt to this' (Bishop & Anderson, 1990, p. 423).

Effecting conceptual change in evolutionary biology is difficult, with some studies showing post-test gains in evolutionary thinking (e.g. Kampourakis & Zogza, 2009; Wallin, 2008), whereas others describe students who revert to nonscientific ideas sometime after instruction (e.g. Banet & Ayuso, 2003). Some students seem to resist conceptual change, which is considered intentional and deeply influenced by learner characteristics, such as achievement of goals, epistemic motivation and beliefs, interest, self-efficacy, affect and emotions (Sinatra & Mason, 2008). Evolution education, in fact, has proved to be a fruitful arena for studying – and attempting to unravel – the complexity and diversity of students' conceptual ecologies and the restructuring that occurs among the various components involved. Thus, research has revealed that students' conceptual ecologies for evolution include thinking dispositions, epistemological beliefs, scientific and religious orientations, and acceptance of evolutionary theory, among other elements (Demastes, Good, & Peebles, 1995; Deniz, Donnelly, & Yilmaz, 2008). Acceptance of evolutionary theory has been studied closely relative to students' understanding of evolution as a factor that might hinder true conceptual change. Research has focused on high school students (e.g. Cavallo & McCall, 2008), college students (e.g. Sinatra, Southerland, McConaughy, & Demastes, 2003), and high school teachers (e.g. Rutledge & Warden, 2000), but the findings have been contradictory (e.g. Kampourakis & Zogza, 2009). In particular, the relationship between acceptance and understanding of evolutionary theory seems to depend on the specific sample and the manner in which acceptance is defined and measured (Konnemann, Asshoff, & Hammann, 2012). As a consequence, researchers focusing on students' understanding of evolution have sometimes deliberately chosen not to investigate this important aspect in favor of in-depth analyses of students' explanatory frameworks (e.g. Kampourakis & Zogza, 2009). Similarly, this study investigates how a particular learning and teaching strategy – providing vertical coherence and thinking across levels – affects students' explanations of evolutionary changes. The argument has been made in support of strengthening the links between knowledge of genetics and evolution (e.g. Banet & Ayuso, 2003; Halldén, 1988; Kampourakis & Zogza, 2009). Therefore, we sought empirical evidence for the importance of interrelating concepts at different levels of biological organization – that is, genotype and phenotype – when so doing.

Promoting thinking across levels of biological organization when teaching selection

Selection is difficult to understand for many students, but it is an important aspect of evolutionary theory (Bishop & Anderson, 1990). To enhance students' understanding of selection, different hands-on lab activities have been developed over the last 40 years. We surveyed selected lab activities that focus on selection in terms of their potential for promoting reasoning across levels. The following two criteria of analysis were used: whether or not the lab activity (a) addresses different levels of organization (i.e. phenotype and genotype) and (b) helps to distinguish between them (e.g. by representing phenotype and genotype as separate conceptual entities). These criteria are interrelated, but they are not identical. In fact, they allowed us to distinguish between type-one lab activities regarding selection, which focus on phenotypic change alone (Lauer, 2000; Maret & Rissing, 1998; Scheersoi & Kullmann, 2007; Stebbins & Allen, 1975), and type-two lab activities, which demonstrate how selection affects both genotypes and phenotypes (Allen & Wold, 2009; Christensen-Dalsgaard & Kanneworff, 2009; Fifield & Fall, 1992; Frey, Lively, & Brodie, 2010).

Type-one lab activities that focus on how selection affects phenotypes

Type-one lab activities involve differently adapted organisms, which are typically represented by colored chips spread out over a piece of fabric, which is intended to represent the habitat (e.g. Maret & Rissing, 1998; Stebbins & Allen, 1975). Additionally, the use of colored beads (Scheersoi & Kullmann, 2007) and differently flavored jelly-beans has been proposed (Lauer, 2000). Students act as predators and select organisms over several generations. There is a range of variations (see, e.g. Stebbins & Allen, 1975), but lab activities classified in this group always remain at the level of the phenotype. Even in the 'mutation' variation, new colored dots are introduced into the population at some stage in the activity without representing the phenomenon at the genetic level (Stebbins & Allen, 1975). In fact one author explicitly described the aim of this type of lab activity as demonstrating how 'selection occurs in phenotypes' (Lauer, 2000, p. 42). By engaging in a type-one lab activity, students can see that some organisms survive and reproduce, whereas others go extinct. Thus, the differences in the organisms' physical properties are shown relative to the likelihood of survival. As a nonrandom factor, selection leads to phenotypic changes, and changes in allele frequencies are not shown.

Type-two lab activities focus on changes in both phenotypes and allele frequencies

Type-two lab activities strive for a fuller understanding of natural selection by demonstrating how natural selection affects both genotypes and phenotypes (Allen & Wold, 2009; Fifield & Fall, 1992). To achieve this goal, type-two lab activities typically represent genotypes and phenotypes as different conceptual entities, showing that natural selection results in changes of both phenotypes and allele frequencies over time. For example, one lab activity consisted of differently colored paper chips with genotypes printed on them, and the activity is expected to help students 'integrate their understanding of alleles, genotypes and phenotypes' (Fifield & Fall, 1992, p. 231). In another lab activity, Lego® bricks are used to construct organisms, with each segment coded by a gene with five different alleles represented by the color of the brick (Christensen-Dalsgaard & Kanneworff, 2009). Thus, type-two lab activities demonstrate to students that changes in phenotypes result from changes in the frequencies of genes and alleles in the gene pool.

Lab activities that confuse levels of biological organization

Our survey of lab activities also revealed the failure to distinguish between phenotypes and genotypes. In one activity focusing on size-selective harvesting, a population of Atlantic cod was represented by a package of dry bean soup, consisting of five varieties of beans (Allen & Wold, 2009). Each variety, problematically, is intended to represent both a particular body size (i.e. when the variety of beans is passed through a sieve to demonstrate the effects of artificial selection on body size) and a particular gene (i.e. when beans are linked to genes in a one-to-one relationship to demonstrate changes in gene frequencies over time). Thus, beans represent both organisms and genes in this lab activity. Although intended to demonstrate the effects of artificial selection on both phenotype and genotype, the lab activity can reinforce the student conception that traits – rather than genes – are passed from one generation to the next (Lewis & Kattmann, 2004; Marbach-Ad & Stavy, 2000). Although confusing levels in a lab activity is problematic in and of itself, linking *one* gene to *one* body size is incorrect as well because body size is a quantitative trait with multiple genes responsible for the phenomenon (polygenic inheritance).

Aims of the study and research questions

Thus far, the impact of type-one and type-two lab activities on students' ability to explain evolutionary change has not been studied systematically. This study investigated the research question of whether type-two lab activities on the topic of selection support students' ability to explain evolutionary change more effectively than type-one lab activities. As noted above, type-two lab activities differ from type-one lab activities insofar as they address different levels of organization (i.e. phenotype and genotype), interrelate them, and help to distinguish between them. Thus, type-two lab activities on selection can be hypothesized to promote thinking across levels more effectively than type-one lab activities in terms of helping students interrelate concepts at different levels of biological organization and to avoid confusing them. To test these hypotheses, we conducted a pre–post-test design study with a comparison group.

Methods

Participants and setting

A total of 197 students (78 female and 119 male; mean age 18 years old, minimum 16 years old, maximum 20 years old) from nine high schools located in five small cities (40,000–70,000 inhabitants) and two large cities (300,000–500,000 inhabitants) in Germany participated in the study. Although data regarding the socioeconomic status of these particular students – as well as data about differences between schools – were not obtained, the students' socioeconomic status could reasonably be estimated to be higher than the average for Germany because the students who participated in this study attended the so-called gymnasium, a type of high school in which students' socioeconomic status ranks higher, on average, than that of students who attend other types of high schools in Germany (Nold, 2010; Prenzel et al., 2007). Additionally, the data collection did not include information about the ethnicity of the students participating in this study. However, because North Rhine-Westphalia is one of the most ethnically diverse states in Germany, where approximately one-quarter of the population is characterized by an immigrant background, the participating students very likely reflected this ethnic diversity.

At the time of the study, all the students were in the second term of Grade 13 and being trained in evolutionary biology, a standard component of the biology curriculum at this stage. All the teachers followed the similar mandatory genetics and evolution curriculum, which was issued by the Ministry of School and Further Education of North Rhine-Westphalia.

Design of the study and data collection procedures

The study had a pre–post-test design with a comparison group. We considered using an additional follow-up test to analyze the extent to which the students were able to maintain their gains on the tests. However, due to the national curricula, evolution is taught right before the final examinations (*Abiturprüfungen*), after which the students leave high school. Therefore, follow-up tests were difficult to implement, and we decided against them.

Nine groups of students (Grade 13) participated in the study, which was undertaken during regular biology lessons, and it was conducted by nine biology teachers who had volunteered to participate. Within each group, the teachers randomly assigned students to the experimental group ($n = 103$; 54 female and 49 male) and to the comparison group ($n = 94$; 65 female and 29 male). The teachers were recommended to use random allocation cards with the numbers 1 and 2 for randomization. The students drew the cards from a small box but were not told which number signified which group. The biology teachers also administered the standardized questionnaire, pre-test, lab activity, and post-test. They had previously been informed by the first author of this paper about the aims and procedures of the study, as part of a one-day teacher training focusing on innovative approaches to teaching evolutionary biology at the high school level. One 90-minute class session was used to administer the questionnaire, pre-test and lab activity. The post-test followed in the next regular biology lesson, that is, one day to five days after the lab activity, depending on when the class was scheduled.

Students in the experimental and comparison groups took the same tests but engaged in different lab activities in groups of no more than five students. Because both lab activities focused on the impact of selection on evolutionary changes, the study was conducted within the first half of the unit on evolution, with slight variations among the nine teachers concerning the specific positioning in the unit. Additionally, the lab activity did not require any teaching by the teachers because all the necessary information was provided to the students in writing.

Data sources

This study had three main data sources:

1. In the questionnaire, the students were asked about their age and gender and whether or not they had been taught genetics, artificial selection and evolution prior to Grade 13. The student responses were double checked by asking the teachers whether they had taught genetics in the semester before the unit on evolution.
2. In the pre-test, the students' prior knowledge of genetics was assessed to ensure that randomization of the students into two groups (i.e. the experimental group and comparison group) was successful in controlling for this factor. In particular, the students were asked to define the terms 'gene' and 'allele' in separate open-response questions because prior knowledge is an important factor of information processing and learning (Bransford, Brown, & Cocking, 2000).
3. The main focus of this study was on the pre–post-test changes in the students' ability to explain evolutionary changes. Research has shown that there are significant item feature effects that must be considered assessing student performance in evolutionary biology (Kampourakis & Zogza, 2009; Nehm & Ha, 2011; Nehm & Reilly, 2007; Nehm & Schonfeld, 2008). To control for item feature effects, we used three structurally and conceptually equivalent items (the pre-test item, post-test reproduction item, and post-test transfer item) based on the cheetah-probe (Bishop & Anderson, 1990). Like the cheetah-probe – the standard in the field – the test items used in this study called for an explanation of evolutionary change. Unlike the cheetah-probe, selection is explicitly mentioned in all three items (e.g. trophy hunting and size-selective harvesting) because this particular aspect is important for analyzing whether or not – and to what extent – students in both groups differ in terms of describing the impact of selection on both phenotypes and genotypes.

In the pre-test, one test item consisting of two questions was administered to assess the ability to explain evolutionary changes. In the stimulus material, the students were informed about the Asian elephants' decrease in average tusk size as a consequence of trophy hunting. The students were asked to explain the phenomenon: 'Why did tusk size diminish as a result of trophy hunting?' Then, the students were informed that, after ending trophy hunting and poaching, average tusk size had increased slowly or not all, depending on the elephant population studied. The students were asked to explain this phenomenon as well by investigating and answering the following question: 'After ending trophy hunting and poaching, why did tusks become longer very slowly or did remain small in the elephant populations?'

In the post-test, students in both groups responded to two test items consisting of two questions each, assessing the ability to explain evolutionary changes in a familiar and an unfamiliar context. As in the pre-test item on trophy hunting, both items presented the evolutionary phenomenon in a stimulus text first. In one item, the students were informed about the decrease in body size in Atlantic cod. The other item addressed the subject of changes in the fur color of mice. Then, the students were asked to explain the phenomena. As with the pre-test item, the students were informed that, after ending selection, reversal of the phenomenon required longer than expected or did not occur at all. The students were asked to explain this phenomenon as well. Because the lab activities of students in the two groups differed (see description below), the test item on Atlantic cod required knowledge reproduction for students in the experimental group and transfer of knowledge for the students in the comparison group. For the test item on changes in the fur color of mice, the situation was reversed.

Data coding and data analysis

To code the participants' definitions of the terms 'gene' and 'allele', separate coding guides were developed. According to Duncan and Reiser (2007), students tend to define genes on two ontologically distinct levels, that is, the informational level (e.g. genes contain information) and the physical level (e.g. genes determine features). In addition, when asked to define genes, students frequently provide general definitions (e.g. genetic material), fail to distinguish between genes and traits (e.g. genes are features) and rarely mention gene products (Lewis & Kattmann, 2004; Lewis, Leach, & Wood-Robinson, 2000; Marbach-Ad & Stavy, 2000). Based on these findings, the students' responses to the open question 'What is a gene?' were coded using four categories (see Table 1). The responses were coded by two

Table 1. Types of students' explanations to the questions 'What is a gene?' and 'What is an allele?' (pre-test).

Type of explanation	Description of the category	Sample responses
What is a gene?		
Informational level	Genes contain information.	A gene codes for a special protein (student 24-5); Genes contain information for the formation of traits of human beings (student 22-18)
Physical level	Genes determine features.	Genes determine our appearance (student 22-2); A gene is a part of the DNA that causes a special trait (student 23-12)
Gene = trait	Genes are traits.	An invisible trait (student 23-3); Carries traits (student 24-12)
Other	Students use alternative conceptions or more general terms, like hereditary material or genetic make-up.	Carries the genetic make-up (student 24-13); DNA-sequence (student 23-20)
What is an allele?		
Forms of a gene	Alleles are versions/forms of a gene.	Alleles represent different versions of a gene. (student 23-4); Is one of several versions of a gene (student 26-19)
Allele = trait	Alleles are traits.	An allele contains a specific trait (student 22-7); An invisible trait (student 23-3)
Other	Students use alternative conceptions or more general terms, like hereditary material, DNA.	Carrier of DNA (student 22-21); An allele is one of the four arms of a chromosome (student 31-8)

persons, and the inter-rater reliability was found to be very high (Cohens' $k = .87$; $p < .001$). In addition, the students' responses to 'What is a gene?' were analyzed for whether or not specific gene products were mentioned. For this coding, the inter-rater reliability was found to be very high (Cohens' $k = .95$; $p < .001$). The inter-rater reliability was Cohens' $k = .83$ ($p < .001$) for 'What is an allele?'.

To code the pre-test item (Asian Elephants) and the post-test items (Atlantic cod; Mice) that assessed the students' ability to reason across levels when explaining evolutionary changes, we developed categories applicable to all three test items (see Table 2). The coding categories were developed both deductively and inductively (Mayring, 2000). To assess the ability to connect different levels of biological organization, we anticipated two major types of student responses: those explaining evolutionary change at the level of the phenotype alone (*type-one explanations*), and responses explaining evolutionary change at both the level of the phenotype and the level of the genotype (*type-two explanations*). After coding a number of student responses, we were able to further distinguish

Table 2. Types of students' explanations of evolutionary phenomena (pre-test and post-test items, sample responses to question 1 // sample response to question 2).

Type of explanation	Description of the category	Sample responses
Type-one explanation	Students explain evolutionary change exclusively at the level of the phenotype. No use of genetic terms or concepts.	Because male elephants with long tusks are killed very often, there are only bulls with smaller tusks. The females mate with these bulls, as there are fewer and fewer bulls with long tusks. That's why the male offspring also have smaller tusks. Thus, tusk size becomes smaller because of hunting. // Possibly, few or no elephants with long tusks survived, so that average tusk size increased only slowly or not at all in the following generations. Also, many hunters ignored the hunting ban, and continued killing elephants with long tusks. (student 22-1, pre-test, Asian Elephant)
Type-two explanation (unspecific)	Evolutionary change is explained mostly at the level of the phenotype. Students use genetic terms and concepts, but do not explain changes in the gene pool in terms of changes in frequencies of specific genes or alleles.	Because elephants with long tusks are hunting trophies, they are shot more frequently than elephants with short tusks. As a consequence, bulls with shorter tusks make-up for a greater part of the gene pool and therefore bulls with shorter tusks predominate when the number of elephants with longer tusks decreases. // Possibly bulls with longer tusks were decimated so strongly that they had no chance to make a substantial contribution to the gene pool any more. Maybe their death rate remained higher than their birth rate even after the ban on hunting. (student 27-35, pre-test, Asian Elephant)
Type-two explanation (specific)	Evolutionary change is explained at the level of the phenotype and the genotype. Students explain phenotypic change in terms of changes in frequencies of specific genes/alleles or loss of genes/alleles.	Cod has several genes which contribute to body size so that it is possible to speak of additive genetic variance. By catching the big fish, alleles are removed which are crucial for body size. Therefore, the next generations of fish are smaller. // It is the alleles, which are responsible for body size, which allow for small fish only. In the population, there are no alleles which can be found in big fish so that these genes cannot be passed on. (student 26-7, post-test, Atlantic cod)

two types of answers among the type-two responses: responses mentioning changes in the gene pool in a rather unspecific way (*unspecific type-two explanations*) and responses making reference to more specific genetic details, such as different genes/alleles and changes in gene frequencies/allele frequencies in the gene pool (*specific type-two explanations*). As each of the three items used in the pre-test and post-test consisted of two questions, responses to both questions were coded into one variable based on to the observation that students provided phenotypical explanations for the first question in many cases, but it was often only after the second question that they considered the genetic level for the explanation of evolutionary change (which was true for approximately one-quarter of the students in the pre-test). Therefore, the students were classified as providing a type-two response, regardless of whether the responses to the first question or to the second question or to both questions contained references to genes/alleles or changes in the gene pool.

The first and last authors developed the coding manual, which was tested using 183 responses. All further responses were double coded. The inter-rater reliability for the first and last authors of this paper was found to be very high (Cohens' $k = .97$; $p < .001$). Two independent coders not involved in developing the manual were trained, and their inter-rater reliability was also found to be very high (Cohens' $k = .90$; $p < .001$).

As described in the introduction to this paper, confusion of levels is a pervasive phenomenon. Thus, in a second coding of student responses to the three test items focusing on evolutionary changes, we analyzed whether the students confused levels by writing that traits – rather than genes – are passed from one generation to the next. The inter-rater reliability between the first and the last authors was found to be very high for confusion of levels (Cohens' $k = .97$; $p < .001$). Two independent coders not involved in the development of the manual were trained to test the inter-rater reliability (Cohens' $k = .72$; $p < .001$). For detailed information about the coding rubrics, see Table 3.

Description of the lab activities

Two different lab activities were developed. 'Why are Atlantic Cod shrinking?' is a type-two lab activity based on an authentic phenomenon, that is, the shrinking of Atlantic cod (*Gadus morhua*), and it was used in the experimental group. This activity illustrates the effects of size-selective harvesting, a pervasive phenomenon shown to have ecological and evolutionary consequences (Conover & Munch, 2002; Fenberg & Roy, 2008). Before the lab activity, the students were informed about the phenomenon and where thus provided with the opportunity to encounter it at the macro-level first (see, e.g. Knippels, 2002; Tsui & Treagust, 2010; Verhoeff et al., 2008). The lab activity involved a population of 10 fish, presented to the students as drawings. (For a full description of the lab activity and its materials, see www.evolution-of-life.com). To promote thinking across levels, phenotypes and genotypes of Atlantic cod were represented as different entities, and the students were encouraged to move among the levels of the population, the organism and the genes. For purposes of simplifying a complex genetic phenomenon (i.e. polygenetic inheritance), three genes were shown to contribute to body size, with two gene loci for each gene represented as circles in the outline drawings of the fish. For each gene, there were four alleles represented by differently colored chips, with color symbolizing the degree to which each allele contributes to the trait. In nature, the body size of

Table 3. Student conceptions of inheritance in students' explanations of evolutionary phenomena (pre-test and post-test items, sample responses to question 1 // sample response to question 2).

Type of explanation	Description of the category	Sample responses
No reference to inheritance	Students do not mention inheritance when explaining evolutionary change.	Because of trophy hunting, bulls with long tusks die. Bulls with smaller tusks can reproduce more often because they have a better chance of finding a mate. // Development of tusks took a long, long time. Observing this phenomenon over a few years only can provide only few definite answers. Maybe further observation over the next 1000 years can provide the expected answers. (student 26-12, pre-test, Asian Elephant)
Inheritance of traits	Students explicitly mention inheritance of traits when explaining evolutionary change.	The tusk length of elephants has decreased, because elephants with longer tusks were hunted at that time and virtually died out. Elephants with shorter tusks survived and reproduced. // Because elephants with long tusks died out, there were only few or none with long tusks left which inherited this trait. (student 22-21, pre-test, Asian Elephant)
Inheritance of traits and genes	Students explicitly mention both inheritance of traits and inheritance of genes when explaining evolutionary change.	Because elephants with short tusks survived, they the next generation inherited short tusks from them. Elephants with long tusks were killed. // The trait for long tusks has almost died out, therefore, only a few carriers of the traits are left, which possess this gene and are able to pass it on. (student 23-5, pre-test, Asian Elephant)
Inheritance of genes/alleles (implicit)	Students describe genetic changes over time (without indicating what exactly is passed on from generation to generation) and thus implicitly mention inheritance of genes or alleles when explaining evolutionary change.	Elephant bulls with long tusks were hunted and killed. Therefore, the percentage of bulls with smaller tusks who mated with females increased. Thus, more elephants with small tusks were born. // Because female elephants primarily mated with bulls with smaller tusks, the number of alleles, which are responsible for long tusks, decreased over time. (student 22-20, pre-test, Asian Elephant)
Inheritance of genes/alleles (explicit)	Students explicitly mention inheritance of genes or alleles when explaining evolutionary change.	Several genes are responsible for body size. Because large fish are caught, alleles for large body size become less frequent. Alleles for small fish are passed on. That's why over many generations alleles for big Atlantic cod have become less frequent in the population. // There are only few alleles left in the population which can contribute to large body size. Further, alleles coding for small fish have become more frequent. (student 26-12, post-test, Atlantic cod)

cod is determined by a far greater – although presently unknown – number of genes and alleles. In this lab activity, however, the number of genes and alleles was reduced to make the lab activity manageable for classroom use. The lab activity does not attempt to illustrate how the genetic information results in the trait.

At the beginning of the lab activity, the gene pool consisted of 60 alleles, with equal numbers for each of the four types of alleles. The alleles were randomly drawn from the gene pool and allotted to the 10 fish. Then, body size was assessed for each fish. Switching to the level of the population, allele frequencies and body sizes were registered for the 10 fish with the help of a table and a diagram. Size-selective harvesting, the next stage in the lab activity, was simulated by identifying the five largest fish and removing their alleles from the gene pool. The remaining fish 'reproduced', and their alleles doubled in number. The resulting 60 alleles represented the gene pool of the next generation, which was randomly allotted to the population of 10 fish. After a few rounds of selection, changes in body size and allele frequencies became clearly visible. Guided by the structure of the lab activity, the students could thus explore the relationships between the levels of biological organization themselves. Therefore, 'Why are Atlantic Cod shrinking?' is classified as a type-two lab activity by focusing on changes in both phenotypes and allele frequencies as the result of artificial selection. The lab activity, however, did not include meta-reflection regarding the question as to which levels were transected. Additionally, the lab activity did not explicitly address the issue of inheritance of genes vs. inheritance of traits, which was also true for 'Mice in the city park', which is described in the next paragraph.

'Mice in the city park' is a type-one lab activity developed exclusively for the purpose of comparing its effects to 'Why are Atlantic Cod shrinking?' 'Mice in the city park' is a variation of the often-used lab activity on selection involving differently colored chips (i.e. organisms) spread out over a background (i.e. habitat), which makes it difficult to see some chips but not others (e.g. Stebbins & Allen, 1975). In contrast to existing type-one lab activities, 10 different shades of gray are used for organisms in 'Mice in the city park'. This change was made to represent a quantitative feature inherited polygenetically, such as body size in cod. The beginning population consisted of 60 mice, with each shade of gray equally represented in the population. The students acted as predators – cats in the city park – and selected 30 mice, which were removed from the population. After sorting the remaining mice according to color, their number was doubled (i.e. reproduction). The next generation of mice was spread out over the city park, and selection began again. The students kept track of the changes in the frequencies of differently colored mice with the help of a diagram.

In summary, 'Mice in the city park' demonstrates how selection occurs in phenotypes, which stands in marked contrast to 'Why are Atlantic Cod shrinking?', in which the students can see that selection impacts both phenotypes and genotypes. There are further differences between the two lab activities, including differences between organisms (mice vs. cod), traits (fur color vs. body size), and in authenticity (fictitious vs. real-life). To mitigate these differences, we originally considered comparing two variants of 'Why are Atlantic Cod shrinking?', that is, the variant described above and a modified variant of 'Why are Atlantic Cod shrinking', illustrating evolutionary change at the level of phenotype alone. This idea proved impossible, however, primarily because of the need to ensure that selection of the largest fish could occur over several generations. For example, if the lab activity involved a fishing net for size-selective harvesting, the population of fish had to be very large such that the first round of selection did not result in elimination of all the largest individuals, thus trivializing the demonstration of the effects of selection. In large populations, such as the Atlantic Ocean, selection is possible over decades because there are always parts of the population not affected by selection, but we were unable to transform this feature into a type-one hands-on lab activity. Further, we wanted to avoid conflating organisms and genes, as was the case in Allen and Wold (2009). Additionally, we aimed at developing a type-one lab activity focusing on a quantitative trait, which we did not find in the survey of lab activities described above. Thus, we developed two lab activities that differed not only in the intended factor but also in other aspects, thus limiting the interpretability of the findings, which will be considered in the discussion.

Pre-piloting of lab activities

Prior to this study, both lab activities were pre-piloted in two independent samples of 29 high school students each (age 14–15 years old). The main aim was to determine whether the lab activities were suitable for classroom use, as well as practicable and interesting to students. Pre-piloting of the lab activities involved teaching specific genetics knowledge, which students at this age level typically do not have. The pre-pilot had two main findings. No changes were deemed necessary for the lab activities, which proved interesting and easy for the students to perform. Additionally, the decision was undertaken to perform the main study with students aged 17–18 years old (Grade 13) rather than with students aged 14–15 years old (Grade 9) to integrate the lab activities into the standard teaching unit on evolution and to benefit from the curriculum for Grade 12, which covered genetics.

Description of the sample

A total of 197 students (Grade 13, age 18 years old) participated in the study. Despite randomization, the gender ratios differed significantly between the groups with male students overrepresented ($n = 49$; expected: $n = 41$) and female students underrepresented ($n = 54$; expected: $n = 62$) in the experimental group. The opposite was true for the comparison group (males: $n = 29$, expected: $n = 37$; females: $n = 65$, expected: $n = 57$). The majority of the students in both groups reported that, prior to Grade 13, they had been taught evolution (experimental group: 70.6%; comparison group: 70.2%), and natural selection (experimental group: 82.4%; comparison group: 75%). Additionally, approximately one-third of the students in both groups reported that they had been taught artificial selection (experimental group: 39.6%; comparison group: 32.3%) before Grade 13. All the students but one reported that they had been taught genetics prior to the study. Chi-square tests revealed no significant differences between the experimental group and the comparison group concerning prior teaching of evolution, natural selection, and artificial selection. The data were analyzed for potential gender differences within the experimental and comparison groups, but no significant differences were found.

Pre-test results

Participants' ability to define basic genetic concepts

Findings for the students' definitions of genes are listed in Table 4. Most of the students in both groups defined genes at the informational level and the physical level, although some students made general responses and sometimes confused genes and traits. The chi-square test revealed no significant differences between the students in the experimental and comparison groups for the four categories of gene definitions ($\chi^2(3) = 2.59$; n.s.). The same outcome was true for the students' mentioning of specific gene products in their definitions of genes. Thirteen students (13.5%) in the experimental group and 15 students (16%) in the comparison group mentioned specific gene products ($\chi^2(1) = .22$; n.s.).

Table 4. Types of student responses to the questions: 'What is a gene?' and 'What is an allele?' (pre-test).

	What is a gene?					What is an allele?			
	Informational level	Physical level	Gene = trait	Other	Total	Forms of a gene	Allele = trait	Other	Total
Experimental group									
Count	52	17	5	22	96	61	14	21	96
Expected count	55.1	14.1	6.6	20.2	96.0	54.9	19.7	21.3	96.0
% within group	54.2	17.7	5.2	22.9	100.0	63.5	14.6	21.9	100.0
Std. residual	−0.4	0.8	−0.6	0.4		0.8	−1.3	−0.1	
Comparison group									
Count	57	11	8	18	94	42	23	19	84
Expected count	53.9	13.9	6.4	19.8	94.0	48.1	17.3	18.7	84.0
% within group	60.6	11.7	8.5	19.1	100.0	50.0	27.4	22.6	100.0
Std. residual	0.4	−0.8	0.6	−0.4		−0.9	1.4	0.1	
χ^2(df); p	$\chi^2(3) = 2.587$; $p = .460$					$\chi^2(2) = 5.016$; $p = .081$			

Further, students confused alleles and traits far more often than genes and traits. For example, one student who defined genes as hereditary features without confusing genes and traits, provided the following definition: 'Alleles are traits, for which – according to Mendelian rules – there exist two states' (student 27-31). In addition, there were more students in the comparison group (23 students) than in the experimental group (14 students) who confused alleles and traits. However, the differences between the two groups of students for the three categories of allele definitions listed in Table 4 were not statistically significant ($\chi^2(2) = 5.016$; n.s.).

Participants' explanations of evolutionary changes: disconnects between levels

Pre-test findings concerning the students' answers to the test item focusing on Asian elephants' decreasing tusk size are listed in Table 5 (left section). These findings revealed disconnects between different levels of biological organization. Thirty-one students (30.4%) in the experimental group and 30 students (34.5%) in the comparison group explained evolutionary change exclusively at the level of the phenotype (type-one explanation; see Table 2). A typical example was the following response: 'Because of trophy hunting, bulls with long tusks die. Bulls with smaller tusks can reproduce more often because they have a better chance of finding a mate' (student 26-12). Other students described connections between phenotypic and genetic changes but in a rather unspecific fashion. This finding was true for 22 students (21.6%) in the experimental group and 17 students (19.5%) in the comparison group. A typical example of a type-two explanation (unspecific) was the following response: 'By killing most of the elephant bulls with long tusks, the gene pool changed. So, bulls with short tusks are more frequent' (student 31-9). Forty-nine students (48%) in the experimental group and 40 students (46%) in the comparison group explained evolutionary change at the level of both genotype and phenotype (specific type-two explanations). A typical student answered in the following manner: 'Because of hunting, only the elephants with small tusks survived. Since they have alleles for small tusks only, only this allele is passed on. Elephants with long tusks cannot pass their allele on' (student 23-7). The chi-square test revealed no significant differences between

Table 5. Students' types of explanations of evolutionary change (pre-test and post-test findings).

	Pre-test: types of explanations				Post-test: types of explanations (reproduction item)				Post-test: types of explanations (transfer item)			
	Type one	Type two (unspecific)	Type two (specific)	Total	Type one	Type two (unspecific)	Type two (specific)	Total	Type one	Type two (unspecific)	Type two (specific)	Total
Experimental group												
Count	31	22	49	102	7	5	91	103	19	5	76	100
Expected count	32.9	20.5	48.6	102	24.3	12.7	66.0	103	31.9	10.1	58.0	100
% within group	30.4	21.6	48.0	100	6.8	4.9	88.3	100	19.0	5.0	76.0	100
Std. residual	−0.3	0.3	0.1		−3.5	−2.2	3.1		−2.2	−1.6	2.4	
Comparison group												
Count	30	17	40	87	39	19	34	92	41	14	33	88
Expected count	28.1	18.0	41.0	87	21.7	11.3	59.0	92	28.1	8.9	51.0	88
% within group	34.5	19.5	46.0	100	42.4	20.7	37.0	100	46.6	15.9	37.5	100
Std. residual	0.4	−0.2	−0.2		3.7	2.3	−3.3		2.4	1.7	−2.5	
χ^2(df); p	$\chi^2(2) = .379$; $p = .827$				$\chi^2(2) = 55.977$; $p < .001$				$\chi^2(2) = 28.644$; $p < .001$			
Cramérs V; p	$V = .045$; $p = .827$				$V = .536$; $p < .001$				$V = .390$; $p < .001$			

Type-one explanations are phenotypic; type-two explanations take the genetic level into account in specific vs. unspecific ways (for sample responses, see Table 2).

the two groups concerning the three types of explanations ($\chi^2(2) = .379$; n.s.). The data were analyzed for potential gender differences within the experimental and comparison groups, but no significant differences were found.

Participants' explanations of evolutionary changes: confusion of levels

Table 6 lists pre-test findings for students' references to inheritance of traits vs. inheritance of genes when explaining evolutionary changes. Twenty-seven students (26.5%) in the experimental group and 20 students (23%) in the comparison group explicitly distinguished between genes and traits and always wrote that alleles or genes were passed from one generation to the next. Implicit mentioning of genetic inheritance was found for 13 students (12.7%) in the experimental group and 17 students (19.5%) in the comparison group. These students explained evolutionary changes by referring to changes in gene frequencies or allele frequencies but without explicitly mentioning that genes or alleles are passed from one generation to the next. Eighteen students (17.6%) in the experimental group and 16 students (18.4%) in the comparison group confused levels and wrote that traits – rather than genes – are passed from one generation to the next. A typical example of confusing levels of organization is the following response: 'Because elephants with long tusks were hunted, only elephants with shorter tusks reproduced. As a consequence smaller tusks are inherited to the next generation' (student 26-11). Confusion of levels was also true for nine students (8.8%) in the experimental group and seven students (8%) in the comparison group, who mentioned both inheritance of genes and traits. Approximately one-third of the students in both groups made no reference to inheritance. This outcome should not be confused with students explaining evolutionary changes at the level of the phenotype alone, that is, students who gave type-one explanations (see Tables 2 and 5). Rather, students who provided type-one explanations were found either to refer to traits being passed from one generation to the next or to make no reference to inheritance at all. Students who provided unspecific type-two explanations were found in all five categories listed in Table 6. Additionally, students, who provided specific type-two explanations, were found in all categories as well, with the exception of 'no reference to inheritance'. The chi-square test revealed no significant differences between the two groups in the pre-test concerning the five categories listed in Table 6 ($\chi^2(4) = 1.797$; n.s.). The data were analyzed for potential gender differences within the experimental and comparison groups, but no significant differences were found.

Post-test results

Participants' explanations of evolutionary changes: disconnects between levels

Opposing trends were observed for all three types of explanations, when comparing the pre- and post-test findings for the experimental and comparison groups (see Table 5). In the experimental group, there was a post-test decrease in the number of students who provided type-one explanations, a post-test decrease in the number of students who provided unspecific type-two explanations, and a post-test increase in the number of students who provided specific type-two explanations. In the comparison group, there was a post-test increase in the number of students offering type-one explanations

Table 6. Number of students who refer to inheritance of traits vs. inheritance of genes when explaining evolutionary change (pre-test and post-test findings).

	Pre-test item explanations						Post-test item explanations (reproduction item)						Post-test item explanation (transfer item)					
	Number of students who refer to inheritance of						Number of students who refer to inheritance of						Number of students who refer to inheritance of					
	No reference to inheritance	Traits	Genes and traits	Genes/ alleles (implicit)	Genes/ alleles (explicit)	Total	No reference to inheritance	Traits	Genes and traits	Genes/ alleles (implicit)	Genes/ alleles (explicit)	Total	No reference to inheritance	Traits	Genes and traits	Genes/ alleles (implicit)	Genes/ alleles (explicit)	Total
Experimental group																		
Count	35	18	9	13	27	102	11	1	4	40	47	103	19	9	6	34	32	100
Expected count	33.5	18.3	8.6	16.2	25.4	102	28	7.9	5.3	31.7	30.1	103	32.4	9.6	5.3	28.7	23.9	100
% within group	34.3	17.6	8.8	12.7	26.5	100	10.7	1.0	3.9	38.8	45.6	100	19	9	6	34	32	100
Std. residual	0.3	−0.1	0.1	−0.8	0.3		−3.2	−2.5	−0.6	1.5	3.1		−2.4	−0.2	0.3	1.0	1.6	
Comparison group																		
Count	27	16	7	17	20	87	42	14	6	20	10	92	42	9	4	20	13	88
Expected count	28.5	15.7	7.4	13.8	21.6	87	25	7.1	4.7	28.3	26.9	92	28.6	8.4	4.7	25.3	21.1	88
% within group	31	18.4	8.0	19.5	23	100	45.7	15.2	6.5	21.7	10.9	100	47.7	10.2	4.5	22.7	14.8	100
Std. residual	−0.3	0.1	−0.1	0.9	−0.4		3.4	2.6	0.6	−1.6	−3.3		2.5	0.2	−0.3	−1.0	−1.8	
χ^2(df); *p*	$\chi^2(4) = 1.797$; $p = .773$						$\chi^2(4) = 60.054$; $p < .001$						$\chi^2(4) = 20.040$; $p < .001$					
Cramérs *V*; *p*	$V = .097$; $p = .773$						$V = .555$; $p < .001$						$V = .326$; $p < .001$					

and a post-test decrease in the number of students offering specific type-two explanations; in contrast, the number of students, who offered unspecific type-two explanations remained stable from pre-test to post-test. Post-test differences between the experimental and comparison groups were analyzed using chi-square tests, which revealed statistically significant differences between the two groups (see Table 5), with strong effects for the reproduction item (Cramérs $V = .536$; $p < .001$) and medium effects for the transfer item (Cramérs $V = .390$; $p < .001$). Standardized residuals are listed in Table 5, showing that the significance of the chi-square tests for the post-test findings referred to all three types of explanations, with fewer students than expected in the categories of 'type-one explanations' and 'type-two explanations (unspecific)' and more students than expected in the category of 'type-two explanations (specific)' in the experimental group. The opposite outcome was true for students in the comparison group. The data were analyzed for potential gender differences within the experimental and comparison groups, but no significant differences were found.

The appendix shows close analyses of changes in explanations from pre-test to post-test. Table A1 (Appendix) reveals that 45 students in the experimental group improved their ability to consider the genetic level when explaining evolutionary changes on the post-test reproduction item. In particular, two students moved from type-one explanations in the pre-test to unspecific type-two explanations in the post-test, 24 students from type-one explanations in the pre-test to specific type-two explanations in the post-test, and 19 students from unspecific type-two explanations in the pre-test to specific type-two explanations in the post-test. All the increases were statistically significant (Wilcoxon $z = -5.588$; $p < .001$). Three students showed a decrease in the quality of their explanations. Further, 47 of 49 students who had provided specific type-two explanations in the pre-test did so in the post-test. Similar trends were found for the transfer item in the experimental group, although they were less pronounced than on the reproduction item because 36 students were able to improve their ability to consider the genetic level and 12 students showed a decrease in the quality of their explanations (see Appendix, Table A2). All the increases were statistically significant (Wilcoxon $z = -2.921$; $p = .003$).

General insights into changes from pre-test to post-test in the experimental group can be gained by examining the following responses from the same student:

> Because of trophy hunting, bulls with long tusks die. Bulls with smaller tusks can reproduce more often because they have a better chance of finding a mate. // The development of tusks took a long, long time. Observing this phenomenon over a few years only can provide only few definite answers. Maybe further observation over the next 1000 years can provide the expected answers.
>
> (type-one explanation provided by student 26-12, pre-test item, response to question 1 // question 2)

> Several genes are responsible for body size. Because large fish are caught, alleles for large body size become less frequent. Alleles for small fish are passed on. That's why over many generations alleles for big Atlantic cod have become less frequent in the population. // There are only few alleles left in the population which can contribute to large body size. Further, alleles coding for small fish have become more frequent.
>
> (type-two explanation, specific, provided by student 26-12, post-test reproduction item, response to question 1 // question 2)

> Predation involves hunting different mice. There are alleles responsible for darker shades of fur color, which become less frequent in number. White mice produce more offspring. This leads to a further increase in the number of alleles that code for lighter shades of fur color. // There are few alleles that are responsible for a large variety, and it is difficult for these alleles return to their original number in the population. At the same time, alleles continue to be passed on that code for a few shades of fur color.
>
> (type-two explanation, specific, provided by student 26-12, post-test transfer item, response to question 1 // question 2)

In the comparison group, opposite trends were observed. Eleven students improved their ability to consider the genetic level, when explaining evolutionary changes (see Appendix, Table A1). In particular, two students moved from type-one explanations in the pre-test to unspecific type-two explanations in the post-test, five students from type-one to specific type-two explanations and four students from unspecific to specific type-two explanations. Twenty-three students showed a decrease in the quality of their explanations. Further, 21 of 38 students who had offered specific type-two explanations in the pre-test also did so in the post-test. None of the changes were statistically significant (Wilcoxon $z = -1.644$; n.s.). Similar trends were found for the transfer item in the comparison group, on which nine students were able to improve their ability to consider the genetic level and 26 students showed a decrease in the quality of their explanations (see Appendix, Table A2). These changes were statistically significant (Wilcoxon $z = 2.028$; $p = .043$).

Participants' explanations of evolutionary changes: confusion of levels

Post-test findings for students' references to inheritance of traits vs. inheritance of genes (Table 6) were characterized by decreases in the numbers of students who always wrote that traits – rather than genes – are passed from one generation to the next for both groups of students. In the experimental group, the number decreased from 18 students (17.6%) in the pre-test to one (1%) on the post-test reproduction item and nine (9%) on the post-test transfer item. In the comparison group, there were 16 students (18.4%) who confused genes and traits in the pre-test, 14 students (15.2%) on the post-test reproduction item and nine students (10.2%) on the post-test transfer item. Detailed analyses of changes in students' references to inheritance of traits vs. inheritance of genes between the pre-test and post-test revealed that students in the experimental group who had originally confused genes and traits did not do so in the post-test, mostly by referring implicitly or explicitly to genes/alleles being passed from one generation to the next (see Appendix, Tables A3 and A4). In contrast, students in the comparison group who had originally argued that traits are passed from one generation to the next either continued to argue that traits are passed on or avoided doing so by not referring to inheritance at all (see Appendix, Tables A3 and A4).

Additionally, for the experimental group, there was a post-test increase in the number of students, who explicitly or implicitly explained that genes are passed from one generation to the next (Table 6). In particular, the number of students who explicitly/implicitly mentioned that genes are passed on rose from 27/13 students (26.5%/12.7%) in the pre-test to 47/40 students (45.6%/38.8%) on the post-test reproduction item and 32/34 students (32%/34%) on the post-test transfer item. In the comparison group, in contrast, the respective number of students either remained stable (for implicit reference to the inheritance of genes) or decreased from pre-test to post-test (for explicit reference to inheritance of genes). In particular, there were 20/17 students (23%/19.5%) in the pre-test, 10/20 students (10.9%/21.7%) on the post-test reproduction item, and 13/20 students (14.8%/22.7%) on the post-test transfer item who explicitly/implicitly mentioned inheritance of genes. These trends coincided with decreasing numbers of students in the experimental group, making no reference to inheritance at all (pre-test: 35 students, 34.3%; post-test reproduction item: 11 students, 11.7%; post-test transfer item: 19 students, 19%). This trend was reversed in the comparison group, in which the respective number of students increased from 27 students (31%) in the pre-test to 42 students (45.7%) on the post-test reproduction item and 42 students (47.7%) on the post-test transfer item. Qualitative analyses of student responses showed that many students did not explicitly or implicitly describe what is passed from one generation to the next because they focused on describing genetic and/or phenotypic changes in the population before and after selection. For example, a typical pattern of response was the following. Before selection, there were many organisms with the phenotypic and/or genetic feature *x*; after selection, there were many organisms with the phenotypic and/or genetic feature *y*. Additionally, the length of responses decreased from pre-test to post-test, which was true for students in both the experimental and comparison groups (see Appendix, Table A5).

Within the experimental group, the changes from pre-test to post-test (reproduction) and from pre-test to post-test (transfer) were statistically significant (Wilcoxon $z = -5.490$; $p < .001$; Wilcoxon $z = -3.081$; $p = .002$). The same outcome was true for changes in student responses from pre-test to post-test (reproduction) within the comparison group (Wilcoxon $z = -2.833$; $p = .005$) but not for the changes between pre-test and post-test (transfer) (Wilcoxon $z = -1.522$; $p = .128$). Post-test differences between the experimental and comparison groups were analyzed using chi-square tests. The analyses revealed significant differences between the two groups (see Table 6), with strong effects for the reproduction item (Cramérs $V = .555$; $p < .001$) and medium effects for the transfer item (Cramérs $V = .326$; $p < .001$). The standardized residuals listed in Table 6 show that the categories 'no reference to inheritance', 'inheritance of traits', 'explicit reference to inheritance of traits', and 'implicit reference to inheritance of traits' contributed to the differences between the groups in the post-test reproduction item. However, the category 'inheritance of traits' did not contribute to the significance of the test in the post-test transfer item, which was obvious because similar numbers of students in both groups confused genes and traits. The data were analyzed for potential gender differences within the experimental and comparison groups, but no significant differences were found.

Discussion and implications

General discussion

As argued in the 'Introduction' section, many students lack the ability to distinguish between different levels of biological organization and to interrelate them when explaining biological phenomena. The so-called micro–macro problem is widespread and it represents a central challenge in teaching biology (see, e.g. Knippels, 2002; Lijnse, Licht, de Vos, & Waarlo, 1990; Verhoeff, 2003). Accordingly, fostering students' abilities to think across levels is understood as an important aim for biology instruction (Parker et al., 2012; Wilensky & Resnick, 1999; Wilson et al., 2006). Essentially, this argument is based on the premise that a coherent understanding of biological phenomena 'includes both interrelating biology concepts at each level of organization and interrelating concepts at the different levels of biological organization, i.e. horizontal and vertical integration respectively' (Verhoeff, 2003, p. 151). The former aspect is also called 'horizontal coherence', whereas the latter is called 'vertical coherence' (Verhoeff et al., 2008).

Building on the distinction between horizontal and vertical coherence, we investigated (the lack of) vertical coherence in the students' understanding of evolutionary biology. Horizontal coherence was not studied. In general, vertical coherence denotes students' ability 'to distinguish the levels of biological organisation [and] to see the relationships between those levels' (Knippels, 2002, pp. 75–76). Congruent with this definition, we researched students' ability to avoid confusion of levels and disconnects between levels when explaining evolutionary changes. Framing the present study with an extensive literature review, we were able to synthesize a large number of research findings under these two aspects. In fact, the research review showed how pervasive the lack of vertical coherence is. Because the majority of biological phenomena – if not all – are multi-leveled, students in general are challenged to transect levels and to interrelate them. When they fail to do so, disconnects between levels and confusion of levels can be observed. The point we thus propose to make is that students struggle with explanatory coherence in general and not just in evolutionary biology. The literature on student conceptions in biology thus lends support to the argument that biology educators must support students in their ability to distinguish between concepts at different levels (i.e. to avoid confusion) and to interrelate them (i.e. to avoid disconnects).

Discussion of pre-test and post-test findings

This study researched whether or not – and to what extent – a lab activity that addresses different levels of biological organization (i.e. phenotype and genotype) helps students to distinguish between them and promotes vertical coherence in the students' understanding of evolutionary changes. The study yielded two main findings. On the one hand, an analysis of the pre-test findings showed that many students lacked vertical coherence – that is, their explanations of evolutionary changes showed disconnects between levels and confusion of levels. This problem has not yet received much attention in the literature on evolution education. On the other hand, the analysis of the post-test findings showed that thinking across levels can be promoted if students are afforded the opportunity to engage with different levels of biological organization and to learn how they are interrelated. In particular, disconnects between levels remained typical of student post-test explanations in the comparison group but not in the experimental group (see Table 5).

The analysis of the pre-test findings helped to assess the extent to which students grapple with thinking across levels at the beginning of a unit on evolution. Whereas one of the problems with which students struggle has been focused on for some time – confusion of genes and traits (e.g. Lewis & Kattmann, 2004; Marbach-Ad & Stavy, 2000) – the analysis of the pre-test findings showed that approximately one-third of the students did not use genetic terms and concepts at all when explaining evolutionary changes, although they had been trained in genetics before the unit on evolution. This finding is important because it reveals that a considerable number of students did not use their knowledge of genetics, which might be interpreted as inert knowledge (Bransford et al., 2000). Additionally, this finding mirrors the difficulties of integrating genetic and evolutionary knowledge reported by Halldén (1988), to whom we referred in the introduction. Biology educators must be aware of this problem and must plan evolution instruction accordingly. Further support for the advantages of integrating genetics and evolution instruction comes from Kampourakis and Zogza (2007, 2008, 2009) as well as Banet and Ayuso (2003), who described decreases in students' teleological explanations after teaching genetics in the context of evolution.

As a second major finding, the post-test effects differed between the two groups. Providing vertical coherence (i.e. interrelating concepts at different levels of organization), the lab activity 'Why are Atlantic Cod shrinking?' demonstrated to the students how selection impacts both phenotypes and allele frequencies, and it helped them to avoid both disconnects between levels and confusion of levels on the post-test reproduction item. In contrast, the students in the comparison group continued to grapple with these problems after engaging in a lab activity that illustrated phenotypic changes alone. On the post-test transfer item, the advantages of a learning environment providing vertical coherence remained evident, but they were less pronounced than on the reproduction item. To some extent, the attenuation of the effects observed in the reproduction item might be the result of similar item formats being used for pre-test and post-test reproduction and transfer items, which might have affected student motivation to write elaborate responses throughout the test, as the consistently shorter responses show (see Appendix, Table A5). Nevertheless, as a main implication of this study, learning activities can be recommended that interrelate concepts at different levels of biological organization, helping students to distinguish between them and promoting thinking across levels.

Practical implications

The focus of this study – promoting reasoning across levels in evolutionary biology – is related to the yo–yo learning and teaching strategy, which has proven helpful for providing coherence in upper secondary students' understanding of genetics and cell biology (Knippels, 2002; Verhoeff, 2003). Using research methods from design-based research, both of these authors described qualitative evidence for the effectiveness of yo–yo learning. Although we did not research the yo–yo learning and teaching strategy itself, the findings of the present study add to this literature. This study offers quantitative support for the argument that thinking across levels is difficult for students and must be fostered with learning activities that promote vertical coherence. However, there are also differences between the yo–yo learning and teaching strategy and the design principles of the lab activity 'Why are Atlantic Cod shrinking?' A key element of yo–yo learning is problem-posing, which involves teachers formulating questions, preferably questions explicitly indicating the level to which the students are expected to ascend and descend. Thus, fostering thinking across levels, as proposed by Knippels (2002) and Verhoeff (2003), involves teachers' guidance, which is considered disadvantageous by the authors themselves because students' thinking across levels might not have occurred independently of the teachers' (Knippels, 2002, p. 116). For the lab activity 'Why are Atlantic Cod shrinking?', in contrast, it was the structure of the lab activity itself – rather than a series of teacher questions – that guided the students' thinking across levels, which allowed for student-centered learning in the lab activity.

Can the lab activity 'Why are Atlantic Cod shrinking?' be used as a model for researchers and biology educators, and can it be developed further? Transferring the lab activity to other fields essentially requires identification of the general design principles that contribute to fostering thinking across levels. Our analysis of lab activities is helpful in this regard. On the one hand, we were able to show that concepts at different levels of biological organization must be represented as separate conceptual entities. On the other hand, student activities must be involved that encourage students to think back and forth between the levels of biological organization. We consider both aspects essential for fostering thinking across levels. The lab activity 'Why are Atlantic Cod shrinking', however, can also be developed further because it did not encourage students to reflect about their thinking on the meta-level, particularly meta-reflection about the question of which levels were transected (see, e.g. Verhoeff et al., 2008). As a next step, lab activities aimed at fostering thinking across levels should include this aspect, and research should focus on the impact of meta-cognition. Meta-cognition, in fact, might be beneficial to thinking across levels because phenomena at the macro-level are visible and concrete. Students who lack meta-cognitive awareness are thus likely to pay phenomena at the macro-level greater attention than concepts located on the micro-level, which are frequently invisible and abstract.

Future directions

Given the importance to biology education of promoting reasoning across levels, empirical support for the effectiveness of learning and teaching strategies that aim at promoting vertical coherence are crucial. The present study offers such proof. Faced with the multi-level nature of biological systems, students struggle with connecting – and avoiding confusing with – concepts at different levels of biological organization. However, vertical coherence in students' understanding can be promoted through learning and teaching strategies that help students distinguish between different levels of organization and to interrelate concepts at different levels of organization.

Thus, our study contributes to the literature an exploration of how different lab activities that focus on the consequences of selection impact students' abilities to reason across levels when explaining evolutionary changes. The strengths of this study are as follows: (a) developing a theory-based approach to identifying disconnects between levels and confusion of levels as specific problems with students' reasoning across levels in evolution education and in other areas of biology instruction; (b) identifying theory-based criteria for analyzing lab activities in terms of their potential to provide vertical coherence in students' evolutionary explanations; and (c) empirically testing the hypothesis that vertical coherence can be promoted by learning activities that address different levels of organization, interrelate them, and help to distinguish between them. Although the present study focused on evolution instruction, the findings are applicable to all 'biological topics that transect the different levels of organisation' (Knippels, 2002, p. 154). Promoting reasoning across levels can thus be recommended for all fields of biology education in which students struggle with disconnects between levels and confusion of levels, such as systems thinking (e.g. Penner, 2000), cell biology (e.g. Flores et al., 2003), genetics (e.g.

Marbach-Ad & Stavy, 2000), physiology (e.g. Brown & Schwartz, 2009), ecology (e.g. Ebert-May et al., 2003), and evolution (e.g. Ferrari & Chi, 1998; Shtulman, 2006).

Limitations of the study

The study met methodological challenges that require discussion. Analysis of the pre-test findings revealed that randomization was successful regarding the baseline characteristics measured: the two groups did not differ in terms of their ability to define central genetic concepts and their ability to explain evolutionary changes. The gender ratios, however, differed within the groups, which might have occurred because teachers, in some cases, randomized groups of girls and boys who wanted to engage in the lab activity together, rather than randomizing them purely as individuals. However, post hoc analyses of gender differences within the groups revealed that failure to randomize students for gender did not result in bias in the pre-test and post-test findings. Additionally, development of the lab activity in the comparison group met constraints such that the two lab activities studied also differed in factors other than providing vs. not-providing vertical coherence. In particular, the two lab activities differed in terms of organisms, traits and authenticity, and no student variables were assessed in this study to account for the potential influence of these variables. However, the post-test findings did not provide any evidence that these factors affected student performance differently in the two groups. Instead, interpretation of the findings for the post-test transfer item showed that students wrote consistently shorter responses in both groups, which might be understood as an indication that students in both groups were struggling equally with motivational aspects when taking the test. Another limitation of this study was that the point in time of the post-test varied from one day to five days after the lab activity, depending on when the class was scheduled. Due to small sample size and diversity of learning trajectories, it was impossible to investigate if variation in point in time affected post-test performance. Further, follow-up tests were not administered because the participating students left high school shortly after the study. Accordingly, it is currently unknown whether the students' post-test gains were maintained.

Conclusions

Considering both its strengths and limitations, this study provided evidence for the argument that learning and teaching strategies that distinguish among different levels of organization and that interrelate concepts at different levels of organization help students to reason across levels of biological organization (Knippels, 2002; Knippels et al., 2005; Parker et al., 2012; Tsui & Treagust, 2013; Verhoeff et al., 2008). It thus adds to the literature on reasoning across levels. It also contributes to the literature on integrative approaches to evolution education, where the argument has been made that 'a complete understanding of evolution requires the knowledge that the result of natural selection is a change in allele frequencies within populations' (White, Heidemann, & Smith, 2013, p. 593). Additionally, in the literature on genetics education, researchers have argued for including more examples of polygenetic inheritance (Dougherty, 2009; Mills Shaw, Van Horne, Zhang, & Boughman, 2008). In more general terms, the empirical findings of this study are central to promoting educators' theory-based understanding of the subject-matter-specific factors that affect explanatory coherence. Additionally, the design principles that inspired the development of 'Why are Atlantic Cod shrinking?' have the potential to contribute to effective biology teaching in other contexts. Beyond the specific example studied in this paper – interrelating phenotype and genotype – empirical evidence must be sought in other areas of biology instruction to support efforts to promote reasoning across levels, which might be considered a major challenge and a central topic of biology instruction.

Acknowledgements

The lab activity 'Why are Atlantic Cod shrinking?' was developed in the context of 'Design and Evaluation of Teaching Materials for the Evolution of Life', a project aimed at developing new teaching materials funded by VolkswagenStiftung on the occasion of the Darwin bicentennial. We would like to thank Erich Bornberg-Bauer, professor of evolutionary biology at University of Münster, for his helpful advice and the participating teachers for their support.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Janina Jördens is a certified biologist currently working on a Ph.D. thesis in biology education at Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Germany. Her dissertation focuses on different aspects of teaching evolutionary biology to high school students.

Roman Asshoff studied Biology and Philosophy at Jena University and Leipzig University, Germany. He received a Ph.D. in Ecology from Basel University, Switzerland. Currently, he is a senior lecturer in biology education at Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Germany. Research areas include student conceptions and scientific inquiry.

Harald Kullmann studied Biology at Bonn University, Germany, where he did his Ph.D. at the Institute for Evolutionary Biology and Ecology. Afterwards, his research focused on several aspects of evolutionary ecology of fishes. Currently, he is a senior lecturer at Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Germany.

Marcus Hammann studied Biology and English at Kiel University, Germany, and Kansas University, USA. He received a Ph.D. in biology education from Leibniz-Institute of Science Education (IPN), Kiel, Germany. Currently he is a professor for biology education at Zentrum für Didaktik der Biologie, Westfälische Wilhelms-Universität Münster, Germany. His research focuses on cognitive and affective aspects of biology teaching and learning.

ORCID

Janina Jördens http://orcid.org/0000-0002-5051-4240
Roman Asshoff http://orcid.org/0000-0002-1001-4479
Harald Kullmann http://orcid.org/0000-0002-6540-2436
Marcus Hammann http://orcid.org/0000-0003-4820-5219

References

Allen, J. H., & Wold, J. (2009). Investigating contemporary evolution via size-selective harvesting. *The American Biology Teacher*, *71*(3), 151–155.

Anderson, C. W., Sheldon, T. H., & Dubay, J. (1990). The effect of instruction on college nonmajors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching*, *27*(8), 761–776.

Banet, E., & Ayuso, G. E. (2003). Teaching of biological inheritance and evolution of living beings in secondary school. *International Journal of Science Education*, *25*(3), 373–407.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role on evolution. *Journal of Research in Science Teaching*, *27*(5), 415–427.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brown, M. H., & Schwartz, R. S. (2009). Connecting photosynthesis and cellular respiration: Preservice teachers' conceptions. *Journal of Research in Science Teaching*, *46*(7), 791–812.

Brumby, M. N. (1984). Misconceptions about the concept of natural selection by medical biology students. *Science Education*, *68*(4), 493–503.

Canal, P. (1999). Photosynthesis and 'inverse respiration' in plants: An inevitable misconception? *International Journal of Science Education*, *21*(4), 363–371.

Cavallo, A. M. L., & McCall, D. (2008). Seeing may not mean believing: Examining students' understandings and beliefs in evolution. *The American Biology Teacher*, *70*(9), 522–530.

Chi, M. (2005). Commonsense conceptions of emergent processes. Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*(2), 161–199.

Christensen-Dalsgaard, J., & Kanneworff, M. (2009). Evolution in Lego®: A physical simulation of adaptation by natural selection. *Evolution: Education and Outreach*, *2*(3), 518–526.

Conover, D. O., & Munch, S. B. (2002). Sustaining fisheries yields over evolutionary time scales. *Science*, *297*, 94–96.

Deadman, J. A., & Kelly, P. J. (1978). What do secondary school boys understand about evolution and heredity before they are taught about the topics? *Journal of Biological Education*, *12*(1), 7–15.

Demastes, S. S., Good, R., & Peebles, P. (1995). Students' conceptual ecologies and the process of conceptual change in evolution. *Science Education*, *79*, 637–666.

Deniz, H., Donnelly, L. A., & Yilmaz, I. (2008). Exploring the factors related to acceptance of evolutionary theory among Turkish pre-service biology teachers: Toward a more informative conceptual ecology for biological evolution. *Journal of Research in Science Teaching*, *45*(4), 420–443.

Dougherty, M. J. (2009). Closing the gap: Inverting the genetics curriculum to ensure an informed public. *The American Journal of Human Genetics*, *85*, 6–12.

Dreyfus, A., & Jungwirth, E. (1989). The pupil and the living cell: A taxonomy of dysfunctional ideas about an abstract idea. *Journal of Biological Education*, *23*, 49–55.

Duncan, R. G., & Reiser, B. J. (2007). Reasoning across ontologically distinct levels: Students' understandings of molecular genetics. *Journal of Research in Science Teaching*, *44*(7), 938–959.

Ebert-May, D., Bazli, J., & Lim, H. (2003). Disciplinary research—strategies of assessment of learning. *BioScience*, *53*, 1221–1228.

Fenberg, P. B., & Roy, K. (2008). Ecological and evolutionary consequences of size-selective harvesting: How much do we know? *Molecular Ecology*, *17*, 209–220.

Ferrari, M., & Chi, T. H. (1998). The nature of naïve explanations of natural selection. *International Journal of Science Education*, *20*(10), 1231–1256.

Fifield, S., & Fall, B. (1992). A hands-on simulation of natural selection in an imaginary organism Platysoma apoda. *The American Biology Teacher*, *54*(4), 230–235.

Flores, F., Tovar, M., & Gallegos, L. (2003). Representation of the cell and its processes in high school students. An integrated view. *International Journal of Science Education*, *25*(2), 269–286.

Frey, F., Lively, C., & Brodie, E. (2010). Selection and evolution with a deck of cards. *Evolution: Education and Outreach*, *3*(1), 114–120.

Halldén, O. (1988). The evolution of the species: Pupil perspectives and school perspectives. *International Journal of Science Education*, *10*(5), 541–552.

Kampourakis, K., & Zogza, V. (2007). Students' preconceptions about evolution: How accurate is the characterization as "Lamarckian" when considering the history of evolutionary thought? *Science & Education*, *16*(3–5), 393–422.

Kampourakis, K., & Zogza, V. (2008). Students' intuitive explanations of the causes of homologies and adaptations. *Science & Education*, *17*(1), 27–47.

Kampourakis, K., & Zogza, V. (2009). Preliminary explanations: A basic framework for conceptual change and explanatory coherence in evolution. *Science & Education*, *18*, 1313–1340.

Kelemen, D. (2012). Teleological minds – how natural intuitions about agency and purpose influence learning about evolution. In K. S. Rosengren, S. K. Brem, E. M. Evans, & G. M. Sinatra (Eds.), *Evolution challenges* (pp. 66–92). Oxford: Oxford University Press.

Knippels, M.-C. P. J. (2002). *Coping with the abstract and complex nature of genetics in biology education: The yo-yo learning and teaching strategy*. Utrecht: CD-β Press.

Knippels, M.-C. P. J., Waarlo, A. J., & Boersma, K. T. (2005). Design criteria for learning and teaching genetics. *Journal of Biological Education*, *39*(3), 108–112.

Konnemann, C., Asshoff, R., & Hammann, M. (2012). Einstellungen zur Evolutionstheorie. Theoretische und messtheoretische Klärungen [Attitudes towards evolutionary theory. Theoretical and psychometric issues]. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 55–79.

Lauer, T. E. (2000). Jelly Belly® Jelly Beans & evolutionary principles in the classroom: Appealing to the students' stomachs. *The American Biology Teacher*, *62*(1), 42–45.

Lewis, J., & Kattmann, U. (2004). Traits, genes, particles and information: revisiting students' understanding of genetics. *International Journal of Science Education*, *26*, 195–206.

Lewis, J., Leach, J., & Wood-Robinson, C. (2000). All in the genes? – young people's understanding of the nature of genes. *Journal of Biological Education*, *34*(2), 74–79.

Lijnse, P. L., Licht, P., de Vos, W., Waarlo, A. J. (Eds.). (1990). *Relating macroscopic phenomena to microscopic particles – a central problem in secondary science education*. Utrecht: CD-ß Press.

Manokore, V., & Williams, M. (2012). Middle school students' reasoning about biological inheritance: Student's resemblance theory. *International Journal of Biology Education*, *2*(1), 1–31.

Marbach-Ad, G., & Stavy, R. (2000). Students' cellular and molecular explanations of genetic phenomena. *Journal of Biological Education*, *34*(4), 200–205.

Maret, T. J., & Rissing, S.W. (1998). Exploring genetic drift & natural selection through a simulation activity. *The American Biology Teacher*, *60*(9), 681–683.

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research*, *1*(2), Art. 20. Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204

Meir, E., Perry, J., Stal, D., Maruca, S., & Klopfer, E. (2005). How effective are simulated molecular-level experiments for teaching diffusion and osmosis? *Cell Biology Education*, *4*(3), 235–248.

Mills Shaw, K. R., Van Horne, K., Zhang, H., & Boughman, J. (2008). Essay contest reveals misconceptions of high school students in genetics content. *Genetics*, *178*(3), 1157–1168.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, *48*(3), 237–256.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, *45*(10), 1131–1160.

Nehm, R., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *BioScience*, *57*(3), 263–272.

Nold, D. (2010). Sozioökonomischer Status von Schülerinnen und Schülern 2008 – Ergebnisse des Mikrozensus [Socioeconomic status of students 2008 – results of the microcensus]. *Statistisches Bundesamt – Wirtschaft und Statistik*, *2*, 138–149.

Parker, J. M., Anderson, C. W., Heidemann, M., Merrill, M., Merritt, B., Richmond, G., & Urban-Lurain, M. (2012). Exploring undergraduates' understanding of photosynthesis using diagnostic question clusters. *CBE-Life Sciences Education*, *11*(1), 47–57.

Penner, D. E. (2000). Explaining systems: Investigating middle school students' understanding of emergent phenomena. *Journal of Research in Science Teaching*, *37*(8), 784–806.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (2007). *PISA 2006 – Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006 – results of the third international comparative study]. Münster: Waxman.

Resnick, M. (1996). Beyond the centralized mindset. *Journal of the Learning Sciences*, *5*(1), 1–22.

Rutledge, M. L., & Warden, M. A. (2000). Evolutionary theory, the nature of science & high school biology teachers: Critical relationships. *The American Biology Teacher*, *62*(1), 23–31.

Scheersoi, A., & Kullmann, H. (2007). Gendrift und Selektion spielerisch vermitteln [Playfully teaching genetic drift and selection]. *Praxis der Naturwissenschaften - Biologie in der Schule*, *56*(7), 45–47.

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

Sinatra, G. M., & Mason, L. (2008). Beyond knowledge: Learner characteristics influencing conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 560–582). New York, NY: Routledge.

Sinatra, G. M., Southerland, S.A., Mc Conaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in student's understanding and acceptance of biological evolution. *Journal of Research in Science Teaching*, *40*(5), 510–528.

Slack, S. J., & Stewart, J. (1990). High school students' problem-solving performance on realistic genetics problems. *Journal of Research in Science Teaching*, *27*(1), 55–67.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: Successful versus unsuccessful performance. *Journal of Research in Science Teaching*, *21*(9), 895–912.

Songer, C. J., & Mintzes, J. J. (1994). Understanding cellular respiration: An analysis of conceptual change in college biology. *Journal of Research in Science Teaching*, *31*(6), 621–637.

Stavy, R., Eisen, Y., & Yaakobi, D. (1987). How students aged 13–15 understand photosynthesis. *International Journal of Science Education*, *9*(1), 105–115.

Stebbins, R. C., & Allen, B. (1975). Simulating evolution. *The American Biology Teacher*, *37*, 206–211.

Tsui, C.-Y., & Treagust, D. F. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073–1098.

Tsui, C.-Y., Treagust, D. F. (2013). Introduction to multiple representations: Their importance in biology and biology education. In D. F. Treagust & C.-Y. Tsui (Eds.), *Multiple representations in biology education* (pp. 3–18). Dordrecht: Springer.

Verhoeff, R. P. (2003). *Towards system-thinking in cell biology education*. Utrecht: CD-ß-Press.

Verhoeff, R. P., Waarlo, A. J., & Boersma, K. T. (2008). Systems modelling and the development of coherent understanding of cell biology. *International Journal of Science Education*, *30*(4), 543–568.

Waheed, T., & Lucas, A. (1992). Understanding interrelated topics: Photosynthesis at age14+. *Journal of Biological Education*, *26*(3), 193–199.

Wallin, A. (2008). One year after teaching, how consistent are students in using the scientific theory of biological evolution by natural selection? In M. Hammann, M. Reiss, C. Boulter, & S. D. Tunnicliffe (Eds.), *Biology in context: Learning and teaching for the twenty-first century. A selection of papers presented at the VIth conference of European Researchers in Didactics of Biology (ERIDOB)* (pp. 52–63). London: Institute of Education.

White, P. J. T., Heidemann, M. K., & Smith, J. J. (2013). A new integrative approach to evolution education. *BioScience*, *63*(7), 586–594.

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and Technology*, *8*(1), 3–19.

Wilson, C. D., Anderson, C. W., Heidemann, M., Merrill, J. E., Merritt, B. W., Richmond, G., … Parker, J. M. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *CBE-Life Sciences Education*, *5*(4), 323–331.

Appendix

Table A1. Changes in students' types of explanations of evolutionary change from pre-test to post-test reproduction item (experimental group and comparison group); numbers in bold show students who improved their explanations.

	Post-test reproduction item – type of explanation							
	Type-one explanation (disconnect)		Type-two explanation (unspecific)		Type-two explanation (specific)		Total	
Pre-test – type of explanation	Experimental group	Comparison group	Experimental group	Comparison group	Experimental group	Comparison group	Experimental group	Comparison group
Type-one explanation (disconnect)	5	23	**2**	**2**	**24**	**5**	31	30
Type-two explanation (unspecific)	1	6	2	5	**19**	**4**	22	17
Type-two explanation (specific)	1	6	1	11	47	21	49	38
Total	7	37	5	18	90	30	102	85

Table A2. Changes in students' types of explanations of evolutionary change from pre-test to post-test transfer item (experimental group and comparison group); numbers in bold show students who improved their explanations.

	Post-test transfer item – type of explanation							
	Type-one explanation (disconnect)		Type-two explanation (unspecific)		Type-two explanation (specific)		Total	
Pre-test – type of explanation	Experimental group	Comparison group	Experimental group	Comparison group	Experimental group	Comparison group	Experimental group	Comparison group
Type-one explanation (disconnect)	9	21	**1**	**0**	**18**	**6**	28	27
Type-two explanation (unspecific)	3	8	2	5	**17**	**3**	22	16
Type-two explanation (specific)	7	10	2	8	40	20	49	38
Total	19	39	5	13	75	29	99	81

Table A3. Changes in students' references to inheritance of traits vs. inheritance of genes from pre-test to post-test reproduction item (experimental group = exp., comparison group = comp.).

	Post-test reproduction item											
	No reference to inheritance		Inheritance of traits		Inheritance of traits and genes		Inheritance of genes (implicit)		Inheritance of genes (explicit)		Total	
Pre-test	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.
No reference to inheritance	9	20	1	3	0	0	13	3	12	1	35	27
Inheritance of traits	0	6	0	7	2	1	6	1	10	1	18	16
Inheritance of traits and genes	0	0	0	1	1	3	3	2	5	1	9	7
Inheritance of genes (implicit)	0	6	0	1	0	0	7	5	6	4	13	16
Inheritance of genes (explicit)	2	8	0	1	1	2	10	6	14	2	27	19
Total	11	40	1	13	4	6	39	16	47	10	102	85

Table A4. Changes in students' references to inheritance of traits vs. inheritance of genes from pre-test to post-test transfer item (experimental group = exp.; comparison group = comp.).

	Post-test transfer item											
	No reference to inheritance		Inheritance of traits		Inheritance of traits and genes		Inheritance of genes (implicit)		Inheritance of genes (explicit)		Total	
Pre-test	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.	Exp.	Comp.
No reference to inheritance	11	15	1	3	1	0	14	3	5	2	32	23
Inheritance of traits	2	8	3	4	1	1	5	0	7	3	18	16
Inheritance of traits and genes	1	1	0	1	1	1	2	1	5	3	9	7
Inheritance of genes (implicit)	1	7	0	1	0	0	7	7	5	1	13	16
Inheritance of genes (explicit)	4	8	5	0	3	2	5	5	10	4	27	19
Total	19	39	9	9	6	4	33	16	32	13	99	81

Table A5. Number of words in students' responses to items requiring explanations of evolutionary change; mean scores M, standard deviations (SD), and *t*-test statistics.

	Pre-test, M (SD)	Post-test reproduction item, M (SD)	Post-test transfer item, M (SD)	*t*-Test (paired samples)
Experimental group	83.62 (25.19)	63.92 (23.69)	47.75 (25.58)	Pre-test/reproduction $t(102) = 6.606$; $p < .001$ Pre-test/transfer $t(102) = 11.126$; $p < .001$ Reproduction/transfer $t(102) = 5.802$; $p < .001$
Comparison group	75.52 (34.60)	65.69 (23.10)	53.95 (29.64)	Pre-test/reproduction $t(90) = 2.643$; $p = .01$ Pre-test/transfer $t(90) = 4.883$; $p < .001$ Reproduction/transfer $t(90) = 4.174$; $p < .001$
t-Test (independent samples)	$t(168) = 1.863$; $p = .064$	$t(192) = -0.525$; $p = .600$	$t(192) = -1.563$; $p = .120$	