

RESEARCH ARTICLE

Open Access



EvoSketch: Simple simulations for learning random and probabilistic processes in evolution, and effects of instructional support on learners' conceptual knowledge

Daniela Fiedler^{1*} , Steffen Tröbst^{2,3}, Jörg Großschedl⁴ and Ute Harms¹

Abstract

Background: Students' knowledge of scientific principles of evolution is often inadequate, despite its recognized importance for understanding biology. Moreover, difficulties associated with underlying abstract concepts such as randomness and probability can hinder successful learning of evolutionary concepts. Studies show that visualizations, particularly simulations together with appropriate instructional support, facilitate the learning of abstract concepts. Therefore, we have developed interactive, web-based simulation software called EvoSketch in efforts to help learners grasp the nature and importance of random and probabilistic processes in evolutionary contexts. We applied EvoSketch in an intervention study comparing four self-directed study conditions: learning with EvoSketch (1) alone, (2) combined with interpretative support, (3) combined with reflective support, and (4) using texts about randomness and probability instead of EvoSketch. All conditions received no support from any instructors. Knowledge about evolution as well as randomness and probability in the context of evolution, time-on-task, and perceived cognitive load were measured. A sample of 269 German secondary school students ($M_{\text{age}} = 15.6$ years, $SD = 0.6$ years) participated in the study.

Results: Learners using EvoSketch without additional support obtained higher follow-up test scores regarding their knowledge of randomness and probability than those using the text-based approach. However, use of the simulations together with given instructional support (interpretative or reflective) did not increase students' performance, relative to the text-based approach. In addition, no significant between-intervention differences were found concerning the knowledge of evolution, while significant differences between the groups were detected concerning students' perceived cognitive load and time-on-task.

Conclusions: From our findings, we conclude that EvoSketch seems to have a very small positive effect on students' understanding of randomness and probability. Contrary to our expectations, additional self-directed instructional support did not improve students' understanding, probably because it was not necessary to understand EvoSketch simulations. When using EvoSketch in the classroom, we recommend increasing the intervention timeframe to several sessions and a variety of evolutionary examples for which EvoSketch serves as an underlying framework.

Keywords: Evolution education, Threshold concepts, Randomness and probability, Web-based simulations, Instructional support, Secondary school students

*Correspondence: fiedler@ipn.uni-kiel.de

¹ Department of Biology Education, Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University, Olshausenstraße 62, 24118 Kiel, Germany

Full list of author information is available at the end of the article



Introduction

Learners have well-documented problems with understanding and learning key scientific concepts like energy (e.g., Opitz et al. 2017; Wernecke et al. 2018), genetics (e.g., Schmiemann et al. 2017; Venville et al. 2005), and evolution (e.g., Gregory 2009; Rector et al. 2013; Rosengren et al. 2012). A shared aspect of these scientific concepts is that spatial and/or temporal dimensions of associated processes and structures prevent their direct perception. Hence, they can only be understood on an imaginary level, like all concepts beyond humans' perceptual (especially visible) dimensions (Lakoff 1987; Lakoff and Johnson 1980). For instance, random mutations in DNA are important sources of variation in the key evolutionary process of natural selection (Heams 2014). However, these mutations are not visible to the naked human eye, although they can be visualized technologically (e.g., using DNA sequencing techniques). The consequent lack of possibilities for students to observe these phenomena in everyday situations may result in misunderstanding of the importance of random processes in evolution (Garvin-Doxas and Klymkowsky 2008). Furthermore, students tend to frequently misunderstand general abstract concepts that underlie biological processes like randomness and probability (Garvin-Doxas and Klymkowsky 2008; Mead and Scott 2010). Thus, it may be essential to address these underlying abstract concepts to overcome problems in learning evolution (Tibell and Harms 2017). Appropriate visualizations such as simulations may help in overcoming these limitations and make the concepts tangible.

Researchers involved in the EvoVis-project (EvoVis: Challenging Threshold Concepts in Life Science - enhancing understanding of evolution by visualization) have developed interactive, web-based simulation software, called EvoSketch, which allows learners to explore random and probabilistic phenomena associated with the process of natural selection. The software generates a line (representing a reproducing organism) that is replicated by the user for 20 generations. After these 20 generations, the line will normally be shifted either to the right or to the left due to the combination of copying error and a selection process.

Our general expectations were that abstracting the processes of randomness and probability in the context of evolution (by EvoSketch) and actively working with EvoSketch will help students to learn these concepts. As research has shown that simulations are seldomly effective in improving knowledge without instructional support (e.g., Eckhardt et al. 2013; Wouters and van Oostendorp 2013), we also expected that additional self-directed instructional support may better facilitate

learners' self-directed simulation-based learning than a learning opportunity without this support.

Background

Learning evolution and the notion of threshold concepts

Over the past decades, a large body of work on evolution education has indicated several difficulties for learning its essential tenets and examined the diversity of students' alternative conceptions (e.g., Baalman et al. 2004; Beggrow and Nehm 2012; Bishop and Anderson 1990; Gregory 2009; Kampourakis and Zogza 2008; Nehm and Schonfeld 2008; Opfer et al. 2012; Shtulman 2006; Spindler and Doherty 2009; Yates and Marek 2015). One problem is that many words in science lessons such as *adaptation* or *fitness* also appear in other contexts or everyday language with slightly different meanings. This can confuse students and lead to misused scientific terminology (Rector et al. 2013; To et al. 2017). If instructors target these alternative conceptions and meanings to cause cognitive conflicts in students, the students are likely to experience conceptual change (e.g., Posner et al. 1982; Sinatra et al. 2008), which means to replace or reorganize old conceptions with new (scientifically accurate) ones.

Current research also mentioned learning difficulties with those evolutionary concepts that are strongly related to underlying abstract concepts like randomness and probability, so-called threshold concepts (Mead and Scott 2010; Ross et al. 2010). Threshold concepts are described as conceptual gateways that, once passed, open up a new way of thinking and are distinguished from "key" or "core" concepts as they are more than mere building blocks toward understanding within a discipline (Meyer and Land 2003, 2006). Threshold concepts are proposed to have eight characteristics: *transformative* (occasioning a shift in perception and practice), probably *irreversible* (unlikely to be forgotten or unlearned), *integrative* (surfacing patterns and connections), often disciplinary *bounded, troublesome* (dealing with counter-intuitive or alien knowledge), *reconstructive* (reconfiguring learners' prior knowledge), *discursive* (extended language usage), and crossing through *liminal space* (chaotic progress across conceptual terrains; Land 2011; Meyer and Land 2003, 2006; Taylor 2006).

Even though evolution is widely considered as troublesome to learn and teach, evolution itself is not suggested to be a threshold concept, but rather consists of a web of interconnected threshold concepts such as temporal scale, spatial scale, probability, and randomness (Ross et al. 2010; Tibell and Harms 2017). Tibell and Harms (2017) developed a two-dimensional framework connecting evolutionary key concepts with these threshold concepts. They propose that a complete understanding of evolution through natural selection requires the

development of knowledge concerning both evolutionary key and threshold concepts, and the ability to freely navigate through this two-dimensional framework. Moreover, they claim that the conceptual change theory can be connected to the notion of threshold concepts. Thus, understanding threshold concepts is a prerequisite for conceptual change concerning the understanding of particular evolutionary concepts; hence changing alternative conceptions to scientifically sophisticated ones (Tibell and Harms 2017).

In this study, we focused on the threshold concepts of randomness and probability, since research reveals that students particularly struggle with the importance and nature of randomness (Garvin-Doxas and Klymkowsky 2008; Robson and Burns 2011). The term randomness is often used in everyday language to explain that a phenomenon is purposeless as well as without order, predictability or pattern, while scientists use the term to suggest unpredictability without referring to purposelessness (Buiatti and Longo 2013; Mead and Scott 2010; Wagner 2012). In fact, the notion of randomness in evolution is rather specific by speaking about events (e.g., mutations or genetic drift) that are independent of an organisms' need or the directionality provided by the process of natural selection (Heams 2014; Mead and Scott 2010). Thus, mutations are called random because they are not directed to an organisms' adaptation, and it cannot be predicted precisely where and when a mutation will appear (Heams 2014). Although it is, of course, possible to predict the *likelihood* of a mutation occurring at a particular site in a specific sequence, this would fit better into the concept of *probability* rather than randomness. In fact, the term probability refers to the likelihood of a particular outcome in a long run (over multiple events), and is assigned a numerical value between zero and one (Feller 1968). The closer a probability value is to one, the more likely the outcome is. In evolution, natural selection itself can be described as a probabilistic process, if the process of selection is defined as individuals' probabilities to survive and reproduce in a specific environment depending on their particular traits (Tibell and Harms 2017). Thus, evolution through natural selection depends on random genetic mutation leading to a heritable variation on which the probabilistic process of selection can act upon (Andrews et al. 2012; Mix and Masel 2014). Therefore, a clear understanding of randomness and probability is essential for understanding evolution.

Computer simulations as tool to enhance understanding

Computer simulations can be effective tools to handle the intangible nature of scientific concepts such as mutations (Ainsworth and VanLabeke 2004; Plass et al. 2012). They also allow students to visualize processes occurring at

spatial and temporal scales that are difficult or impossible to observe directly (Rutten et al. 2012). Simulations have several advantages over reading textbooks or attending lectures, because they provide opportunities to explore theoretical situations, interact with a simplified version of the focal process(es), and/or change time-scales of events (van Berkum and de Jong 1991). However, research on simulation-based learning has revealed that learners may encounter difficulties during the learning process for two contrasting reasons (de Jong and van Joolingen 1998). One is that simulations can involve complex learning environments, which may overwhelm the learner due to the high amount of information that is conveyed and must be processed (Wouters and van Oostendorp 2013). In stark contrast, minimizing guidance (and thus reducing the amount of information) may reduce the effectiveness of simulation based-learning (Rutten et al. 2012). Therefore, instructional support may be needed to provide suitable learning environments and overcome students' learning difficulties (Kombartzky et al. 2010; Urhahne and Harms 2006). Several kinds of support may be provided in different phases of the learning process in efforts to enhance self-directed simulation-based learning (Zhang et al. 2004):

Interpretative support, given before the interaction, can provide scaffolding for learners to activate prior knowledge, and generate appropriate hypotheses. One way to provide effective interpretative support is to offer accessible domain-specific background information (Reid et al. 2003). Research by Leutner (1993) and Lazonder et al. (2010) indicated that the timing of providing these background information is a critical aspect. Students gained higher knowledge outcomes when the domain-specific background information were accessible before and during the learning process. Then, providing worked examples can also have positive effects on learning outcomes (Spanjers et al. 2011; Yaman et al. 2008). Worked examples consist of a problem followed by a worked-out solution, normally presented in a step-by-step format to the learner (Renkl 2005). A study by Lee et al. (2004) state that students receiving worked examples scored higher in a common assessment, while students working with inquiry discovery scored lower.

Experimental support is provided during an interaction and can scaffold learners' process of scientific inquiry during simulation-based learning by helping them to design verifiable experiments, predict and observe the outcomes, and draw appropriate conclusions. Students often have inefficient experimentation behaviors (e.g., vary too many variables at the same time; de Jong and van Joolingen 1998). Effective experimental support for knowledge acquisition may include gradual, cumulative introductions to handle a simulation and/or requests for

learners to predict and describe the outcome (Urhahne and Harms 2006; Wang et al. 2017). Such experimental prompts are particularly effective for learners with low ability and inefficient discovery learning strategies (Chang 2017; Veenman and Elshout 1995).

Reflective support is provided after an interaction and may foster learners' integration of their discoveries. Such support scaffolds the integration of new information arising from discoveries after learners' interaction with a simulation. It involves promoting reflective processes (sometimes also connected to metacognitive processes), which may be done through a reflective assignment tool or opportunities to discuss the results (de Jong and van Joolingen 1998; White and Frederiksen 1998). Indeed, studies by Eckhardt et al. (2013) and Zhang et al. (2004) concluded that prompting students to reflect upon and justify their experimental activities and outcomes raise their self-awareness and contribute to higher knowledge acquisition. Moreover, Chang and Linn (2013) showed that criticizing someone else's experiment can foster students to recognize poorly arranged experiments so that they can create better experiments on their own, and hence knowledge acquisition is enhanced.

Simulations to support students' understanding of evolution and threshold concepts

Although the number of available online educational videos increases, they often lack explanations regarding underlying threshold concepts or, if mentioned, they are communicated orally only (Bohlin et al. 2017). For evolution education, there are few computer simulations available for free such as *Evolve* (Price and Vaughn 2010), *Avida-ED* (Pennock 2007, 2018), or evolution readiness activities (Concord Consortium 2018). The conducted research studies indicated positive learning gains after using these simulations (Horwitz et al. 2013; Soderberg and Price 2003; Speth et al. 2009). Nevertheless, they were designed to focus on evolutionary (key) concepts without focusing on particular underlying threshold concepts such as randomness and probability. For instance, the activities of evolution readiness focus on the process of (natural) selection, variation within species (without referring to the origin of variation), and inheritance of various traits (Horwitz et al. 2013). This also counts for *evolve*, which is designed to focus on the effects of selection, genetic drift, and migration of a population over time without modeling mutations or their random nature (Soderberg and Price 2003). In contrast, *Avida-ED* includes random mutations occurring in the organisms' genome, while students can also observe evolution in action (Speth et al. 2009). Still, the above-mentioned simulations do not imply the underlying threshold concepts such as randomness or probability. In addition, these

simulations are rather time-consuming (e.g., it takes some time to handle the software properly). Particularly *Avida-ED* seems to work well for lecturers in universities, where students can use this tool across several lab lessons, but this simulation software might be too complex and time-consuming to be used by teachers in ordinary school lessons. Therefore, there is a need for simulation software that is easy to handle and visualizes the notion of randomness in evolution.

EvoSketch

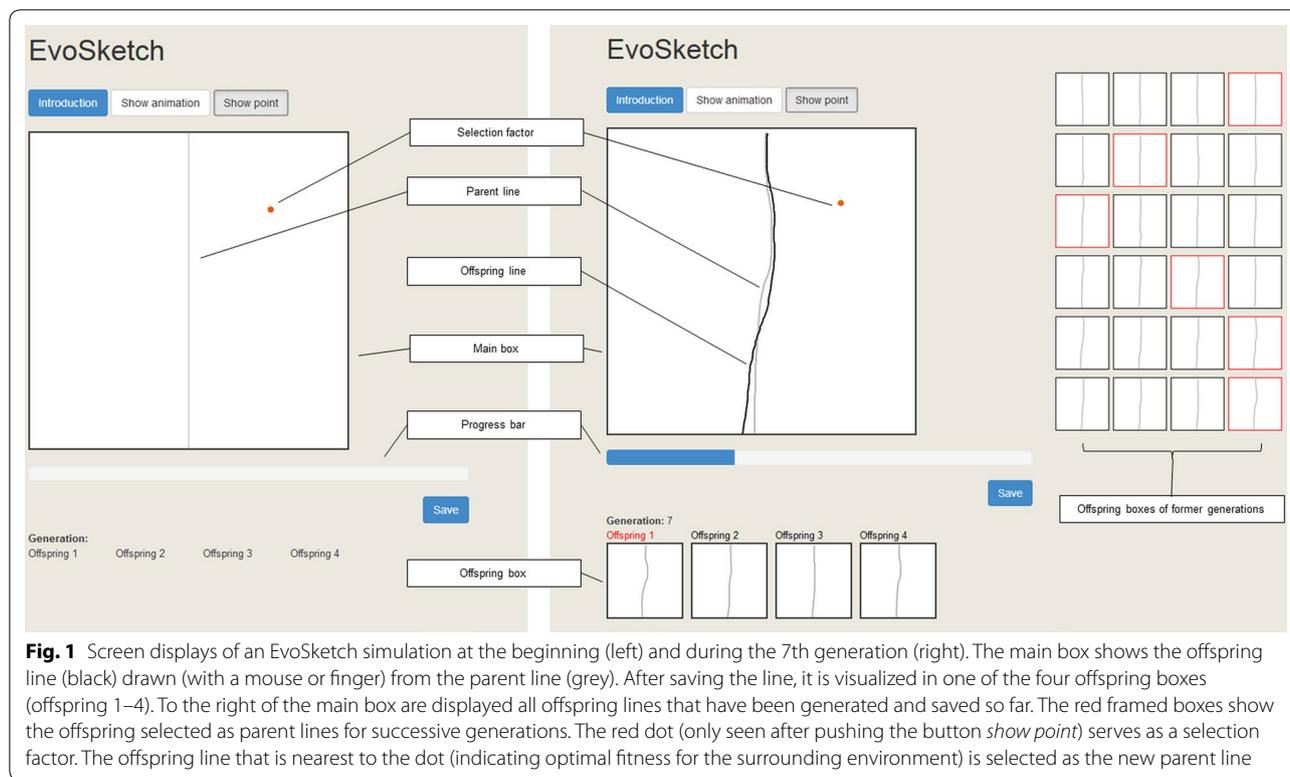
The simulation software

EvoSketch is a project-developed, interactive, web-based simulation software, free of charge and available online in an English (*EvoSketch English* 2018) or German version (used in this study; *EvoSketch German* 2018), that allows learners to explore random and probabilistic phenomena associated with the process of natural selection. The software (which can be used on various electronic devices, such as smartphones, tablets, laptops, and desktop computers) generates a line (representing a reproducing organism) that is replicated by the user for 20 generations.

Every generation consists of four replications of a parent line drawn with a mouse or finger, resulting in four offspring lines (Fig. 1). Since copying errors inevitably occur while drawing, each replication varies and drifts slightly to the right or left of the parent line. These shifts in offspring lines represent the concept of the origin of variation, and hence random processes in evolution. After each generation has been completed by drawing four replications, one of the four offspring lines is selected (by the software) to continue the parent line, and thus represents the next reproducing "organism" in the simulation. The selected line is closest to a point (indicated by the red dot in Fig. 1) indicating optimal fitness for the offspring in the surrounding environment. The organism represented by the selected line has the highest probability to survive and reproduce, and there is selective pressure on the line to move towards the point (probabilistic processes). After 20 generations, the line will normally be shifted either to the right or to the left due to the combination of copying error and selection.

The idea behind the simulation

The idea for *EvoSketch* emerged from a video clip focusing on the visualization and importance of randomness for natural selection to occur (BBC and Open University 2011). Although this video clip was easy to understand, we wanted to create a hands-on activity (i.e., *EvoSketch*) for students to experience these changes on their own. As—according to the theory of embodied cognition (Gropengießer 2007; Lakoff 1987; Lakoff and Johnson



1980)—repeated action within a specific environment (e.g., simulation) helps to create an understanding of scientific concepts. We intended to provide users with a possibility to realize on their own how even tiny mistakes (copying errors) could change the shape of a line across several generations. In addition, students have to draw four offspring lines in contrast to just one line after the other. This should include the aspect of variation across offspring and serves as a basic framework for the process of natural selection.

Although realistic visualizations may facilitate the recognition of the visualized process in the real world (e.g., Höffler and Leutner 2007), realistic visualizations often also entail irrelevant details resulting in distractions of the learners from the relevant parts (Dwyer 1976). In contrast, nonrealistic or schematic visualizations may present such aspects in a way that is easier to realize by the learner (Scheiter et al. 2009). Moreover, decreasing the number of dispensable elements in the learning material might help in reducing extraneous cognitive load (Sweller 1994). Thus, we used a nonrealistic visualization approach for EvoSketch to focus on the notion of randomness and probability.

EvoSketch worksheets

In general, users are guided through an EvoSketch exercise by an accompanying worksheet (EvoSketch

worksheet). This worksheet consists of two introductory texts explaining random processes (specifically, mutations) in evolution [English version: 132 words; German version: 128 words (used in this study)], and the probabilistic process of natural selection [English version: 107 words; German version: 98 words (used in this study)]. Both texts are directly followed by a task asking learners to make predictions about the outcome of the simulation, run the simulation and/or observe the outcome, and explain the outcome (predict-observe-explain strategy; White and Gunstone 1992). The English version of the EvoSketch Worksheet is available as Additional file 1.

Research aim

Our general expectations were that abstracting the processes of randomness and probability in the context of evolution (by EvoSketch) and actively working with EvoSketch will help students to learn these concepts. Further, additional instructional support may better facilitate learners’ self-directed simulation-based learning. Since EvoSketch software provides integrated experimental support (EvoSketch Worksheet tasks), we addressed the potential utility of additional interpretative and reflective instructional support in this study. To evaluate the effectiveness of EvoSketch for teaching and learning the roles and importance of randomness and probability in evolutionary contexts, we used knowledge test performance

on three occasions, time-on-task, and perceived cognitive load of the learners. We compared pre-, post- and follow-up performance scores of students who learned with EvoSketch, with and without additional instructional support (i.e., interpretative or reflective), to those of students who used text-based learning of the same topics.

Methods

Design and interventions

The main aim of this study was to assess the effectiveness of EvoSketch for fostering students' conceptual knowledge of randomness, probability, and evolution. An additional aim was to identify which type of instructional support (if any) most effectively promotes self-directed learning with EvoSketch. For these purposes, we used an experimental repeated measures design approach and assigned students to the following four kinds of self-directed learning interventions (i.e., no additional support of any instructor): text-based, simulation-based, simulation-based with interpretative support, and simulation-based with reflective support. Students participating in all four groups received an overview of the topic of evolution by means of a short, standardized introductory text (cf. Neubrand et al. 2016) to reactivate prior knowledge. The students of each group subsequently individually addressed the following worksheets and tasks:

Text-based intervention (hereafter, text)

Textbooks are still a central teaching resource in science education, and thus, learning is often organized around text-based instructions (McDonald 2016). Learners of this intervention group worked with a worksheet (including the two introductory EvoSketch Worksheet texts mentioned above) and a Powerpoint presentation on the roles of randomness (specifically, mutation) and probability (specifically, selection) in evolution. The presentation included only written texts and pictures (i.e., no audio or video components). Afterwards, learners were asked to answer three questions regarding the information given in the presentation, and two questions regarding evolution. The respective questions as English version are available as Additional file 2.

Simulation-based intervention (hereafter, simulation)

Learners in this group were asked to follow the instructions of the EvoSketch worksheet (mentioned above in section EvoSketch software). They started by reading the introductory text on the topic of randomness in evolution and worked through the first task. During this task, they also progressed through the EvoSketch simulations. They then read the second text on the role of probability in evolution and addressed the second task regarding

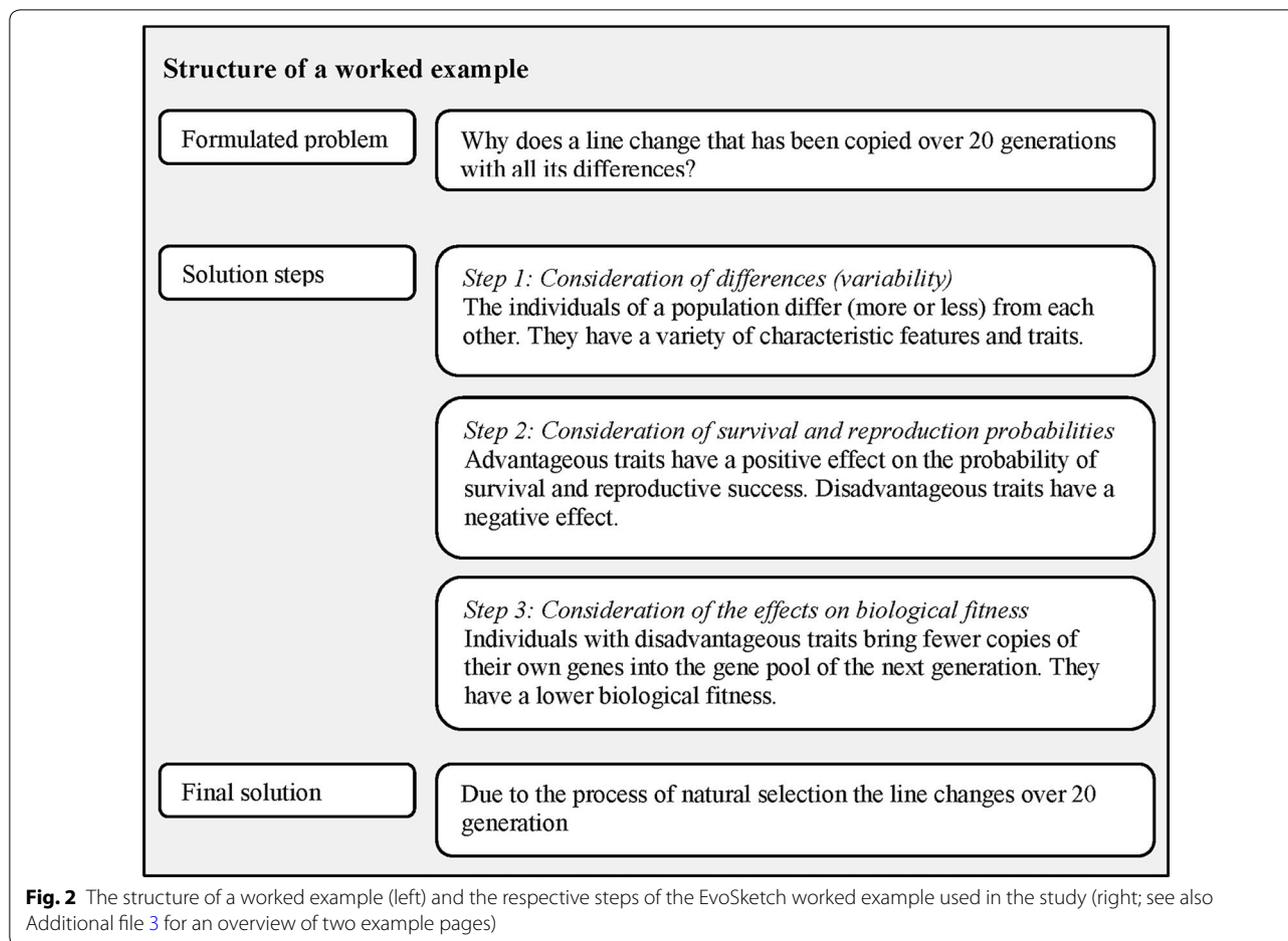
selective pressure (indicated by the distance from the red point in their simulation). Learners in this group did not receive additional instructional support but had to solve on their own how the basic information of randomness and probability in evolution (i.e., texts in EvoSketch worksheet) and EvoSketch simulations were connected to each other.

Simulation-based intervention with interpretative support (hereafter, sim-interpret)

This learning group was identical to the simulation intervention, except that learners were provided interpretative support in the form of a worked example on the roles of randomness and probability in evolution before starting to work with the EvoSketch simulations. Clark et al. (2011) described a worked example as “a step-by-step demonstration of how to perform a task or how to solve a problem” (p. 190; see also Atkinson et al. 2000). Thus, worked examples can help novices (i.e., non-experts) to understand how a formulated problem can be solved through introducing the formulated problem, the relevant solution steps, and the final solution (Renkl 2005; see also Fig. 2). We used a German worked example created by Neubrand et al. (2016) with revised and supplementary sections added in efforts to increase the focus on randomness and probability aspects, and to establish helpful connections to EvoSketch simulations. Concerning the threshold concepts (i.e., randomness and probability) and EvoSketch simulations, our worked example starts with explaining the conditional factors of evolution through natural selection (i.e., origin of variation, individual variation, heredity, and differential reproduction and survival; see also Fig. 2) with explanations connected to EvoSketch simulations (see Additional file 3 for an overview of two example pages). This followed a worked example of the peppered moths' evolution. By reading the worked examples, learners already received the information of how to connect threshold concepts and evolutionary concepts in the context of the EvoSketch software as well as in a biological context, before beginning EvoSketch simulations.

Simulation-based intervention with reflective support (hereafter, sim-reflect)

The last group of learners, the sim-reflect intervention group, also worked through the mentioned EvoSketch worksheet (including EvoSketch simulations). However, in contrast to the simulation and sim-interpret groups, learners received additional reflective support in the form of reflective questions after each task while working with EvoSketch simulations (e.g., “Explain the role of randomness or random processes in the line's evolution.”).



These learners have to describe and interpret their own simulation outcomes with respect to the two threshold concepts in question.

Participants

The sample consisted of 14 classes from nine comprehensive schools (“Gemeinschaftsschulen”) in northern Germany. In total, 269 tenth grade students aged between 14 and 18 years ($M = 15.6$ years, $SD = 0.6$ years; 47.19% female) participated in the study. Students of each class were randomly assigned to one of the four intervention groups: text ($n = 43$), simulation ($n = 70$), sim-interpret ($n = 79$), and sim-reflect ($n = 77$). The study was conducted during regular science lessons between November 2016 and March 2017. All students were informed that participation was voluntary and that their results would not affect their final grades. Students had received no formal instruction on evolutionary theory before. Nevertheless, we assume that they had some fragmentary knowledge on topics related to aspects of evolutionary theory (e.g., genetics), although evolutionary theory is not specifically

included in the German curriculum before the tenth grade (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany 2005).

Instruments

The instruments we used to study the effectiveness of the interventions and potentially influential variables are outlined below, while additional descriptions of the test instruments and item fit values are available as Additional file 4.

Randomness and probability test in the context of evolution (RaProEvo)

RaProEvo is a test instrument designed to measure students’ conceptual knowledge of randomness and probability in evolutionary contexts (Fiedler et al. 2017). It comprises 21 items (16 multiple-choice, three open response and two matching items) that focus on five aspects in which randomness and probability play important roles: the origin of variation, accidental death, random phenomena, the process of natural selection, and

the probability of events. Fiedler et al. (2017) reported that instrument validation was originally performed using expert rating, psychometric analyses of university student responses using item response theory (Rasch modeling), and criterion-related validity measures, and that the test had satisfactory reliability. The items of the German version are scored dichotomously, and we used a reduced set of 19 items (excluded the two matching items). The internal consistency (reliability) measured by Kuder-Richardson 20 (KR-20) for the data presented in this study was moderate ranging from 0.44 and 0.63.

Conceptual inventory of natural selection (CINS)

The CINS is a diagnostic test designed to assess students' understanding of evolution through natural selection (Anderson et al. 2002). It consists of 20 multiple-choice questions that focus on common misconceptions pertaining to 10 key conceptual aspects of natural selection, variation, and speciation. The inventory is structured so that each of the 10 concepts is assessed once in items 1–10 (CINS-A) and once again in items 11–20 (CINS-B). The original test instrument was verified by independent content experts (i.e., face validity evidence), student interviews, and statistical analyses based on classical test theory with satisfactory reliability (Anderson et al. 2002). A German translation of the CINS was prepared for a previous study with university students (Großschedl et al. 2014, 2018). The authors used the reverse translation method to generate a valid translation of the target instrument (e.g., Berry 1989; Su and Parham 2002). The results indicated that the translated test instrument generated reliable and valid inferences of university students' evolutionary knowledge. We used the translated CINS-A and CINS-B sets of items in the pretests and posttests, respectively, to minimize the influence of pretest on post-test scores and students' fatigue by reading the same items. All items are dichotomously scored, and KR-20 ranged from 0.23 and 0.40, suggesting that effects in this study may be somewhat misestimated due to lower-than-desired reliability.

General biological content knowledge test (GBCK)

The German GBCK test was designed to measure tenth grade students' existing content knowledge of biological topics included in up to tenth grade curricula (such as genetics or plant and animal ecology), and consists of 19 dichotomously scored items (16 multiple-choice items, two matching items, and one open response item; Neubrand et al. 2016). We used the GBCK to control for differences in students' existing prior knowledge of the subject and to test if this knowledge is related to the learning of randomness, probability, and evolution. The

results obtained with our students indicate that the test has an internal consistency (KR-20) of 0.36, lower than the level (0.51) reported by Neubrand et al. (2016) in applications with other samples of tenth grade German students.

Students' general language proficiency (C-test)

C-tests are designed to measure students' general language proficiency (Eckes and Grotjahn 2006), which may affect their performance in other diagnostic test instruments (Härtig et al. 2015). Therefore, we assessed our students' general German language proficiency using C-tests based on two German texts, each including 20 words with missing letters (Wockenfuß and Raatz 2006). Since learners' ability to read items or texts and produce answers is highly relevant in a study such as this, the responses were screened for both orthographical and grammatical errors. The students' answers were dichotomously scored, and KR-20 reliability of the test was found to be 0.78.

Perceived cognitive load (PCL)

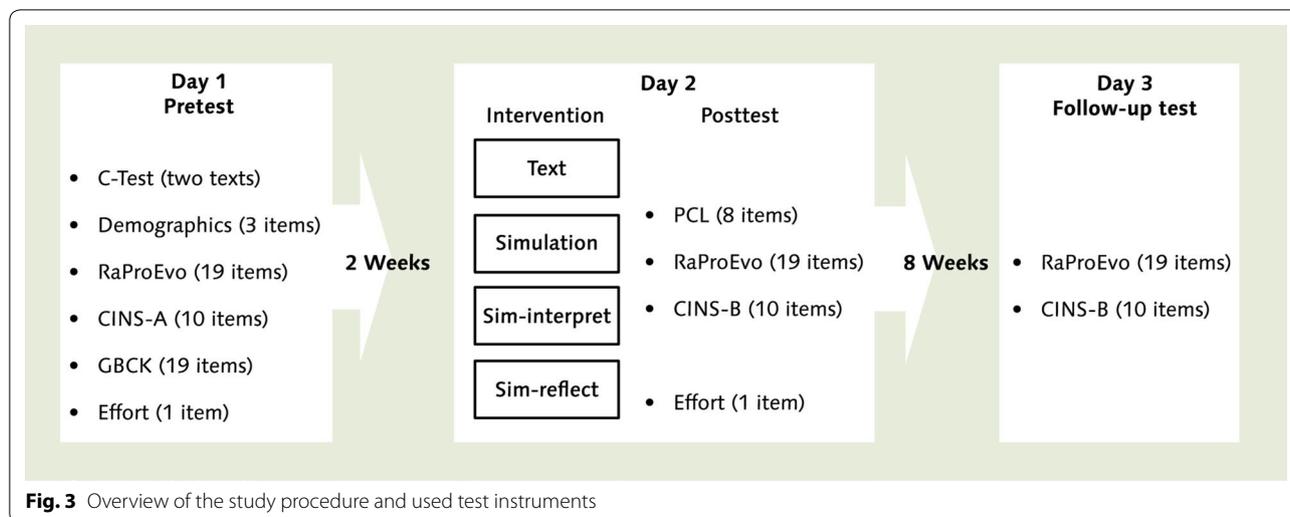
Cognitive load can affect learning (Sweller 1994), but it can be reduced by providing instructional support for learning with simulations (Leutner 1993). Therefore, students' PCL during the intervention was assessed using an adapted 5-point rating scale instrument (Urhahne 2002) consisting of eight items that allow differentiation of participants' PCL with a Cronbach's Alpha of 0.87.

Self-reported test-taking effort (effort)

Scores obtained by takers of any test are likely to depend on the effort they expend while taking it (Wise and Kong 2005). Thus, students' self-reported test-taking effort was appraised on one 10-point scale item (Organization for Economic Co-operation and Development [OECD] 2010), after they completed both the pretests and posttests.

Procedure

Prior to the intervention (day 1), every student took pretests consisting of the targeted randomness, probability and evolutionary knowledge tests (RaProEvo and CINS-A), and instruments designed to capture information on the control variables: general biological content knowledge (GBCK), language proficiency (C-test), self-reported test-taking effort (effort), and demographic data (i.e., age, sex, and biology grade; see also Fig. 3). Roughly 2 weeks later (day 2: intervention day) every student of each intervention group worked alone through their own EvoSketch worksheet on a single laptop. Laptops were all of the same types and provided by the Leibniz



Institute for Science and Mathematics Education (IPN) at Kiel University. Students of all intervention groups had 45 min to complete their worksheet tasks (intervention). On average, learners spent 30 min ($SD=7$ min; range 13–52 min) completing their tasks. Immediately after completing their worksheet, students took posttests consisting of the knowledge tests (RaProEvo and CINS-B) and items asking about their effort and PCL during the learning process. Roughly eight school weeks later (day 3), students took follow-up tests consisting of the targeted knowledge tests (RaProEvo and CINS-B). The study was conducted by the first author, with support from a university student who set-up and removed the laptops on the second day. All instruments and materials applied in this study were in German language.

Statistical analysis

The unequal sample size of the different groups might create problems in terms of homogeneity of variance across groups. Therefore, we performed Levene’s test to see if our groups have roughly the same variance on the investigated variables. Depending on the test result, we compared groups based on either Hochberg’s GT2 post hoc tests (if Leven’s test $p > 0.05$) or Games-Howell post hoc tests (if Levene’s test $p < 0.05$; Field 2018).

We analyzed the CINS-B and RaProEvo responses with generalized linear mixed models (GLMM) featuring a logistic link function, crossed random effects for participants and items, and an additional random effect for class (Baayen et al. 2008). The random effects for participants and items were included to account for differences in participants’ general ability and items’ general difficulty, respectively. The random effect for class also controlled for possible discrepancies in average ability

between classes. To uncover systematic effects of the experimental conditions on the development of students’ knowledge, dummy-coded variables for intervention, assessment, and their interaction were incorporated as fixed effects in the models. The text group served as the reference category for intervention, while posttest and pretest, respectively, served as the reference category for the CINS-B and RaProEvo assessments. This approach ensured simultaneous generalization of significant effects to new samples of both participants and items (Raaijmakers et al. 1999). As a measure of effect size (i.e., an expression of how much the respective method is better than the alternative one; Furukawa and Leucht 2011), we computed Cohens’ d . Cohens’ d provides information on the magnitude of the effect relative to the standard deviation (Cohen 1988). For instance, an effect of $d=0.25$ would mean that the difference is one quarter of a standard deviation. The greater the value of Cohens’ d , the greater the effect. Cohen (1988) also suggested a rule of thumb for interpreting the results with a small effect starting at 0.2, a medium effect starting at 0.5, and a large effect starting at 0.8. We used the lme4-package (Bates et al. 2011) for the statistical computing environment R 3.0.0 (R Core Team 2013) for all these statistical analyses.

Results

Baseline equivalence

At first, we conducted one-way analyses of variance (ANOVAs) to detect possible significant differences between intervention groups in pretest performance (i.e., CINS-A and RaProEvo scores) or the control variables (demographic variables and either C-test or GBCK scores). Values of these variables for each of the groups are listed in Table 1. Levene’s test of each variable showed

Table 1 Control variables and pretest performance scores (means with standard deviations; minimum and maximum in parentheses)

	Intervention				Total ^a (n = 244)
	Text (n = 39)	Simulation (n = 66)	Sim-interpret (n = 70)	Sim-reflect (n = 69)	
Age	15.39 ± 0.55 (15–17)	15.62 ± 0.74 (14–18)	15.57 ± 0.61 (15–17)	15.65 ± 0.66 (15–17)	15.58 ± 0.65 (14–18)
Biology grade ^b	2.51 ± 0.73 (1–4)	2.61 ± 0.91 (1–5)	2.77 ± 0.75 (1–5)	2.77 ± 0.75 (1–5)	2.65 ± 0.85 (1–5)
C-test	13.05 ± 2.88 (8–20)	11.58 ± 3.34 (2–18)	12.90 ± 2.78 (7–17)	12.74 ± 2.77 (7–18)	12.52 ± 2.99 (2–20)
GBCK	6.41 ± 2.61 (2–11)	5.86 ± 2.31 (0–12)	6.23 ± 2.31 (1–12)	5.90 ± 2.10 (2–10)	6.07 ± 2.30 (0–12)
RaProEvo	10.08 ± 2.51 (3–15)	9.29 ± 2.92 (3–18)	9.93 ± 2.57 (3–17)	9.57 ± 2.78 (3–18)	9.68 ± 2.72 (3–18)
CINS-A	3.51 ± 1.68 (0–7)	3.45 ± 1.84 (0–10)	3.90 ± 1.87 (0–9)	3.84 ± 1.78 (1–9)	3.70 ± 1.81 (0–10)

GBCK general biological content knowledge, C-test general language proficiency, RaProEvo conceptual knowledge of randomness and probability in evolutionary context, CINS-A conceptual inventory of natural selection

^a The total sample is smaller than N = 269 because 25 participants were absent during the pretests

^b Grades are measured within the possible range of 1 (good performance) and 6 (poor performance)

that the four groups had statistically equivalent variances ($F_s < 1.96$, $p_s > 0.121$), and ANOVA results indicated that the groups do not significantly differ in any of the relevant variables (Welch's $F_s < 2.66$, $p_s > 0.051$). Thus, the random assignment of learners to the four intervention groups caused no apparent bias in terms of any of these variables.

Intervention effects on CINS-B and RaProEvo scores

CINS-B

As already stated, students' responses to the CINS-B items were analyzed with a generalized linear mixed model featuring a logistic link function, crossed random effects for participants and items, and a random effect for class. Dummy-coded variables for intervention (with text as the reference category) and assessment (with posttest

Table 2 Intervention effects on CINS-B and RaProEvo scores

Effects	CINS-B			RaProEvo		
	b	SE	Cohen's d	b	SE	Cohen's d
Fixed						
Intercept	- 1.16***	0.22	0.65	0.10	0.26	0.05
Simulation	0.14	0.15	0.08	- 0.21	0.15	0.12
Sim-interpret	- 0.01	0.15	0.00	- 0.01	0.15	0.01
Sim-reflect	- 0.02	0.15	0.01	- 0.14	0.15	0.08
Post	-	-	-	- 0.09	0.11	0.05
Follow-up	0.25	0.16	0.14	- 0.04	0.12	0.02
CINS-A	0.12***	0.02	0.06	-	-	-
Simulation*post	-	-	-	0.04	0.15	0.02
Sim-interpret*post	-	-	-	- 0.04	0.14	0.02
Sim-reflect*post	-	-	-	0.09	0.14	0.05
Simulation*follow-up	- 0.25	0.20	0.06	0.34*	0.15	0.19
Sim-interpret*follow-up	- 0.09	0.19	0.14	0.09	0.15	0.05
Sim-reflect*follow-up	- 0.01	0.20	0.05	0.16	0.15	0.09
		Var	SD		Var	SD
Random						
Participants _(intercept)		0.08	0.28		0.30	0.55
Items _(intercept)		0.28	0.52		0.94	0.97
Class _(intercept)		0.02	0.13		0.05	0.23

CINS-A/B pretest/posttest of the conceptual inventory of natural selection, RaProEvo pretest of conceptual knowledge of randomness and probability in evolutionary context

* $p < 0.05$, *** $p < 0.001$

as the reference category) were included as fixed effects. Moreover, a fixed effect for students' CINS-A pretest performance was included as a covariate. A main effect for CINS-A ($b=0.12$, $SE=0.02$, $p<0.001$, $d=0.06$) was detected, but no other significant fixed effects (Table 2). No significant differences were detected, at posttest or follow-up, between intervention groups in understanding of evolution through natural selection. The inclusion of the GBCK results, C-test scores, and time-on-task data as further covariates did not alter this pattern of results.

RaProEvo

Students' RaProEvo performance was explored with a similar model, but with pretest as the reference category for assessment. No significant main effects of intervention were detected, indicating that there were no substantial differences between intervention groups in conceptual knowledge of randomness and probability in evolutionary context at the outset of the study (Table 2). Similarly, there was no significant general improvement in students' performance across assessments. However, a significant interaction revealed that students in the simulation group outperformed students in the text group at follow-up, $b=0.34$, $SE=0.15$, $p=0.024$, $d=0.19$. Incorporation of the GBCK results, C-test scores, and time-on-task data as covariates did not change this pattern of results.

Time-on-task

Differences in the time learners spend on tasks in their interventions may influence learners' knowledge acquisition. Levene's test indicated statistically inequivalent variances of the four groups regarding time-on-task, $F(3, 265)=3.93$, $p=0.009$. In addition, one-way ANOVA indicated a significant effect of intervention on time-on-task: Welch's $F(3, 132.69)=26.74$, $p<0.001$. Games-Howell post hoc tests revealed that learners in the text group ($M=30$ min, $SD=6$ min, $n=43$) worked significantly longer than learners in the simulation intervention group ($M=26$ min, $SD=5$ min, $n=70$, $p=0.006$, $d=0.74$). Furthermore, simulation intervention learners spent significantly less time with the material than the sim-interpret ($M=33$ min, $SD=7$ min, $n=79$, $p<0.001$, $d=1.14$) and sim-reflect intervention ($M=34$ min, $SD=7$ min, $n=77$, $p<0.001$, $d=1.31$) learners.

Perceived cognitive load (PCL)

Since cognitive load can influence learners' knowledge acquisition (Sweller 1994), the students' PCL was measured directly after each intervention. Levene's test showed that the four groups had statistically equivalent variances on PCL, $F(3, 265)=0.66$, $p=0.576$, while one-way ANOVA indicated significant differences

between intervention groups: Welch's $F(3, 134.55)=5.40$, $p=0.002$. Hochberg's GT2 post hoc tests showed that average PCL was higher in the text intervention group ($M=1.31$, $SD=0.53$, $n=43$) than in the simulation ($M=0.92$, $SD=0.63$, $n=70$, $p=0.004$, $d=0.66$) and sim-reflect intervention ($M=0.95$, $SD=0.58$, $n=77$, $p=0.008$, $d=0.64$) groups. However, no significant differences in this respect between the other pairs of interventions were detected (in all remaining cases, $p>0.05$; PCL sim-interpret: $M=1.08$, $SD=0.60$, $n=79$).

Self-reported test-taking effort

Test performance may also depend on the test-taking effort, as low effort is likely to result in test scores underrepresenting learners' true level of knowledge (Wise and Kong 2005). We applied repeated-measures ANOVA to investigate differences between intervention groups in self-reported test-taking effort. Levene's test showed that the four groups had statistically equivalent variances on pretest effort, $F(3, 229)=1.75$, $p=0.158$, but inequivalent variances on posttest effort, $F(3, 229)=4.08$, $p=0.008$. The results of the repeated-measures ANOVA showed a significant main effect of self-reported test-taking effort: $F(1, 229)=30.86$, $p<0.001$. Repeated contrasts also revealed that learners self-reportedly spent significantly more effort in the pretests ($M=7.03$, $SD=1.84$) than in the posttests ($M=6.19$, $SD=2.07$, $n=233$; $p<0.001$, $d=0.40$). Nevertheless, no significant main effect of group or interaction effect between group and effort was detected: $F(3, 229)=1.49$, $p=0.218$, and $F(3, 229)=0.46$, $p=0.710$, respectively.

In addition, Spearman's correlation coefficients were calculated to assess relationships between effort (pre- and posttest) and test performance (RaProEvo and CINS-A/B scores). A significant positive association was detected between pretest effort and RaProEvo pretest performance ($r_s=0.14$, $p=0.031$, $n=236$). Significant positive relationships were also found between posttest effort and both RaProEvo and CINS-B posttest scores ($r_s=0.18$, $p=0.007$, $n=240$, and $r_s=0.14$, $p=0.028$, $n=240$, respectively).

Discussion

The main aim of this study was to assess the effectiveness of EvoSketch simulations for improving students' knowledge about randomness and probability in evolutionary contexts, and their evolutionary knowledge. Since instructional support may reportedly improve the effectiveness of simulations, and EvoSketch Worksheets provide experimental support, we also examined and compared effects of additional interpretative and reflective support (worked example and reflective questions,

respectively) on self-directed learning with EvoSketch simulations.

We found that the overall mean posttest scores were lower, but mean follow-up test scores were higher than pretest scores. Concerning RaProEvo learning gains from the pretest to the follow-up test (not posttest), findings indicate that learners in the simulation intervention group (but not those in the simulation with additional self-directed support groups) acquired more knowledge than text-based learners. However, this positive effect was very small (i.e., Cohens' d of 0.19), which means that the difference between the simulation group and the text group was only one-fifth of a standard deviation. Expressed in other words (i.e., number needed to treat; Furukawa and Leucht 2011), it would mean that if 100 participants worked with EvoSketch (without additional self-directed support), only six more of them would have a greater RaProEvo score compared to students who only received the text material. Concerning the CINS scores, we could not find differences between the intervention groups. In contrast, we detected significant differences between intervention groups in both time spent on the material and PCL. Learners in the simulation groups with additional support (sim-interpret and sim-reflect) worked significantly longer on their tasks than learners in the simulation group (without additional support). Still, these groups did not differ in PCL. Students in the text group spent an intermediate amount of time on their worksheets but reported a significantly higher PCL than students of the simulation and sim-reflect groups.

The capacity of humans' working memory is limited and learning is likely to be hindered when tasks require too much cognitive load (Chandler and Sweller 1991; Paas and Sweller 2014; Sweller 1988). Thus, too much (new) information that is not aligned with the learner's prior knowledge, as well as inadequately designed learning material, can result in a high load of the working memory, which is detrimental for the learning process (de Jong 2010; Kirschner et al. 2006). The high PCL of the text group could have resulted from aspects of the intervention material. These students were not only supported with the two introductory texts of the EvoSketch Worksheet, but also received a text as Powerpoint presentation, to which three questions regarding topics covered in this text and two questions regarding evolution in a broader sense had to be answered. These students had to understand the concepts of randomness and probability in evolution based on the information given in the text only. In other words, they had no supporting simulation that visualized these concepts in connection with the evolutionary concept of variation or the relevance of random processes (i.e., mutations) for the probabilistic process of natural selection. Consequently, text-based

learners had to build up a simulation of these processes in their mind. This may have caused them to perceive a higher cognitive load than students of the other intervention groups.

Based on the performance tests, our participating secondary school students had on average a low (i.e., CINS) to medium (i.e., RaProEvo) test score, which can be interpreted as low to medium prior knowledge of the focal topics. This is potentially problematic as learners may be overwhelmed by the high amounts of abstract information conveyed in the simulations (Rutten et al. 2012; Wouters and van Oostendorp 2013). The slight improvements in delayed knowledge acquisition of simulation groups, relative to the text-based learners, may indicate that EvoSketch is probably not too abstract (nonrealistic) for fostering learners' knowledge about randomness and probability, but it does not seem to foster broad evolutionary knowledge in just one school lesson.

The limited duration of the learning session (the intervention time was roughly 45 min) may have affected several of the performance results, particularly CINS scores. There are a few studies indicating that learning evolutionary concepts in very short time (i.e., one or two lessons or hours) can result in higher knowledge (e.g., Beardsley et al. 2012; Bohlin 2017; Lee et al. 2017; Yamanoi and Iwasaki 2015). However, evolution education research also shows that the theory of evolution presents severe problems to learners, which have not been effectively solved by teaching strategies applied to date (e.g., Kampourakis and Zogza 2008; Rosengren et al. 2012). Introducing abstract, counter-intuitive concepts (i.e., randomness) in addition to these problems (particularly in a brief intervention) may partly explain the lack of learning gains directly after the intervention and the generally weak between-intervention differences in students' learning.

Moreover, additional instructional support in either the interpretative or the reflective forms did not lead to improvements in the performance (i.e., RaProEvo and CINS) of simulation-based learners relative to text group learners. One explanation might be that students perceived the additional material similar to normal textbook work (i.e., reading the worked example or answering the reflective questions on paper), which might have lowered the effect of self-directed learning with a simulation. Maybe the results could have been different if the additional supports were integrated as computer-based exercises.

Another factor for our lack of learning gains may be tracked back to the expertise reversal effect (Kalyuga et al. 2003). The effect explains why some instructional support may be highly effective for learners with low knowledge, while losing its effectiveness and resulting

in negative learning consequences for high knowledge learners, and vice versa (e.g., Kalyuga et al. 2003; Scott and Schwartz 2007). For instance, additional instructional support (e.g., worked example) may have been trivial for students with high prior knowledge (high pretest scores; see also maximum values in Table 1) to understand how threshold concepts and evolutionary concepts are connected in EvoSketch simulations, and even might have caused negative learning results for these students.

At last, the high amount of additional information provided in these interventions could have overwhelmed the students, deterred learners with low interest, and—in turn—reduced their motivation (Amabile et al. 1994; Pintrich and Schrauben 1992). Participants did not receive any credit for their test performance, and their results did not influence their final grade. Thus, their inherent learning motivation was likely correlated with motivation to address the large amount of material (e.g., worked example, large numbers of test items), thereby introducing a substantial random behavioral response factor in the posttest results (e.g., Meijer 2003). Accordingly, results of correlation analyses indicated a significant positive correlation between self-reported test-taking effort and posttest scores. Moreover, students reported significantly lower effort in the posttests than in the pretests, but no differences were detected among the groups, which may explain the lack of learning gains directly after the intervention. Since we did not ask the students a third time for their self-directed test-taking effort, we cannot clarify the connection of test performance and effort in the follow-up testing.

Limitations

This study is the first experimental approach to (1) examine the relationships of threshold concepts (i.e., randomness and probability) and evolution knowledge, and (2) foster the understanding of these concepts by use of an abstract visualization (i.e., EvoSketch). Visualizing random and probabilistic processes through EvoSketch seems to have a very small positive effect on students' conceptual knowledge of randomness and probability in evolutionary contexts, although it is unclear if this is due to higher understanding or rather an aspect of variability in the sample.

Nevertheless, our statistical analyses are limited by the lower than desired reliability of the knowledge test instruments (RaProEvo, CINS, and GBCK). The GBCK test's reliability was not expected to be high because it covers a large range of biological topics (Neubrand et al. 2016), but its internal consistency was unsatisfactorily low. The internal consistency of the CINS instrument was similarly low, possibly because the tenth

grade students had not received formal instruction on evolutionary theory before the intervention and may have been overstrained by the complexity of the presented items. In contrast, the internal consistency of the RaProEvo instrument was higher, but still not satisfactory. In addition, some of the test instruments (i.e., RaProEvo or CINS) were originally developed and validated to measure post-secondary students' knowledge of the respective context. However, at least the RaProEvo seems to be applicable to be used with secondary school students based on the range of item difficulty (see Additional file 4).

Another factor that potentially has affected the results is the limited duration of the learning session. We worked with tenth grade school students in several participating schools. The respective class teachers could decide on their own if they wanted their class (or classes) to participate or not. Since the research was performed during regular school lessons (depending on the particular school a lesson was between 45 and 60 min), we were unable to extend the research to more than the respective three days (i.e., five lessons including the time of the tests). Concerning this, the timeframe for the intervention was highly restricted. Nevertheless, studies indicated that even short learning periods could improve student's knowledge acquisition (e.g., Bohlin 2017; Eckhardt et al. 2013; Yamanoi and Iwasaki 2015). The timeframe should have been fine for investigating the effectiveness of EvoSketch simulations and the respective self-directed instructional support.

At last, in this study, we only focused on the effect of self-directed learning and did not incorporate teachers' support. Although teachers can be powerful for students' learning (e.g., Hattie 2009), students' performance is also likely to be influenced by teacher's professional knowledge (e.g., Mahler et al. 2017; Sadler et al. 2013). Our first step was to examine if EvoSketch can be effective on its own (without any teachers' support), but including teacher's support in further studies (e.g., in form of class discussions or one-by-one support) may be helpful for students' understanding of abstract threshold concepts, particularly directly after working with EvoSketch.

Implications and future research

Our findings may be useful for further research on how to visualize randomness and probability in evolution and for implementing EvoSketch in school sessions. Adequate knowledge of evolutionary concepts, and particularly related abstract concepts such as randomness and probability, is essential for students to critically address numerous issues associated with their environment and everyday life.

Thus, when using EvoSketch in the classroom, we recommend increasing the intervention timeframe to incorporate interventions on several days or weeks to foster students' understanding of randomness and probability in the context of evolution. Moreover, working with EvoSketch on a variety of real evolutionary examples (e.g., resistance in bacteria, peppered moth evolution), may help novices to realize that the same processes are relevant in different organisms to evolve, and are based on the same evolutionary principles. Studies of Nehm and Ridgway (2011), and Kampourakis and Zogza (2009) indicate that novices' explanations are often based on concrete surface features (e.g., running speed of cheetah) and have multiple explanatory models, while experts' explanations are based directly on the main domain principles (e.g., natural selection). Therefore, using an underlying framework (e.g., EvoSketch) and explaining different real case studies based hereupon might help to develop a coherent understanding of the respective concepts across case studies.

Additionally, deep learning is often more strongly supported by small group learning than individual learning (Dori and Belcher 2005; Springer et al. 1999). EvoSketch could be used in group settings with each individual working initially on their own and subsequently discuss observations in the group. Such discussions could also be extended to class discussions with the teacher. Learners with little prior knowledge could receive additional instructional support through worked examples or reflective prompts.

This study was only a first step to examine how the understanding of abstract concepts may be fostered by the use of visualizations. In future studies, we intend confirming whether learners apply a better understanding when (1) using an abstracted example in contrast to using a real example, and (2) students received additional teacher support. Moreover, we also want to gather more qualitative data on how students actually use EvoSketch simulations and how this changes their experience by using eye-tracking methods, think-aloud protocols, and interviews afterwards. We hope to increase the research in evolution education by focusing on understanding underlying abstract concepts such as randomness and probability.

Conclusions

We developed the simulation software EvoSketch to allow learners to explore—in a non-realistic way—random and probabilistic phenomena associated with the process of natural selection. Although the simulation group received higher knowledge gains in the follow-up test than the comparison text group, the effect size was

very small. Moreover, the additional self-directed learning supports (i.e., worked example or reflective questions) did not seem to improve students' knowledge. In fact, there was no immediate learning gain directly after the interventions. Our suggestion is that if students were able to work on EvoSketch for several rounds (i.e., going through the 20 generations more than once), and playing with the (new) knowledge they have (e.g., drawing known inaccurate lines while the fitness value influences the selection process), they would probably gain a more intuitive understanding of the random and probabilistic processes. Additionally, incorporate teacher's support may increase students' learning.

Additional files

Additional file 1. EvoSketch worksheet.

Additional file 2. Text intervention worksheet tasks.

Additional file 3. Two example pages of the worked example used in sim-interpret intervention.

Additional file 4. Table S4. Test instrument descriptions and item fit values.

Abbreviations

ANOVA: analysis of variance; CINS: conceptual inventory of natural selection; C-Test: students' general language proficiency; GBCK: general biology content knowledge test; PCL: perceived cognitive load; RaProEvo: randomness and probability test in the context of evolution; Text: text-based intervention; Simulation: simulation-based intervention; Sim-interpret: simulation-based intervention with interpretative support; Sim-reflect: simulation-based intervention with reflective support.

Authors' contributions

The objective and the rationale of the study was developed by UH (one of the project's principal investigators). With support from UH and JG, DF designed the study, collected the data, and analyzed them with support from ST. DF wrote the first draft of the manuscript, that was developed further and edited by DF, UH, JG, and ST. All authors read and approved the final manuscript.

Author details

¹ Department of Biology Education, Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University, Olshausenstraße 62, 24118 Kiel, Germany. ² Institute of Educational Sciences, Kiel University, Olshausenstraße 75, 24098 Kiel, Germany. ³ Department of Educational Research, Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University, Olshausenstraße 62, 24118 Kiel, Germany. ⁴ Department of Science and Mathematics Education, Institute for Biology Education, University of Cologne, Herbert-Lewin-Straße 10, 50931 Cologne, Germany.

Acknowledgements

We thank the EvoVis team, particularly Andreas Görrenson, for developing the simulation software. Special thanks are also due to the participating teachers for providing access to classes, the school students who participated in the study, and the university students Denise Bock and Charlotte Hölterhoff for their help with the laptops. We are also very grateful to John Blackwell for language review. Parts of this article (in a pre-peer reviewed manuscript version) are part of the publication-based doctoral thesis of the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Relevant data (in summarized form) and material are available upon request to the first author.

Ethics approval and consent to participate

All student participants were provided with an informed consent form for their parents as most were under the age of 18. Parents had to sign and return the form to their child's teacher if they agree to their child's participation. All student participants were instructed to write a predefined code on their worksheets and paper-based tests, so that data was already de-identified during collection. Research plan, procedure, and instruments were reviewed and approved by the Ministry for School and Vocational Training of Schleswig-Holstein, Germany (approved October 11, 2016).

Funding

The study reported here was part of the Swedish-German cooperation project "EvoVis: Challenging Threshold Concepts in Life Science—enhancing understanding of evolution by visualization", supported by the Swedish Research Council (Grant No. VR 2012:5344, LT). The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 June 2018 Accepted: 22 November 2018

Published online: 28 November 2018

References

- Ainsworth S, VanLabeke N. Multiple forms of dynamic representation. *Learn Instr.* 2004;14(3):241–55.
- Amabile TM, Hill KG, Hennessey BA, Tighe EM. The work preference inventory: assessing intrinsic and extrinsic motivational orientations. *J Pers Soc Psychol.* 1994;66(5):950–67.
- Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. *J Res Sci Teach.* 2002;39(10):952–78.
- Andrews TM, Price R, Mead L, McElhinny T, Thanukos A, Perez K, et al. Biology undergraduates' misconceptions about genetic drift. *CBE Life Sci Educ.* 2012;11(3):248–59.
- Atkinson RK, Derry SJ, Renkl A, Wortham D. Learning from examples: instructional principles from the worked examples research. *Rev Educ Res.* 2000;70(2):181–214.
- Baalmann W, Frerichs V, Weitzel H, Gropengießer H, Kattmann U. Schülervorstellungen zu Prozessen der Anpassung—Ergebnisse einer Interviewstudie im Rahmen der Didaktischen Rekonstruktion (Students conceptions of adaptation processes—results of an interview study within the framework of didactic reconstruction). *ZfDN.* 2004;10(1):7–28.
- Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. *J Mem Lang.* 2008;59(4):390–412.
- Bates DM, Maechler M, Bolker B. lme4: linear mixed-effects models using Eigen and Eigen. R package version 0.999999-2; 2011.
- BBC & Open University, producers. Bang goes the theory—evolving lines (video clip). <http://www.bbc.co.uk/programmes/p00wvffs>. 2011. No longer available.
- Beardsley PM, Bloom MV, Wise SB. Challenges and opportunities for teaching and designing effective K-12 evolution curricula. In: Rosengren KS, Brem SK, Evans EM, Sinatra GM, editors. *Evolution challenges: integrating research and practice in teaching and learning about evolution*. New York: Oxford University Press; 2012. p. 287–310.
- Beggrow EP, Nehm RH. Students' mental models of evolutionary causation: natural selection and genetic drift. *Evol Educ Outreach.* 2012;5(3):429–44.
- Berry JW. Introduction to methodology. In: Triandis H, Berry JW, editors. *Handbook of cross-cultural psychology*, vol. 2. Boston: Allyn & Bacon; 1989. p. 1–28.
- Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. *J Res Sci Teach.* 1990;27(5):415–27.
- Bohlin G. *Evolving germs—antibiotic resistance and natural selection in education and public communication*. Linköping: LiU-Tryck; 2017.
- Bohlin G, Göransson A, Höst GE, Tibell LA. A conceptual characterization of online videos explaining natural selection. *Sci Educ.* 2017;26(7–9):975–99.
- Buiatti M, Longo G. Randomness and multilevel interactions in biology. *Theory Biosci.* 2013;132(3):139–58.
- Chandler P, Sweller J. Cognitive load theory and the format of instruction. *Cogn Instr.* 1991;8(4):293–332.
- Chang H-Y. How to augment the learning impact of computer simulations? The designs and effects of interactivity and scaffolding. *Interact Learn Environ.* 2017;25(8):1083–97.
- Chang H-Y, Linn MC. Scaffolding learning from molecular visualizations. *J Res Sci Teach.* 2013;50(7):858–86.
- Clark RC, Nguyen F, Sweller J. *Efficiency in learning: evidence-based guidelines to manage cognitive load*. New York: Wiley; 2011.
- Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum; 1988.
- Concord Consortium. *Evolution readiness activities (simulation software)*. <https://learn.concord.org/evolution-readiness>. 2018. Accessed 13 June 2018.
- de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. *Instr Sci.* 2010;38(2):105–34.
- de Jong T, van Joolingen WR. Scientific discovery learning with computer simulations of conceptual domains. *Rev Educ Res.* 1998;68(2):179–201.
- Dori YJ, Belcher J. How does technology-enabled active learning affect undergraduate students' understanding of electromagnetism concepts? *J Learn Sci.* 2005;14(2):243–79.
- Dwyer FM. Adapting media attributes for effective learning. *Educ Technol.* 1976;16(8):7–13.
- Eckes T, Grotjahn R. A closer look at the construct validity of C-tests. *Lang Test.* 2006;23(3):290–325.
- Eckhardt M, Urhahne D, Conrad O, Harms U. How effective is instructional support for learning with computer simulations? *Instr Sci.* 2013;41(1):105–24.
- EvoSketch English. *EvoSketch English version (simulation application)*. <http://learninglabs.se/evolution/randomlineEN/>. 2018. Accessed 13 June 2018.
- EvoSketch German. *EvoSketch German version (simulation application)*. <http://learninglabs.se/evolution/evosketchde/>. 2018. Accessed 13 June 2018.
- Feller W. *An introduction to probability theory and its applications*. London: Wiley; 1968.
- Fiedler D, Tröbst S, Harms U. University students' conceptual knowledge of randomness and probability in the contexts of evolution and mathematics. *CBE Life Sci Educ.* 2017;16(2):ar38.
- Field A. *Discovering statistics using IBM SPSS statistics*. London: SAGE Publications Ltd.; 2018.
- Furukawa TA, Leucht S. How to obtain NNT from Cohen's d: comparison of two methods. *PLoS ONE.* 2011;6(4):e19070.
- Garvin-Doxas K, Klymkowsky MW. Understanding randomness and its impact on student learning: lessons learned from building the biology concept inventory (BCI). *CBE Life Sci Educ.* 2008;7(2):227–33.
- Gregory TR. Understanding natural selection: essential concepts and common misconceptions. *Evol Educ Outreach.* 2009;2(2):156–75.
- Gropengießer H. Theorie des erfahrungsbasierten Verstehens (The theory of embodied cognition). In: Krüger D, Vogt H, editors. *Theorien in der biologiedidaktischen Forschung*. Berlin: Springer; 2007. p. 105–16.
- Großschedl J, Konnemann C, Basel N. Pre-service biology teachers' acceptance of evolutionary theory and their preference for its teaching. *Evol Educ Outreach.* 2014;7(1):1–16.
- Großschedl J, Seredzsus F, Harms U. Angehende Biologielehrkräfte: Evolutionsbezogenes Wissen und Akzeptanz der Evolutionstheorie (Pre-service biology teachers: Knowledge and acceptance of evolutionary theory). *ZfDN.* 2018. <https://doi.org/10.1007/s40573-018-0072-0>. Accessed 6 Nov 2018.
- Härtig H, Heitmann P, Retelsdorf J. Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik—Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit (Analysis of tasks for the evaluation of the educational standards in physics—differentiation of literary skills and professionalism). *ZfE.* 2015;18(4):763–79.

- Hattie J. *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge; 2009.
- Heams T. Randomness in biology. *Math Struct Comp Sci*. 2014;24(03):e240308.
- Höffler TN, Leutner D. Instructional animation versus static pictures: a meta-analysis. *Learn Instr*. 2007;17(6):722–38.
- Horwitz P, McIntyre CA, Lord TL, O'Dwyer LM, Staudt C. Teaching 'evolution readiness' to fourth graders. *Evol Educ Outreach*. 2013;6(21):1–12.
- Kalyuga S, Ayres P, Chandler P, Sweller J. The expertise reversal effect. *Educ Psychol*. 2003;38(1):23–31.
- Kampourakis K, Zogza V. Students' intuitive explanations of the causes of homologies and adaptations. *Sci Educ*. 2008;17(1):27–47.
- Kampourakis K, Zogza V. Preliminary evolutionary explanations: a basic framework for conceptual change and explanatory coherence in evolution. *Sci Educ*. 2009;18(10):1313–40.
- Kirschner PA, Sweller J, Clark RE. Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ Psychol*. 2006;41(2):75–86.
- Kombartzky U, Ploetzner R, Schlag S, Metz B. Developing and evaluating a strategy for learning from animations. *Learn Instr*. 2010;20(5):424–33.
- Lakoff G. *Women, fire, and dangerous things*. Chicago: University of Chicago Press; 1987.
- Lakoff G, Johnson M. *Metaphors we live by*. Chicago: University of Chicago Press; 1980.
- Land R. There could be trouble ahead: using threshold concepts as a tool of analysis. *Int J Acad Dev*. 2011;16(2):175–8.
- Lazonder AW, Hagemans MG, de Jong T. Offering and discovering domain information in simulation-based inquiry learning. *Learn Instr*. 2010;20(6):511–20.
- Lee KM, Nicoll G, Brooks DW. A comparison of inquiry and worked example web-based instruction using physlets. *J Sci Educ Technol*. 2004;13(1):81–8.
- Lee TW, Grogan KE, Liepkalns JS. Making evolution stick: using sticky notes to teach the mechanisms of evolutionary change. *Evol Educ Outreach*. 2017;10(11):1–13.
- Leutner D. Guided discovery learning with computer-based simulation games: effects of adaptive and non-adaptive instructional support. *Learn Instr*. 1993;3(2):113–32.
- Mahler D, Großschedl J, Harms U. Using doubly latent multilevel analysis to elucidate relationships between science teachers' professional knowledge and students' performance. *Int J Sci Educ*. 2017;39(2):213–37.
- McDonald CV. Evaluating junior secondary science textbook usage in Australian schools. *Res Sci Educ*. 2016;46(4):481–509.
- Mead LS, Scott EC. Problem concepts in evolution part II: cause and chance. *Evol Educ Outreach*. 2010;3(2):261–4.
- Meijer RR. Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychol Methods*. 2003;8(1):72–87.
- Meyer JH, Land R. Threshold concepts and troublesome knowledge: linkages to ways of thinking and practising within the disciplines. In: Rust C, editor. *Improving student learning: theory and practice ten years on*. Oxford: Oxford Centre for Staff and Learning Development (OCSLD); 2003. p. 412–24.
- Meyer JH, Land R. Threshold concepts and troublesome knowledge: an introduction. In: Meyer JH, Land R, editors. *Overcoming barriers to student understanding: threshold concepts and troublesome knowledge*. Abingdon: Routledge; 2006. p. 3–18.
- Mix LJ, Masel J. Chance, purpose, and progress in evolution and christianity. *Evolution*. 2014;68(8):2441–51.
- Nehm RH, Ridgway J. What do experts and novices "see" in evolutionary problems? *Evol Educ Outreach*. 2011;4(4):666–79.
- Nehm RH, Schonfeld IS. Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. *J Res Sci Teach*. 2008;45(10):1131–60.
- Neubrand C, Borzikowsky C, Harms U. Adaptive prompts for learning evolution with worked examples—highlighting the students between the "Novices" and the "Experts" in a classroom. *Int J Environ Sci Educ*. 2016;11(14):6774–95.
- Opfer JE, Nehm RH, Ha M. Cognitive foundations for science assessment design: knowing what students know about evolution. *J Res Sci Teach*. 2012;49(6):744–77.
- Opitz ST, Neumann K, Bernholt S, Harms U. How do students understand energy in biology, chemistry, and physics? Development and validation of an assessment instrument. *Eurasia J Math Sci Technol Educ*. 2017;13(7):3019–42.
- Organisation for Economic Co-operation and Development, OECD. *PISA computer-based assessment of student skills in science*. Paris: OECD; 2010.
- Paas F, Sweller J. Implications of cognitive load theory for multimedia learning. In: Mayer RE, editor. *The Cambridge handbook of multimedia learning*. 2nd ed. New York: Cambridge University Press; 2014. p. 27–43.
- Pennock RT. Learning evolution and the nature of science using evolutionary computing and artificial life. *McGill J Educ*. 2007;42(2):211–24.
- Pennock RT. Avida-ED (educational application). <https://avida-ed.msu.edu/avida-ed-application/>. 2018. Accessed 13 June 2018.
- Pintrich PR, Schrauben B. Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In: Schunk DH, Meece JL, editors. *Student perceptions in the classroom*. Hillsdale: Erlbaum; 1992. p. 149–83.
- Plass JL, Milne C, Homer BD, Schwartz RN, Hayward EO, Jordan T, et al. Investigating the effectiveness of computer simulations for chemistry learning. *J Res Sci Teach*. 2012;49(3):394–419.
- Posner GJ, Strike KA, Hewson PW, Gertzog WA. Accommodation of a scientific conception: toward a theory of conceptual change. *Sci Educ*. 1982;66(2):211–27.
- Price F, Vaughn VG. *Evolve (simulation software)*. https://bioquest.org/BQLibrary/library_details.php?product_id=2367. 2010. Accessed 13 June 2018.
- Raaijmakers JG, Schrijnemakers JM, Gremmen F. How to deal with "the language-as-fixed-effect fallacy": common misconceptions and alternative solutions. *J Mem Lang*. 1999;41(3):416–26.
- R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2013.
- Rector MA, Nehm RH, Pearl D. Learning the language of evolution: lexical ambiguity and word meaning in student explanations. *Res Sci Educ*. 2013;43(3):1107–33.
- Reid DJ, Zhang J, Chen Q. Supporting scientific discovery learning in a simulation environment. *J Comput Assist Learn*. 2003;19(1):9–20.
- Renkl A. The worked-out examples principle in multimedia learning. In: Mayer RE, editor. *The Cambridge handbook of multimedia learning*. Cambridge: Cambridge University Press; 2005.
- Robson RL, Burns S. Gain in student understanding of the role of random variation in evolution following teaching intervention based on Luria–Delbruck experiment. *J Microbiol Biol Educ*. 2011;12(1):3–7.
- Rosengren KS, Brem SK, Evans EM, Sinatra GM. *Evolution challenges: Integrating research and practice in teaching and learning about evolution*. New York: Oxford University Press; 2012.
- Ross PM, Taylor CE, Hughes C, Whitaker N, Lutze-Mann L, Kofod M, et al. Threshold concepts in learning biology and evolution. *Biol Int*. 2010;47:47–52.
- Rutten N, van Joolingen WR, van der Veen JT. The learning effects of computer simulations in science education. *Comput Educ*. 2012;58(1):136–53.
- Sadler PM, Sonnert G, Coyle HP, Cook-Smith N, Miller JL. The influence of teachers' knowledge on student learning in middle school physical science classrooms. *Am Educ Res J*. 2013;50(5):1020–49.
- Scheiter K, Gerjets P, Huk T, Imhof B, Kammerer Y. The effects of realism in learning with dynamic visualizations. *Learn Instr*. 2009;19(6):481–94.
- Schmiemann P, Nehm RH, Tornabene RE. Assessment of genetics understanding. *Sci Educ*. 2017;26(10):1–31.
- Scott BM, Schwartz NH. Navigational spatial displays: the role of metacognition as cognitive load. *Learn Instr*. 2007;17(1):89–105.
- Secretariat of the standing conference of the ministers of education and cultural affairs of the Länder in the Federal Republic of Germany. *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10) (Beschluss der Kultusministerkonferenz vom 16.12.2004)* [National educational standards for the subject of biology concerning the 10th grade]. Munich, Germany: Luchterhand; 2005.
- Shtulman A. Qualitative differences between naive and scientific theories of evolution. *Cogn Psychol*. 2006;52(2):170–94.
- Sinatra GM, Brem SK, Evans EM. Changing minds? Implications of conceptual change for teaching and learning about biological evolution. *Evol Educ Outreach*. 2008;1(2):189–95.

- Soderberg P, Price F. An examination of problem-based teaching and learning in population genetics and evolution using EVOLVE, a computer simulation. *Int J Sci Educ*. 2003;25(1):35–55.
- Spanjers IA, Wouters P, van Gog T, van Merriënboer JJ. An expertise reversal effect of segmentation in learning from animated worked-out examples. *Comput Hum Behav*. 2011;27(1):46–52.
- Speth EB, Long TM, Pennock RT, Ebert-May D. Using Avida-ED for teaching and learning about evolution in undergraduate introductory biology courses. *Evol Educ Outreach*. 2009;2(3):415–28.
- Spindler L, Doherty J. Assessment of the teaching of evolution by natural selection through a hands-on simulation. *Teach Issues Exp Ecol*. 2009;6:1–20.
- Springer L, Stanne ME, Donovan SS. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. *Rev Educ Res*. 1999;69(1):21–51.
- Su C-T, Parham LD. Generating a valid questionnaire translation for cross-cultural use. *Am J Occup Ther*. 2002;56(5):581–5.
- Sweller J. Cognitive load during problem solving: effects on learning. *Cogn Sci*. 1988;12(2):257–85.
- Sweller J. Cognitive load theory, learning difficulty, and instructional design. *Learn Instr*. 1994;4(4):29–312.
- Taylor C. Threshold concepts in biology: do they fit the definition? In: Meyer JH, Land R, editors. *Overcoming barriers to student understanding: threshold concepts and troublesome knowledge*. Abington: Routledge; 2006. p. 87–99.
- Tibell LAE, Harms U. Biological principles and threshold concepts for understanding natural selection. *Sci Educ*. 2017;26(7–9):953–73.
- To C, Tenenbaum HR, Hogh H. Secondary school students' reasoning about evolution. *J Res Sci Teach*. 2017;54(2):247–73.
- Urhahne D. *Motivation und Verstehen (Motivation and understanding)*. Münster: Waxmann Verlag; 2002.
- Urhahne D, Harms U. Instruktionale Unterstützung beim Lernen mit Computersimulationen (Instructional support for learning with computer simulations). *Unterrichtswissenschaft*. 2006;34(4):358–77.
- van Berkum JJA, de Jong T. Instructional environments for simulations. *Educ Comput*. 1991;6(3):305–58.
- Veenman MV, Elshout JJ. Differential effects of instructional support on learning in simulation environments. *Instr Sci*. 1995;22(5):363–83.
- Venville G, Gribble SJ, Donovan J. An exploration of young children's understandings of genetics concepts from ontological and epistemological perspectives. *Sci Educ*. 2005;89(4):614–33.
- Wagner A. The role of randomness in Darwinian evolution. *Philos Sci*. 2012;79(1):95–119.
- Wang J-Y, Wu H-K, Hsu Y-S. Using mobile applications for learning: effects of simulation design, visual-motor integration, and spatial ability on high school students' conceptual understanding. *Comput Hum Behav*. 2017;66:103–13.
- Wernecke U, Schwanewedel J, Harms U. Metaphors describing energy transfer through ecosystems: helpful or misleading? *Sci Educ*. 2018;102(1):178–94.
- White BY, Frederiksen JR. Inquiry, modeling, and metacognition: making science accessible to all students. *Cogn Instr*. 1998;16(1):3–118.
- White RT, Gunstone RF. *Probing understanding*. London: Routledge; 1992.
- Wise SL, Kong XJ. Response time effort: a new measure of examinee motivation in computer-based tests. *Appl Meas Educ*. 2005;18(2):163–83.
- Wockenfuß V, Raatz U. Über den Zusammenhang zwischen Testleistung und Klassenstufe bei muttersprachlichen C-Tests (About the relationship between test performance and grade level in native language C-tests). *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, Empirical Research, Applications*. 2006:211–42.
- Wouters P, van Oostendorp H. A meta-analytic review of the role of instructional support in game-based learning. *Comput Educ*. 2013;60(1):412–25.
- Yaman M, Nerdel C, Bayrhuber H. The effects of instructional support and learner interests when learning using computer simulations. *Comput Educ*. 2008;51(4):1784–94.
- Yamanoi T, Iwasaki WM. Origami bird simulator: a teaching resource linking natural selection and speciation. *Evol Educ Outreach*. 2015;8(14):1–11.
- Yates TB, Marek EA. A study identifying biological evolution-related misconceptions held by prebiology high school students. *Creative Educ*. 2015;6(8):811–34.
- Zhang J, Chen Q, Sun Y, Reid DJ. Triple scheme of learning support design for scientific discovery learning based on computer simulation: experimental research. *J Comput Assist Learn*. 2004;20(4):269–82.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

