Conceptual Characterization of Threshold Concepts in Student Explanations of Evolution by Natural Selection and Effects of Item Context

Andreas Göransson,[†]* Daniel Orraryd,[‡] Daniela Fiedler,[§] and Lena A. E. Tibell[†]
[†]Department of Science and Technology and [‡]Department of Behavioural Sciences and Learning, Linköping University, SE-601 74 Norrköping, Sweden; [§]Department of Biology Education, IPN–Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany

ABSTRACT

Evolutionary theory explains a wide range of biological phenomena. Proper understanding of evolutionary mechanisms such as natural selection is therefore an essential goal for biology education. Unfortunately, natural selection has time and again proven difficult to teach and learn, and students' resulting understanding is often characterized by misconceptions. Previous research has often focused on the importance of certain key concepts such as variation, differential survival, and change in population. However, so-called threshold concepts (randomness, probability, spatial scale, and temporal scales) have also been suggested to be important for understanding of natural selection, but there is currently limited knowledge about how students use these concepts. We sought to address this lack of knowledge by collecting responses to three different natural selection items from 247 university students from Sweden and Germany. Content analysis (deductive and inductive coding) and subsequent statistical analysis of their responses showed that they overall use some spatial scale indicators, such as individuals and populations, but less often randomness or probability in their explanations. However, frequencies of use of threshold concepts were affected by the item context (e.g., the biological taxa and trait gain or loss). The results suggest that the impact of threshold concepts, especially randomness and probability, on natural selection understanding should be further explored.

INTRODUCTION

Decades of education research have yielded extensive knowledge about the teaching and learning of evolution, especially the process of natural selection. Hence, there is extensive knowledge of factors that influence understanding and assessment of evolution (Smith, 2009a,b). Nevertheless, evolution by natural selection remains conceptually challenging for learners, and many teaching and learning strategies tend to fail or have modest effects (Gregory, 2009; Smith, 2009a). In addition, test items have been developed for probing conceptual understanding of evolution (e.g., Bishop and Anderson, 1990; Anderson *et al.*, 2002; Nadelson and Southerland, 2009; Nehm *et al.*, 2012). However, they usually focus on learners' use of key concepts of natural selection such as origin of variation or differential reproduction (hereafter, key concepts; Nehm and Reilly, 2007). This may be inadequate, because recent research indicates that a set of "threshold" concepts could be as vital as key concepts for understanding natural selection (Ross *et al.*, 2010; Fiedler *et al.*, 2017, 2019; Tibell and Harms, 2017). Threshold concepts, originally proposed by Meyer and Land (2003), are concepts that, once understood, transform the way learners understand or interpret subject matter or their worldview. Without grasping a threshold concept, the learner cannot progress in understanding. Grasping threshold concepts entails a changed view, which is not necessarily the case with core concepts (i.e., essential conceptual building

Jennifer Momsen, *Monitoring Editor*
Submitted Mar 14, 2019; Revised Oct 11, 2019; Accepted Oct 22, 2019
CBE Life Sci Educ March 1, 2020 19:ar1
DOI:10.1187/cbe.19-03-0056

*Address correspondence to: Andreas Göransson (andreas.c.goransson@liu.se). ORCID: 0000-0001-5038-9630.

blocks). We propose, on the basis of previous research, that randomness, probability, and temporal and spatial scales are threshold concepts vital for understanding natural selection (Ross *et al.*, 2010; Fiedler *et al.*, 2017, 2019; Tibell and Harms, 2017).

At present, there is little evidence of whether and how these threshold concepts are used by students in explanations of natural selection and what factors influence how and when they are used. Because these concepts are important constituents of evolutionary theory in general and natural selection in particular, regardless of whether they are in fact threshold concepts, more knowledge of how they are understood and applied by learners is needed to improve evolution education research, assessment, and instructional practices. In this paper, we use results from an empirical study to explore how learners express and use the proposed threshold concepts of randomness, probability, temporal scale, and spatial scale in explanations of several examples of natural selection.

THEORETICAL BACKGROUND

Conceptual Foundations of Natural Selection

Natural selection was presented by Darwin in 1859 as a major mechanism explaining evolution and is regarded as central to biology (Dobzhansky, 1973). It is often formulated around three major principles: variation, selection, and inheritance (e.g., Lewontin, 1970; Tibell and Harms, 2017). According to Nehm and Ha (2011), variation includes the presence and causes of variation, selection comprises differential reproduction and/or survival, and inheritance is the inheritance of variation (Nehm and Ha, 2011). Some scholars in science education describe these principles (variation, selection, and inheritance) as core concepts of natural selection (Nehm and Ha, 2011; Opfer *et al.*, 2012). However, additional concepts are often used to explain natural selection, such as biotic potential (i.e., a population's growth capacity), selection pressure (imposed by limitations of resources and competition), and change of distribution/frequency of traits or genes within a population (e.g., Nehm and Reilly, 2007). Together, these additional concepts and the core concepts are often referred to as "key concepts" (Bishop and Anderson, 1990; Anderson *et al.*, 2002; Nehm and Reilly, 2007), a terminology we chose to adhere to in this paper. However, as will be argued in the coming sections, key concepts do not emphasize important aspects such as the random and stochastic rather than deterministic and directed nature of natural selection.

Students' Understanding of Key Concepts

It is well known that the theory of natural selection is challenging for learners, because it entails the integration and coordination of several concepts (Catley *et al.*, 2005). Therefore, it is not surprising that students' explanations are often simplistic and frequently reveal misconceptions and misunderstandings, for example, that organisms teleologically adapt to meet their needs (Gregory, 2009; Smith, 2009b). The scientific concepts used by students to explain natural selection have been extensively studied (Ferrari and Chi, 1998; Nehm and Schonfeld, 2008; Nehm and Ha, 2011; Nehm and Ridgway, 2011; Opfer *et al.*, 2012). Most commonly, learners seem to apply the concepts of differential survival and variation, while origin of variation and inheritance tend to be less frequently used. This indicates that concepts linked to genetics may be more challenging for learners due to the invisible nature of genes and/or the multitude of organizational levels needed to link genes to phenotypes. Also, genetics and evolution tend to be treated as separate topics in biology teaching and textbooks.

In addition, surface features of natural selection problems (e.g., the types of trait or organisms involved) affect how many key concepts learners use and how consistently they use those concepts in their explanations (Nehm and Ha, 2011; Nehm *et al.*, 2012; Federer *et al.*, 2015). Unfortunately, definitions of specific key concepts tend to vary in the literature, as do the scoring procedures applied in testing their use. For example, sometimes the presence and causes of variation are separated into different concepts—individual variation and origin of variation (Nehm and Ridgway, 2011)—and sometimes they are treated as a single concept—individual variation (Opfer *et al.*, 2012). Some studies have also considered other important aspects, such as the randomness of mutations (including both point mutations and transfer of larger DNA fragments), but the results provide little indication of how extensively students participating in those studies used them (Nehm and Ha, 2011). This lack of clarity in analyses of student responses is troubling, as key concepts are complex and require integration with other concepts. Without such integrated understanding of the key concepts, there are risks of learners developing oversimplified understandings of mechanisms of natural selection. For example, learners might think that variation is a response to needs of an organism or species. It is insufficient to understand merely that there is individual variation in a population. Rather, it is necessary to understand that variation is continuously generated by random processes at the molecular level to avoid misconceptions about need- or goal-based changes.

Key concepts thus do not capture some essential aspects of the natural selection mechanism, such as randomness in the above example. In the next section, we present additional aspects not captured by key concepts. Such aspects have not received systematic attention until quite recently, but the studies by Ross *et al.* (2010) and Tibell and Harms (2017) suggest that they are so-called threshold concepts. Considerably less is known about how learners understand and use these threshold concepts, as opposed to key concepts, in explanations of evolution and natural selection. We believe that students' incomplete understanding of threshold concepts such as randomness, probability, spatial scale, and temporal scale may be a potential source of learners' difficulties and misconceptions (Ross *et al.*, 2010; Tibell and Harms, 2017). If so, threshold concepts require more attention to improve our understanding of how students learn them and to enhance teaching and learning opportunities.

Threshold Concepts

Threshold concepts can be described as conceptual portals or gateways that, once passed, open up new and previously unavailable ways of thinking, leading to a transformed view of subject matter (Meyer and Land, 2003). In addition to being transformative, they are characterized by being integrative, irreversible, and potentially troublesome (Meyer and Land, 2003). For example, grasping that novel variation occurs

randomly can lead to a transformed view or change in conceptual understanding of natural selection from a directed or need-based to a probabilistic process. Other concepts suggested to be threshold concepts in natural selection, aside from randomness, are probability, spatial scales, and temporal scale (Ross *et al.*, 2010; Tibell and Harms, 2017). As noted in the following sections, these concepts are abstract, often not readily perceivable, and thus likely to be challenging for learners.

In this paper, we address four suggested threshold concepts: randomness, probability, and spatial and temporal scales. In the following sections, we describe these threshold concepts in more detail and provide arguments for their inclusion in the conceptual framework used in this study.

Randomness and Probability

Living organisms, or specific parts of them like eyes, may seem to have been purposefully "designed," but are the result of random variation and probabilistic selection. In general, randomness plays a crucial role in evolution and natural selection, especially in the origin of variation, which is a prerequisite for natural selection. Mutations occur (in principle) at random positions and can therefore affect different genes, giving rise to different phenotypes. In each individual within a generation, mutations occur at different positions, causing a population of individuals with various phenotypes. The outcome of mutations is random with respect to their selective value in a given environment, that is, the environment does not cause the specific mutations needed. While the presence of variation has been the focus of previous research, and is even argued to be a threshold concept by some (Ross *et al.*, 2010; Batzli *et al.*, 2016), we propose that the underlying difficulty lies not in understanding the presence of variation (which can be readily observed in many populations) but rather in understanding the underlying processes causing and changing the variation (e.g., random mutations and probabilistic selection). Thus, we propose that the real change in conception of natural selection occurs when learners grasp that novel variation arises due to random factors and not in response to an organism's need. Hence, we propose randomness to be a threshold concept, while variation will be regarded as key concept.

The selective value of a phenotype is dependent on the specific environmental factors present but also on random events such as mating and accidental death. Therefore, selection is probabilistic. After usually several to many generations, a new phenotype can occasionally become predominant in a population—this is the process of natural selection (i.e., differential survival and reproduction of individuals due to differences in phenotypes).

However, research has indicated that students have trouble incorporating aspects such as random mutations in their explanations of natural selection (Bishop and Anderson, 1990; Demastes *et al.*, 1995; Settlage, 2007; Garvin-Doxas and Klymkowsky, 2008; Bray Speth *et al.*, 2009; Fiedler *et al.*, 2017). It appears that many people have particular difficulties accepting that something so apparently well designed and efficient as an eye could result from processes with random components (Wallin, 2004; Garvin-Doxas and Klymkowsky, 2008; Larsson and Tibell, 2014). Instead, many students tend to use teleological explanations, such as changes occurring in response to selective pressure or need (Zohar and Ginossar, 1998). It has also been argued that learners are unfamiliar with causal models that include randomness and probability and that this underlies a number of challenges experienced in learning science concepts (Perkins and Grotzer, 2005). Unfortunately, many large-scale studies on understanding of natural selection have not addressed randomness or probability explicitly (Bishop and Anderson, 1990; Nehm and Reilly, 2007; Nehm and Schonfeld, 2008), and the key concepts framework does not include randomness and probability as important conceptual aspects of natural selection. Consequently, the community has incomplete knowledge of how and to what extent students include randomness and probability in their explanations of natural selection and whether test items designed to probe natural selection actually can capture these concepts. In addition, we lack knowledge of how understanding of such concepts is linked to the use of different item contexts. Hence, we integrated randomness and probability as two concepts in the conceptual framework of this study. More specifically, we added randomness in the origin of variation, differential survival (accidental death), and reproductive success (random mating).[1] Probability is connected to differential survival (probability of an organism surviving) and reproductive success (probability of having offspring).

Spatial Scale and Organizational Levels

Evolutionary processes occur in hierarchically organized biological systems spanning many magnitudes of spatial scale at molecular, genetic, protein, cellular, tissue, organ, organism, population, and species levels. These levels are often organized in three overarching categories: submicroscopic (biochemical), microscopic (cellular), and macroscopic (organismal; Tsui and Treagust, 2013). Consequently, many biological phenomena such as evolution involve mechanisms at various organizational levels, for example, genetic mutation in DNA molecules is a prerequisite for population- or species-level evolutionary change.

Viewing the key concepts through the lens of spatial scale and organizational levels reveals a number of crucial points. Perhaps the most crucial cross-level relations to understand are the causal relationships between genes, proteins, cells, and organisms' traits. Individual variation in a population arises through molecular-scale processes that cause variation in the genetic makeup of individuals within a population. Thus, while a novice might observe superficial similarities among individuals of a species and hence assume corresponding genetic similarity, there is always a range of genetic variation (not directly observable). The genetic variation in turn causes a variation in phenotype (often observable) when the genes and environment interact. Because all organisms have DNA and essentially the same basic machinery of replication, transcription, and translation of genetic information, all populations of all taxa are subject to the same fundamental evolutionary processes. For evolutionary change to occur within a population, the genetic composition must change. However, natural selection operates on phenotypes, and thus only indirectly on the genetic material. In fact, variation is present on the submicroscopic scale, in the form of molecular variation in DNA and proteins (for example);

[1]We are aware that many other processes of evolution involve randomness. In our framework, we included the most significant random processes for a basic understanding of natural selection.

the microscopic scale, as variation in cellular form and function; and simultaneously at the macroscopic scale as larger-scale phenotypic differences.

Accordingly, the ability to identify phenomena and reason about them and their effects across multiple levels of organization is suggested to be an essential but challenging skill to develop in biology (Wilensky and Resnick, 1999; Knippels, 2002; Mohan *et al.*, 2009; Elmesky, 2013; Tsui and Treagust, 2013). For example, while research-based learning progressions in genetics propose that even younger children should be able to learn reason across levels of organization (Elmesky, 2013), there are indications that even adults at the university level are having issues with this (e.g., Jördens *et al.*, 2016). There is also evidence about learners' abilities to work with scale from other domains such as geoscience. Cheek *et al.* (2017) performed an extensive review on student learning about spatial and temporal scales and found few studies in the domain of biology. In addition, the authors failed to find studies giving insight into when learners develop concepts of scale over the course of their education. However, there is evidence that younger students appear to have less sophistication in scales smaller than a person compared with experts, who seem to have developed more efficient categories for smaller scales. In addition, the number of scale categories, and the precision of these, increase with educational level. Numeracy and mathematical knowledge are also associated with better ability to grasp scale concepts (Cheek *et al.*, 2017). Thus, increased sophistication of scale conception seems to develop with age and education, but we know to date very little of when and how this occurs. In addition, Swarat *et al.* (2011) also raise concern about the effectiveness of current instruction for teaching scales.

This is concerning, because inadequate understanding and skill in working with organizational levels in the context of evolution and natural selection can be associated with misconceptions such as essentialism (i.e., focus on species level rather than intraspecific variation on the individual level) and teleology (i.e., focus on species or individuals, ignoring, for instance, the importance of random genetic mutation; Bishop and Anderson, 1990; Samarapungavan and Wiers, 1997). Interestingly, the general principle of genetic origin of variation is not used consistently across examples of natural selection differing in surface features (i.e., trait gain or loss and biological taxa; Nehm and Ridgway, 2011). This indicates that the use of genetic concepts is context bound to some extent.

The relevance of organizational levels for understanding natural selection is supported by studies showing that understanding increases and misconceptions diminish when organizational levels are explicitly addressed (e.g., Kampourakis and Zogza, 2008; Jördens *et al.*, 2016). In addition, focusing on the genetic level can aid transfer of ideas about natural selection from one context to another (Jördens *et al.*, 2016; Bohlin *et al.*, 2017b).

In summary, reasoning across organizational levels is a central skill for understanding natural selection and should be included in assessments. We integrated these aspects under the term "spatial scale" in the conceptual framework of the study presented here to capture how participating learners used and linked organizational levels across spatial scales. In the framework, we organize the levels as follows: submicro (molecular, genetic, protein), micro (cellular), macro (individual), and supermacro (population, species, and higher taxa; see Table 2 later in this article). We use the term "supermacro" to group higher organizational levels into a superordinate category, including entities beyond a single organism, such as populations and species.

Temporal Scales

Evolution includes processes that occur over timescales ranging from an extremely short time for mutations (submillisecond) to deep time (millions of years) for macroevolution of species and higher taxa. Analogous to spatial scales, some of the timescales relevant for evolution are far beyond direct human perception, and thus more challenging to conceptualize (Catley and Novick, 2009). Previous studies on educational aspects of time in the context of evolution have focused mostly on the issue of deep time and concluded that students have difficulties with both short and long timescales (Catley and Novick, 2009). This is not surprising, because humans tend to overestimate the length of short durations and underestimate those of longer durations (Lee *et al.*, 2010). This is concerning, because many of the most important evolutionary processes have very short timescales (e.g., mutations) or large timescales (e.g., repeated selection over many generations and speciation). In addition, the studies that have considered time have tended to focus on students' ability to place important macroevolutionary events in time, for example, the origin of life, nucleated cells, and photosynthesis. However, we propose that the main conceptual obstacle or threshold for learners is the relation between deep time and evolutionary mechanisms such as natural selection. For example, the probability of even minute evolutionary changes such as a single mutation is astoundingly low for a single reproductive event (e.g., 10^{-8} per replication). However, in sufficiently large populations and/or time frames, such a mutation becomes highly probable or almost inevitable. Thus, an important competence is the ability to consider processes with different timescales, which is related to the competence to work with large numbers and reason about proportional relationships (Cheek, 2012). To do so, learners must have the ability to translate large timeframes into numbers of generations and connect them with population numbers, mutation frequencies, and so forth. Thus, this is directly related to the ability to reason about evolutionary mechanisms such as natural selection and time (e.g., how unlikely events become probable with a large enough timeframe).

The continuously ongoing process of natural selection and evolution also seems prone to misconception. Many learners conceptualize natural selection as an event that ends when adaptation is "achieved," rather than a continuous process (Ferrari and Chi, 1998). Hence, they fail to distinguish evolutionary adaptation as a process of genetic change that occurs over many generations rather than individual adaptations that occur within one generation though nongenetic changes. Similar misconceptions have been identified in learners' descriptions of equilibrium processes in chemistry (Perkins and Grotzer, 2005).

In conclusion, several aspects of time can be problematic for learners but important for understanding natural selection. Unfortunately, students' understanding of time has often been studied in isolation to determine their conceptions of natural selection and has typically not been part of conceptual frameworks used to score learners' explanations of natural selection (see key concepts used in, e.g., Nehm and Reilly, 2007). Therefore, we integrated temporal scales into the conceptual

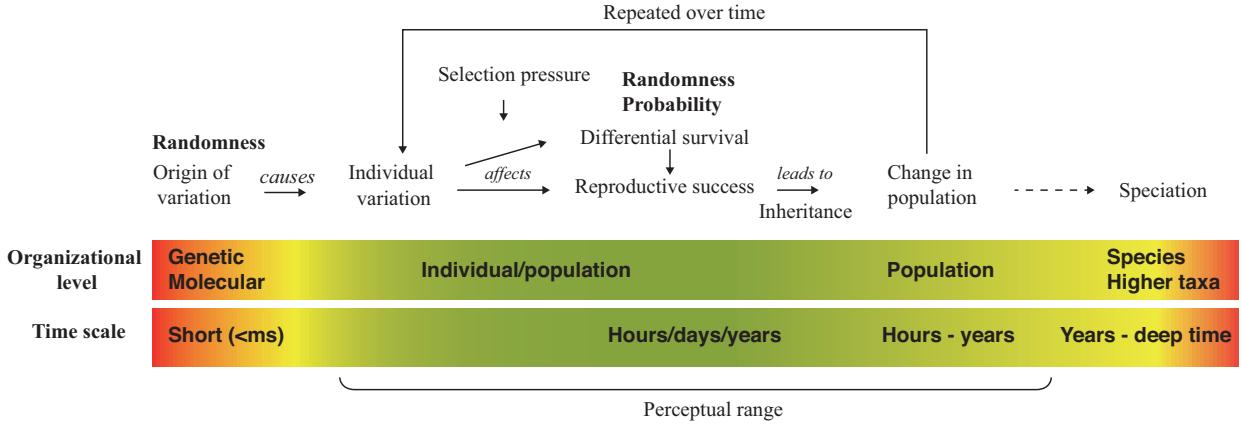

FIGURE 1. Schematic representation of the relation between the different key and threshold concepts. The process starts with the random generation of variation (origin of variation) by mutations. This occurs on the short timescale at an imperceptible scale level. The novel genetic variation this introduces in a population can manifest as individual phenotypic differences (individual variation). These differences, in combination with the selective pressure from the environment, can result in differential survival and reproduction. If the differences are inherited, this can give rise to a change in population. On the longer timescales, these changes can ultimately give rise to new species.

framework of our study to explore how students included time aspects in their explanations of natural selection.

An Integrated Framework of Key and Threshold Concepts

The previously described key and threshold concepts can be integrated into a conceptual framework (Figure 1), as suggested by Tibell and Harms (2017), wherein key concepts constitute the first dimension and threshold concepts the second dimension. In this study, we included the following key concepts: origin of variation, individual variation, inherited variation, differential survival, reproductive success, selection pressure (limited resources and competition), and change in population. The threshold concepts included were randomness, probability, spatial scale, and temporal scale. This proposed framework was used to analyze answers to items frequently used to assess students' understanding of natural selection mainly in terms of the proposed threshold concepts. However, it is known that the context of an assessment item (such as trait gain or loss and biological taxa) tends to influence the concepts elicited (Nehm and Ha, 2011). Accordingly, this third dimension should also be considered when proposing a conceptual framework for assessing natural selection.

Effects of Test Item Contexts

As already mentioned, the context of a test item tends to influence students' responses to it. An item's context comprises a number of features that may vary without affecting its basic problem structure, so many variants of items may potentially be structurally "isomorphic." Item features such as the type of organism and direction of evolution (trait gain or loss) involved are known to influence frequencies of both core concepts and misconceptions in students' responses (e.g., Nehm and Ha, 2011; Großschedl *et al.*, 2018). This likely applies to the less-studied threshold concepts as well. To gain a more nuanced understanding of students' threshold concept knowledge, it is therefore important to investigate whether the features of any items used affect the results and, if possible, relate the effects to dimensions of our framework. Therefore, we also addressed this possibility.

Aim and Research Questions

The aim of this study was to explore whether commonly used assessment items on natural selection also elicit threshold concepts in students' explanations of natural selection and to characterize the way students expressed those threshold concepts. We also aimed to explore whether surface features affect the use of threshold concepts.

The following research questions guided our study:

1. How do students apply and express threshold concepts in their written explanations of evolution by natural selection?
2. How consistent is the use of threshold concepts across examples of natural selection with differing surface features?
3. Can relations between items' surface features and the students' expression of threshold concepts be discerned?

MATERIAL AND METHODS

Data Collection

Data-Collection Instrument. To gauge learners' use of threshold concepts in explanations of natural selection, we chose to use open-response items for the following reasons. First, they provide more robust measures of students' knowledge than multiple-choice items, because recall of information rather than recognition is required (Opfer *et al.*, 2012). Second, they ask students to produce an explanation from the recalled information, inviting them to apply and integrate specific concepts into an explanation. Therefore, written answers provided by the students should reflect their own understanding, rather than discriminating between alternatives in a multiple-choice test. An available instrument for natural selection fulfilling these criteria is the Open Response Instrument (ORI; Bishop and Anderson, 1990; Nehm and Reilly, 2007), which also reportedly has excellent agreement with oral interviews in terms of key concepts (Nehm and Schonfeld, 2008) and better agreement with interviews in terms of alternative conceptions than the commonly employed CINS (Conceptual Inventory of Natural Selection) multiple-choice test (Anderson *et al.*, 2002). For these reasons, we used three items from the ORI:

1. Explain why some bacteria have evolved a resistance to antibiotics (that is, the antibiotics no longer kill the bacteria).
2. Cheetahs (large African cats) are able to run faster than 60 miles per hour when chasing prey. How would a biologist explain how the ability to run fast evolved in cheetahs, assuming their ancestors could run only 20 miles per hour?
3. Cave salamanders (amphibian animals) are blind (they have eyes that are not functional). How would a biologist explain how blind cave salamanders evolved from ancestors that could see?

We selected only a subset of the ORI items that seem to be representative examples of natural selection problems to explore potential variation in concept use across a variety of items while avoiding test fatigue. In addition, the other three of the original six ORI items were less relevant for our study, because they did not refer to a specific context but posed questions regarding the definition of natural selection or accelerating evolution without referring to a concrete evolutionary example.

The three items are all framed in an evolutionary context and are isomorphic in structure. Thus, they are expected to induce similar explanations from students with a good understanding of natural selection. However, the items differ in surface features such as biological taxa, type of trait, and gain or loss of trait (Table 1). The first item is different from the second and third regarding the type of organism (unicellular and prokaryotic organism versus multicellular and eukaryotic organism). The second and third items concern multicellular animals that probably are more familiar to learners. In addition, the familiarity with the trait type should be higher for running speed in cheetahs and sight in salamanders compared with drug resistance in bacteria, which is confined to subcellular components such as changes in proteins and enzymes. It is also worth noting that items 2 and 3 involve evolutionary developmental changes that affect morphological and metabolic features, thus increasing the complexity of a satisfactory scientific explanation. In addition, the scale of evolutionary change described in the items differs. Item 1 concerns microevolution, while items 2 and 3 are framed in a macroevolutionary context. Item 3 also differs from items 1 and 2 in that it deals with speciation.

Data-Collection Procedure. Data were collected by administering the test electronically to volunteers from universities in Sweden and Germany. The test-collection procedure was designed to be completely anonymous, and the participants were asked to provide as elaborate answers as possible. The items were administered in the same order (i.e., bacteria, cheetah, and salamander) and in the participants' native languages (i.e., Swedish or German). Thus, respondents answered in Swedish (within Sweden) or German (within Germany).

In Sweden, students attending introductory-level courses in biology or biochemistry were asked to participate in the online survey. The volunteers participated without any incentives for their participation.

In Germany, biology students from different universities were made aware of the online survey via the biology student council's home page (Spring 2016). All these respondents were given the opportunity to participate in a lottery for 10 vouchers, each worth €50 (approximately US$54 at the time of data collection). In Summer 2016, biology students from Kiel University were also asked to participate in an intervention study via course visits and postings at notice boards. Another 32 respondents volunteered (15.3% of this sample) and received €30 (approximately US$32 at the time of data collection) for their participation.

Sample Characteristics. Before any analyses, 10 German and four Swedish respondents were excluded from the sample due to low response behavior (i.e., only one item was answered), resulting in an overall sample size of 247 university students from Sweden and Germany, of whom 140 students were at an introductory level (i.e., 1–2 years at university/college), 60 students were at an advanced level (i.e., 3–4 years at university/college), and 47 students were at a graduate level (i.e., more than 5 years at university/college).

Swedish Sample. The Swedish sample included 38 university undergraduate students (average age = 23.7 years, SD = 2.25 years) from a southern Sweden university. Four of the students attended a primary-teacher education program and the other 34 attended various chemistry- or biology-oriented programs. Regarding evolution instruction, all students had been exposed to basic evolutionary theory (i.e., natural selection, micro- and macroevolution, and genetics) in their preceding upper-secondary education according to the Swedish curricula (Skolverket, 2011). In addition, the students attending the biology-oriented education program (11) had received an additional introductory course to biology covering evolutionary theory.

German Sample. The German sample of the study included 209 biology students from 21 German universities (average age = 23.0 years, SD = 3.3 years). Ninety-seven of the students were biology majors, of whom 71 attended undergraduate (leading to a bachelor's degree) and 26 graduate (leading to a masters' degree) courses. The other 112 were preservice biology teachers: 52 taking undergraduate courses, 42 taking graduate courses, and 18 taking basic foundation courses (leading to the first state exam). Regarding evolution (or evolutionary theory), biology majors and preservice biology teachers (depending on the university) are normally

TABLE 1. Comparison of the surface features of the three items used in our study

Item	Biological taxa	Trait gain/loss	Selective factor	Scale of change
1. Antibiotic resistance in bacteria	Bacteria (prokaryotic, unicellular)	Gain	Antibiotics	Microevolution Hours–days[a]
2. Cheetah running speed	Animal (eukaryotic, multicellular)	Gain	Running speed of prey	Macroevolution 10^6–10^7 years[a]
3. Loss of sight in cave salamanders	Animal (eukaryotic, multicellular)	Loss	Light and food (severely restricted)	Macroevolution 10^6 years[a]

[a]Typical time ranges.

TABLE 2. Analytical categories for the threshold concept spatial scale

Scale (perceptual) level	Definition	Biological organization level	Illustrative events or processes
Submicro	Molecular/biochemical level. No direct experience possible. Imagination or abstract/symbolic representation necessary. Indirect imaging techniques.	Molecule	Base pairing
		DNA	Mutation
			Translocation
			Cross-over
		Gene	Genetic variants
		Protein	Protein function
Micro	Cellular/subcellular level. Visible under light microscope. Outside directly perceptible range but within limits of light.	Cell	Cellular function
		Cell (single individual organism)[a]	
Macro	Biological structures or processes visible to the naked eye.	Individual organism	Individual fitness
Supermacro	Level of scaling beyond single organism. Sometimes beyond perceptual range (e.g., an entire species or large population are not directly observable). Abstractions and symbolic representations.	Population	Change in composition of population
			Intraspecies variation
		Species	Variation between populations
		Higher taxa	

[a]Because the individual and cellular levels coincide for unicellular organisms, we added this code to distinguish between mentions of cells as individual organisms and references to cellular functions or components.

exposed to the topics of 1) mechanisms of evolution, 2) micro- and macroevolution, 3) evolutionary theories, and 4) abiotic and biotic factors during their bachelor's programs (see Appendix A in the Supplemental Material). As evolution is also described as an organizing principle for the life sciences and explicitly stated as a learning goal in the German middle school standards (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005), both sets of students should ideally have a shared knowledge regarding evolutionary changes through natural selection.

Data Analysis

To address the first research question, we extracted variables by content analysis (Krippendorff, 2013). We used a directed content analysis approach with deductive use of theory (Hsieh and Shannon, 2005) and an inductive coding of the expression and context of threshold concepts (see section *Expression of Threshold Concepts* for additional information). Descriptive and inferential statistics were then used to establish measures that allowed us to answer research questions 2 and 3.

Coding of Variables. Our deductive coding scheme was used to identify sets of variables corresponding to the two dimensions, key concepts and threshold concepts (Tibell and Harms, 2017), of the conceptual framework described earlier (see Appendix B in the Supplemental Material for code descriptions and examples). The variables were operationalized as binary variables, coding concepts as present (1) or not present (0). To pilot the coding schema, three raters (A.G., D.O., and D.F.—two for each sample) independently analyzed the overlap of answers from both samples (Swedish: overlap of 43 of 110 answers;[2] German: overlap of 50 of 209 answers), which is higher than the recommended overlap of at least 10% of the total sample (Krippendorff, 2013). Reliability was checked by calculating Guilford's *G* (Holley and Guilford, 1964), which performs more consistently for variables with low occurrence than frequently used measures, such as Scott's π or Cohen's κ (Xu and Lorber, 2014). Variables lacking satisfactory reliability ($G > 0.7$) were discussed, and definitions in the final codebook were refined. The revisions were checked by recoding these variables within the entire sample. The final reliabilities can be found in Appendix C in the Supplemental Material.

The instances of threshold concepts found in the first coding round were categorized in a second inductive round of analysis. The inductive analysis was systematic, and a code memo was constructed and added to the codebook, acting as a constant comparison tool. Links between threshold concepts and key concepts were identified by considering the entire response and judging whether the threshold concept described any aspect of a key concept (e.g., variation occurs by random mutations links mutation [spatial scale: gene level] and randomness with origin of variation).

For randomness and probability, synonyms such as "chance" or "likely" were also included. The inductive analysis of randomness and probability was categorized according to the key concepts they were linked to (see Tables 4 and 5 later in this article).

For spatial and temporal scales, we also performed a more fine-grained analysis of the intervals used, that is, the time units and/or spatial scales (organizational levels) mentioned (see Table 2 and Table 6 later in this article). The organizational levels used in the analysis of spatial scale were derived from the literature. We grouped biological levels of organization largely following Johnstone (1991) with the addition of the microscopic level (Tsui and Treagust, 2013). However, this categorization system lacks a category for entities larger than the macro level, for example, populations, species, or higher taxonomic units, which mostly fall outside the perceptual range used to delineate the macro level. Therefore, we chose to make this explicit by introducing the supermacro level for these entities. This resulted in the categories shown in Table 2, which served

[2]The original Swedish sample consisted of 110 students, in total, including 72 upper secondary students. The upper secondary students were later dropped from the analysis, but reliability calculations were based on data from the entire sample.

as the analytical subcategories for the threshold concept spatial scale.

Because we considered linking between different organizational levels an important threshold for understanding natural selection, for example, that variation arises from processes on a lower scale level (submicro) than the outcome (macro and supermacro), we also analyzed how students linked different levels in their answers. We defined indications of a causal relation or mechanism between two levels as a link. For example,

> "A certain cheetah has once acquired a mutation [submicro/genetic] which caused this individual to easily run faster [macro/individual]." (SWE7)

In the quoted responses, coded concepts are indicated in square brackets, and parentheses indicate quoted participants in terms of sample (GT/GI for German or SWE for Swedish) and an assigned number.

Statistical Analysis

The coded variables were exported from MaxQDA 2018 to IBM SPSS 24 for further analysis. Cochran's *Q* was used to test for significant differences in proportions of participants using a specific concept across the three items (concerning bacteria, cheetahs, and salamanders). The test is suitable for comparing a dichotomous outcome variable in related samples, such as differences in pass/fail frequencies on different test items (Siegel and Castellan, 1988).

Cochran's *Q* test was applied for each concept (key and threshold) and Bonferroni correction was used to adjust for the number of comparisons within each sample (12 different variables), $\alpha = 0.05/12 = 0.004$. If a significant effect of item context was found, subsequent pairwise comparisons were performed (Dunn's post hoc test, nonparametric) using built-in alpha adjustment for multiple comparisons in SPSS.

We also analyzed the verbosity of the answers by calculating the average number of words, sentences, and sentence length (see Appendix B in the Supplemental Material).

RESULTS

Application of Key and Threshold Concepts

Our findings show that the frequency of the various threshold concepts present in student answers varied (see Figure 2). All four of the threshold concepts and all seven of the analyzed key concepts were found in the sample. Threshold concepts generally had a lower presence than key concepts in students' explanations (Figures 2 and 3). Moreover, there was no single item that generally elicited more key or threshold concepts (Figures 2 and 3). The threshold concepts randomness and probability were least frequently used. Approximately 25% of all students mentioned these concepts at least once across the three items (Table 3). Temporal scale was mentioned slightly more frequently (32%) and spatial scale most frequently (39%). However, "probability" roughly co-occurred with "individual variation," "differential survival," and "reproductive success," in accordance with our expectations.

The consistency of concept application across the items was generally low (Table 3, column 3). Further analysis of the consistency in use of the key and threshold concepts revealed significant between-item differences in frequencies for all concepts except the key concept change in population and threshold concept temporal scale (Figures 2 and 3). In addition, the item that elicited the most frequent use of a concept depended on the concept, and none of the concepts was mentioned by most of the students in responses to all three items. The most consistently used concepts were selection pressure and individual variation, while probability, randomness, and inheritance were the least consistently used (Table 3).

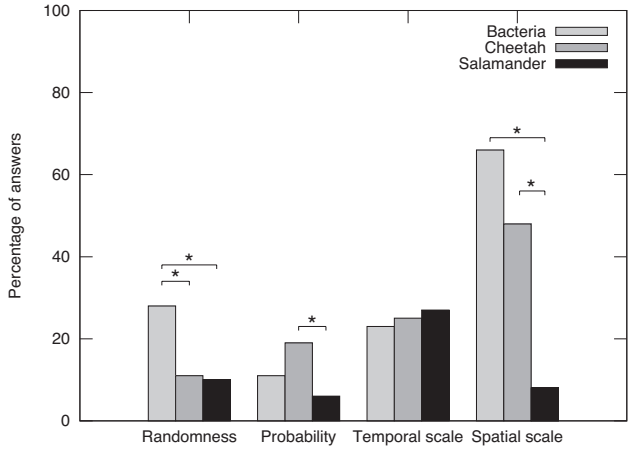

FIGURE 2. Frequencies of inclusion of the threshold concepts in responses of the 247 students to the three items. An asterisk (*) denotes a significant difference according to Dunn's post hoc test.

We also tested the possibility that between-item differences in length of the responses could explain the inconsistency of concept application. However, the length of the students' answers did not differ dramatically between items in terms of word count, number of sentences, or sentence length (see Appendix D in the Supplemental Material). Hence, the inconsistent application of concepts across the items does not appear to be an artifact linked to differences in verbosity in responses to the items.

Expression of Threshold Concepts

In the next step of the analysis, we compared items in which each threshold concept was used, mainly (as already mentioned) in relation to the key concepts. In the following sections, we present results of this analysis together with illustrative responses.

Randomness

Overall, the bacteria item elicited significantly more frequent (almost threefold) use of randomness than the other items. In responses to the bacteria and cheetah items, the dominant concept linked to randomness was mutation (see Table 4):

> "There occurred random mutations that enabled a cheetah to run faster than others [randomness, origin of variation, individual variation]." (GT003)

In responses to the salamander item, random mutations or random appearance of a trait was the most common concept. In addition, randomness was linked to genetic drift and death in a few cases (only in responses to the salamander item):

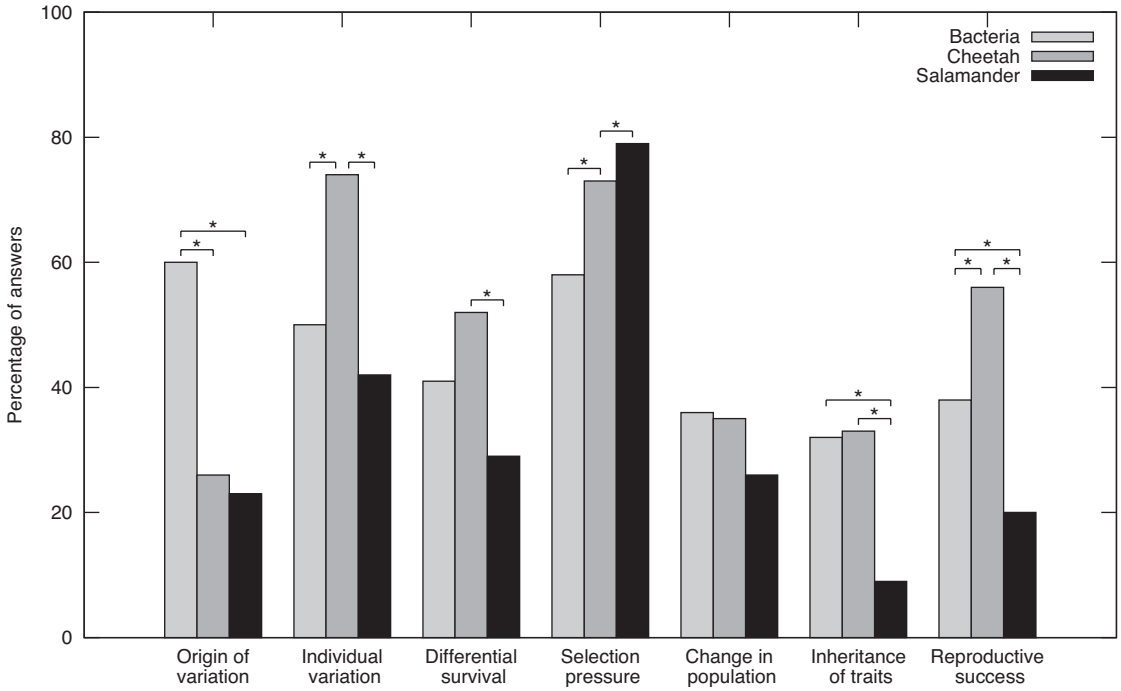

FIGURE 3. Frequencies of inclusion of the key concepts in responses of the 247 students to the three items. An asterisk (*) denotes a significant difference according to Dunn's post hoc test.

> "In their offspring, there were randomly some who got regressed/no eyes [randomness, individual variation, random appearance of trait]." (GT008)

> "A few of the seeing salamanders got blind by random mutations [randomness, origin of variation, individual variation, random mutation]." (GT127)

Probability

Overall, responses to the cheetah item contained the most mentions of probability (Table 5). In responses to all items, mention of probability mostly occurred in connection with survival, for example,

> "Cheetahs that due to mutation had the ability to run faster, like 60 km/h, could hunt prey more efficiently and had a higher probability of surviving [individual variation, origin of variation, differential survival, probability]." (GT024)

A few students also connected reproduction to probability, for example,

> "As individuals with a mutation for less developed eyes had an 'energy advantage.' These individuals therefore had a higher probability of reproducing [probability, reproductive success]." (GT034)

TABLE 3. Consistency of threshold and key concept application in students' responses

	Students using the concept in responses to: (categories are mutually exclusive)			Total (concept used at least once in responses to the three items)
	1 item	2 items	3 items	
Threshold concepts				
Randomness	24%	9%	2%	35%
Probability	25%	4%	1%	30%
Temporal scale	32%	16%	4%	52%
Spatial scale	39%	34%	5%	78%
Key concepts				
Origin of variation	36%	18%	13%	67%
Individual variation	26%	31%	26%	83%
Differential survival	33%	28%	11%	72%
Selection pressure	15%	32%	44%	91%
Change in population	33%	19%	9%	61%
Inheritance of traits	34%	16%	2%	52%
Reproductive success	30%	27%	10%	67%

TABLE 4. Frequencies of contextual links of randomness concepts in students' responses (percentages of the 247 students with numbers in parentheses)[a]

Randomness concepts	Bacteria	Cheetah	Salamander
Total randomness	28% (70)	11% (26)	10% (25)
Origin of variation			
Random mutation	20% (50)	7% (18)	4% (10)
Random appearance of trait	8% (19)	3% (8)	5% (13)
Change in population			
Random drift	0% (0)	0% (0)	1% (2)
Random death	0% (0)	0% (0)	0.4% (1)

[a]Note that codes are overlapping.

In addition, a small proportion of the students mentioned probability of mutations or novel traits, most frequently in responses to the bacteria item:

> "Due to the high reproduction numbers of bacteria the probability that they develop antibiotic resistance by mutations is relatively high [origin of variation, probability, probability of mutation]." (GT115)

Temporal Scale

We found that time was mentioned approximately equally frequently in responses to all three items. Generally, most uses of time were unspecific and connected to the idea that adaptation takes time (see Table 6):

> "The eyes have adapted to their life situation over time [temporal scale, unspecified time]." (SWE28)

The most frequently mentioned specific timescales were relative times in terms of generations, while there were few (or no) uses of absolute timescales, such as years or shorter scales:

> "The genes of these cheetahs propagated to the next generation. The cheetahs in this generation who were the fastest runners could transfer their genes to the next generation etc., which led to cheetahs running faster and faster because it is an advantage for survival [individual variation, inheritance, differential survival, temporal scale/generations]." (SWE55)

TABLE 5. Frequencies of contextual links of probability concepts in students' responses (percentages of the 247 students with numbers in parentheses)[a]

Probability concepts	Bacteria	Cheetah	Salamander
Total probability	11% (27)	19% (48)	6% (14)
Differential survival			
Survival probability	5% (12)	11% (27)	2% (6)
Chance of catching prey	N/A	6% (16)	0% (0)
Reproductive success			
Reproduction probability	2% (4)	5% (13)	2% (5)
Chance of providing for offspring	N/A	1% (2)	0% (0)
Origin of variation			
Probability of novel trait	2% (6)	1% (2)	1% (2)
Mutation probability	2% (6)	0% (0)	0% (0)
Inheritance of traits			
Probability of inheritance	1% (3)	2% (5)	1% (2)

[a]Note that codes are overlapping.

TABLE 6. Frequencies of contextual links of timescale concepts, expressed as percentages of temporal scale coding in students' responses (percentages of the 247 students with numbers in parentheses)

Temporal scale concepts	Bacteria	Cheetah	Salamander
Temporal scale total	23% (58)	25% (61)	27% (67)
Origin of variation			
Mutations over time	6% (14)	2% (4)	1% (3)
Selection pressure			
Selection duration	2% (4)	1% (3)	0% (0)
Change in population			
Accumulation of traits	0% (1)	1% (2)	0% (0)
Reproductive success			
Reproduction rate	4% (10)	3% (7)	1% (3)
Other			
Adaptation takes time or traits evolve over time	9% (22)	15% (36)	22% (55)
Temporal scale linking			
Generation time affects rate of evolution	0% (1)	0% (0)	0% (0)
Timescales			
Unspecified time			
Over time or within an unspecified time frame	11% (27)	14% (37)	18% (44)
Relative time			
Generation time	11% (28)	0% (0)	0% (0)
Generations	2% (6)	7% (17)	7% (18)
Absolute time			
Years	0% (0)	2% (5)	2% (5)
Days	0.4% (1)	0% (0)	0% (0)
Hours or shorter time	0% (0)	0% (0)	0% (0)

Spatial Scale

The population and individual levels of spatial scales were used in some responses to all items (Figure 4). The gene level was used mostly in responses to the bacteria and cheetah items, and less frequently in responses to the salamander item. An illustrative example is

> "Earlier, the ancestors could only run at 20 km/h, but a few could run faster due to a genetic change and therefore had an advantage because they could capture prey better and also escape enemies faster [spatial scale: gene, individual, population]." (GT004)

In contrast, references to the DNA, protein, and cell levels were (in principle) only found in the responses to the bacteria item:

> "By a genetic mutation, or more likely by uptake of a plasmid for antibiotic resistance (can degrade antibiotics/not

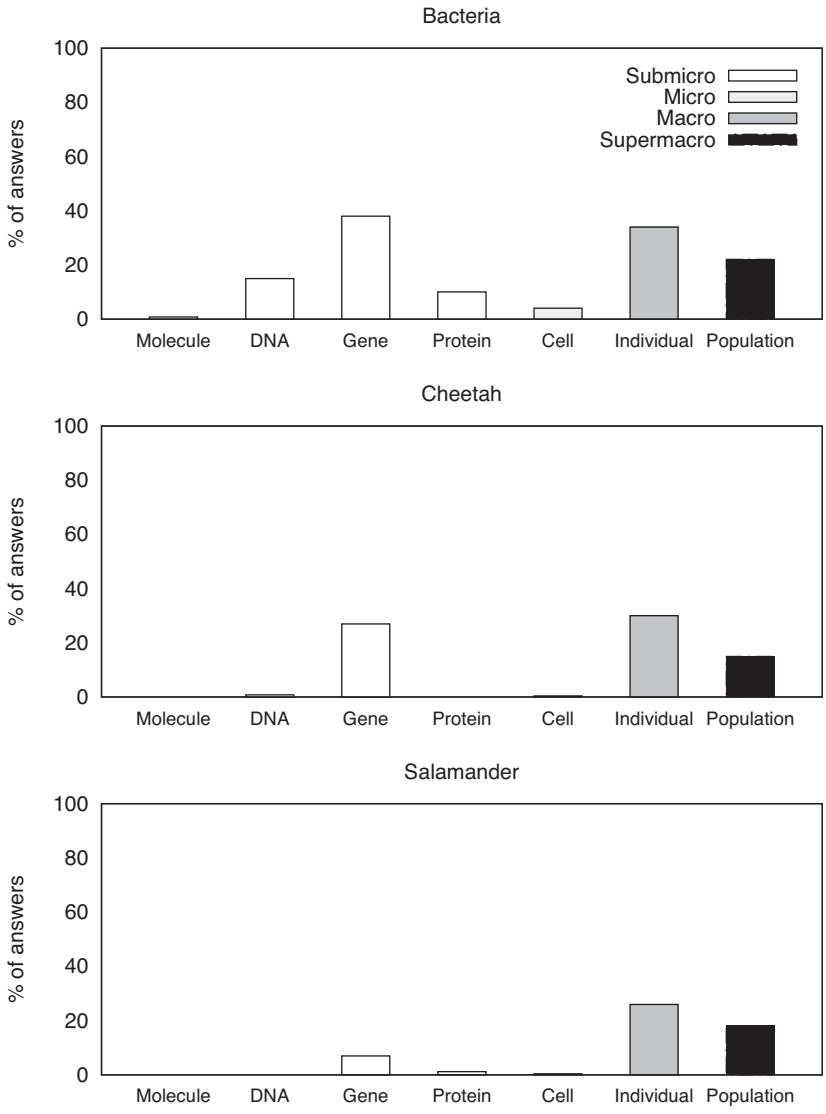

FIGURE 4. Frequencies (%) of mentions of major categories of organizational levels in the 247 students' responses to the three items.

attackable), the bacteria are not killed by antibiotics any longer. Mostly, antibiotics are not taken up anymore, or an enzyme can break it down, before the cell is damaged [origin of variation, spatial scale, submicro/DNA, submicro/gene transfer, submicro/protein, micro/cell, macro/individual, submicro → micro, micro → macro]." (GT054)

We also found large between-item differences in connections between the organizational levels (Figure 5). In responses to the bacteria item, we found examples of connections between all levels (the difference between micro and macro levels in bacterial contexts is defined in Table 2), but in responses to the cheetah and salamander items, there were no connections to the micro level. In responses to the cave salamander item, the most frequent connections were between the macro and supermacro levels. In total, the number of links in responses to the bacteria item was almost twice the number in responses to the cave salamander and cheetah items.

DISCUSSION

The items used in the study (i.e., bacteria, cheetah, and cave salamander) elicited all the probed threshold concepts to various degrees. The capacity of the instrument to elicit natural selection key concepts has been established by prior research, and the frequencies of key concepts in our sample were comparable to those in previous studies (Bishop and Anderson, 1990; Nehm and Reilly, 2007; Nehm and Schonfeld, 2008). In addition, the instrument has been cross-checked against the commonly used CINS test (Anderson *et al.*, 2002), showing good agreement in the diversity and magnitude of concepts (Nehm and Schonfeld, 2008). Thus, we conclude that our results concerning threshold concepts were obtained from typical undergraduate responses to natural selection items. In the following sections, we discuss the results and their implications.

How Do Students Express and Apply Key Concepts and Threshold Concepts in Written Explanations of Evolution by Natural Selection?

Overall, we found that 1) threshold concepts were relatively seldom used by students in their explanations compared with key concepts, and 2) students' use of both threshold and key concepts was sensitive to the items' context (Figures 2 and 3). The ways and situations in which the threshold concepts were expressed in our participants' responses are discussed in the following sections.

Randomness

Overall, a minority of the students used randomness in their explanations, a third or less in responses to each item (Figure 2). In addition, only 2% of the participants used randomness consistently in responses to all three items (Table 3). Randomness was associated with the genetic level in the explanations and was almost three times more frequent in responses to the bacteria item than the other two items (Figure 2). In responses to the bacteria and cheetah items, randomness was generally connected to mutations, but in responses to the salamander item, it was more frequently associated with appearance of novel traits and less frequently with mutations (Table 4). A few occurrences of randomness were linked to genetic drift or random death, and only in responses to the salamander item. Thus, most connections between randomness and natural selection were in the context of genetic-level events. Because the genetic level was mostly associated with the bacteria item, it appears to explain the relatively high occurrence of randomness in responses to this item.

Because variation is "the fuel" for natural selection and is ultimately dependent on random genetic variation, we expected randomness to occur mostly in association with the key concepts origin of variation or individual variation in the

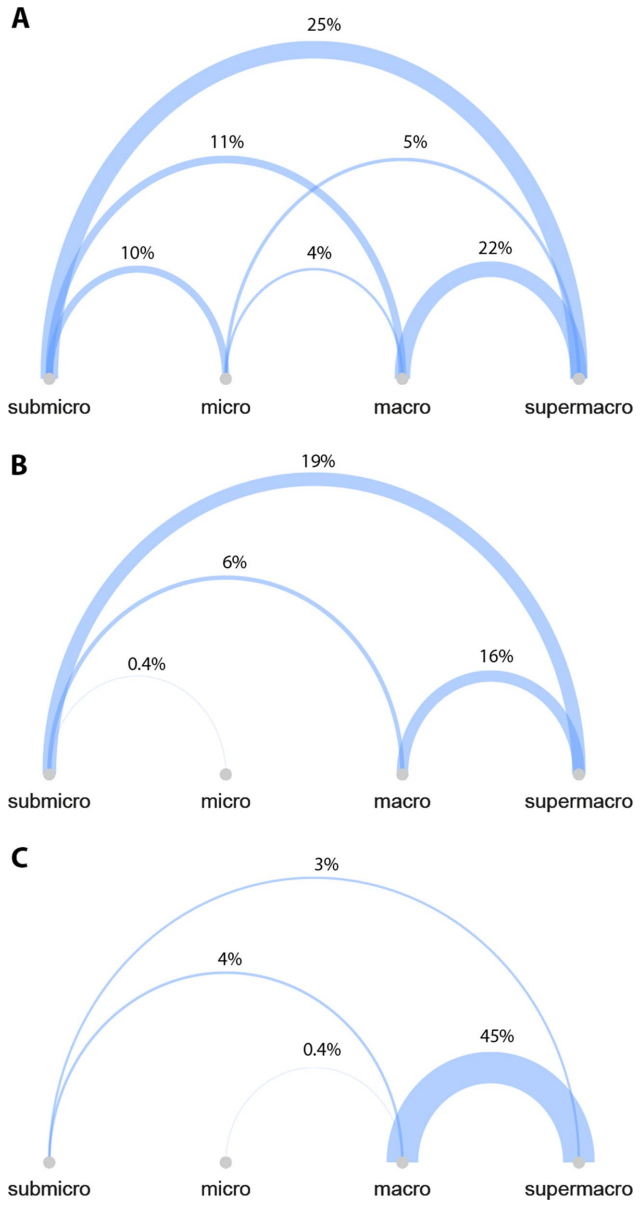

FIGURE 5. Percentages of answers with organizational-level links in responses to the bacteria (A), cheetah (B), and cave salamander (C) items. The thickness of a line indicates the number of links found.

explanations. However, far from all students who used origin of variation mentioned randomness (as shown by a comparison of Figures 2 and 3). This indicates that most of the students were either unaware of the role of randomness in novel variation or did not consider it a central concept to mention in an explanation of natural selection. Hence, it is not surprising that other studies often find the misconception that novel traits arise in response to need (Garvin-Doxas and Klymkowsky, 2008; Kampourakis and Zogza, 2008; Gregory, 2009).

Earlier studies, for example, Bishop and Anderson (1990) and Bizzo (1994), have indicated that random variation as well as probabilistic survival and reproduction seem especially difficult for learners to connect to natural selection. In addition, learners tend to prefer deterministic explanations over probabilistic explanations (Metz, 1998) and experience difficulties in using probabilistic rather than deterministic causation (Grotzer *et al.*, 2017). This certainly seems to apply to biological phenomena. For example, Garvin-Doxas and Klymkowsky (2008) found that learners tend to consider random processes inefficient and thus less tempting to use in biological explanations. Further, textbooks and educational videos rarely emphasize randomness and probability in evolution and natural selection (Aleixandre, 1994; Bohlin *et al.*, 2017a). Hence, it is not surprising that stochastic aspects were relatively scarce in the explanations. In addition, the close association between the genetic level and randomness in our participants' responses indicates that randomness may not be emphasized in instruction about evolution unless genetic-level phenomena are included. As understanding of randomness and probability is negatively associated with teleological misconceptions (Kampourakis and Zogza, 2008), it seems particularly important to address learners' understanding of randomness in evolution.

Probability

Probability occurred at substantially lower frequencies than any of the key concepts relevant to it (differential survival, reproductive success, and inheritance; Figure 2). In fact, only 1% of the students used probability consistently in responses to all three items (Table 3).

As expected, probability occurred in connection to differential survival and reproductive success, often expressed in terms such as "survival chances" or "chance of reproducing" (Table 5). This was most common in responses to the cheetah and salamander items, possibly because cheetahs and salamanders reproduce sexually, and cheetahs also care for their offspring. Thus, the higher frequency of probability in responses to the cheetah item may be related to the more frequent use of differential survival and reproductive success in them. Probability was more seldom connected to origin of variation, and mostly in responses to the bacteria item. Both the probability of novel traits and mutations were used in almost a quarter of the cases.

Our results indicate a positive association between trait gain and expression of probability in the cheetah item. A possible reason for this is that quantitative traits such as running speed may be easier to connect to probability, because a population includes gradual variation in quantitative traits (in contrast to discontinuous traits, such as some instances of antibiotic resistance). Most of the students were probably familiar with cheetahs and could easily conceptualize how running speed is related to survival and reproductive success. In addition, students probably have a nondeterministic model of hunting, that is, that you do not always succeed in games such as run and catch (tag games). Thus, the students could easily transfer this nondeterministic model to the cheetah context. Conversely, it is not surprising that the bacteria item elicited less use of probability, because learners have no direct experience of how bacteria function.

However, we noted that, rather than describing survival as a phenomenon with stochastic components, several of the responses included deterministic formulations like "only the most fit survive." It is worth noting that the widely used

metaphor "survival of the fittest" could be misleading for learners, especially from a probabilistic viewpoint.

In summary, probabilistic reasoning was generally rare and varied across the items, indicating an effect of item context.

Temporal Scale

It is well known that deep time is challenging for students (Hidalgo and Otero, 2004; Catley and Novick, 2009; Cheek, 2012; Johnson *et al.*, 2014). However, most studies focused on declarative knowledge such as the age of the earth or the timing of important evolutionary events like the emergence of photosynthesis. Much less attention has been paid to students' understanding of the role of the huge temporal scales in evolutionary processes such as natural selection.

Indication of time was present in roughly equal proportions (25–28%) of responses to the three items (Figure 2), in accordance with findings by others (e.g., Nieswandt and Bellomo, 2009). However, our consistency analysis revealed that this was not due to the same students repeatedly applying the concept in their responses to the three items (Table 3).

The in-depth analysis of time (Table 5) revealed that most of the students did not integrate the time aspect into their explanations. Most of the time mentions were vague and unspecific, typically "adaptation takes time" or "traits evolve over time." Such mentions were most common in responses to the salamander item and least common in the bacteria item. The mentions of time were not directly connected to any natural selection key concept, that is, "adaptation takes time" or "traits evolve over time." Surprisingly, very few of the answers mentioned "accumulation of traits" over time. A small minority included the central aspect that natural selection takes places across generations in a population and mostly in responses to the cheetah and salamander items. Because bacteria have short generation times, it is not surprising that generation time was only mentioned in answers to this item. The presence of "generations" in responses to the cheetah and salamander items could be an effect of familiarity with animals reproducing in distinct generations, while bacterial reproduction is less familiar.

Shorter timescales (days or shorter) were not mentioned in the responses, while longer timescales such as years were mentioned at low frequencies, and only in responses to the cheetah and salamander items. This is not very surprising, because the items did not focus on the time aspect per se. However, our results give valuable indications about what to expect from students in terms of addressing shorter timescales when using typical natural selection items.

In conclusion, we found that time aspects were scarce and poorly integrated in students' explanations of natural selection. Thus, if time is an important cross-cutting concept in science education and a threshold concept, it is problematic that roughly two-thirds of the students failed to mention that evolutionary change occurs across generations. Furthermore, understanding that natural selection occurs only over generations is crucial for distinguishing the process from developmental processes or physiological responses to the environment. In addition, generation times strongly influence the rapidity of evolution.

The fact that we found no example of students reasoning explicitly with large time frames in our sample also raises concerns. In addition, students rarely linked any components of natural selection to time. This indicates that students either are generally unaware of the importance of time or that they do not consider time aspects such as accumulated changes important in explanations. To further explain this finding, we suggest that research should be undertaken to elucidate whether students are unaware of the significance of multiple generations for evolutionary change or whether they just consider this aspect unimportant in explanations of natural selection. We also regard the ORI items to be of limited use for evaluating students' reasoning about the role of time in natural selection. Hence, new items should be developed that are better suited to assess students' time reasoning skills.

Spatial Scale

In our initial analysis of participants' use of spatial scale, we focused on instances in which learners made connections between at least two organizational levels (Figure 2), because we consider this an important threshold-crossing step. The detailed analysis also examined mentions of objects and processes on specific organizational levels, regardless of whether connections were made to other levels (Figure 4). Overall, the explanations mostly focused on the macro (individual) and supermacro (population) levels. Although change in population composition is the main outcome of natural selection, mentions of populations were less common than mentions of individuals in responses to all three items (Figure 4). Interestingly, the bacteria item elicited most of the mentions of the lower organizational levels (submicro and micro), which explains the high overall occurrence of spatial scale in responses to this item. Organizational levels between the genetic level (protein and cellular) and individual level were (in principle) referred to only in explanations of the bacteria item. The bacteria context also yielded explanations with more links across organizational levels and inclusion of a wider range of levels (Figure 5). In fact, the bacteria context generated approximately twice as many links as either the cheetah or salamander context, and this was closely connected to mutations and the randomness concept. However, the genetic level was at most present in roughly a third of the explanations for the bacteria and cheetah items, but no more than 8% of the explanations for the salamander item. This suggests, somewhat surprisingly to us, that genetic causes are less associated with trait loss than with trait gain. In addition, only 3% of the responses to the salamander item showed evidence of linking from the genetic level. Thus, linking genetic changes with phenotypic changes seems more challenging for students in connection with trait loss than with trait gain.

This is interesting, because earlier research showed that students have problems incorporating genetic aspects correctly in evolutionary explanations (Duncan and Reiser, 2007; Jördens *et al.*, 2016) and in linking mutations to effects on higher organizational levels (Nieswandt and Bellomo, 2009). Marbach-Ad and Stavy (2000) found that bacteria-related questions elicited a higher level of submicroscopic concepts than animals or plants in student explanations of genetic phenomena. Thus, our results are consistent with these earlier findings, but we also found a link between the inclusion of genetic-level explanations and the tendency to include randomness.

Unfortunately, addressing these difficulties with instruction has proven challenging. Even with interventions targeting the

connection between the submicro/micro level and macro levels, these relationships seem challenging for students to grasp (Bray Speth *et al.*, 2009).

The Relation between Item Type and Concept Use

As indicated in the preceding section, we found that the use of threshold concepts was related to the item context. Effects of item features on students' explanations of natural selection have been observed in previous studies (e.g., Nehm and Ha, 2011; Opfer *et al.*, 2012; Heredia *et al.*, 2016; Großschedl *et al.*, 2018), but the relationship between item type and threshold concept use has not been previously surveyed. Features such as trait polarity (gain or loss) and type of taxa have been found to affect learners' explanations of natural selection. Familiarity of the trait and the example organism (animal or plant) also has proven significant. However, fungi and micro-organisms as well as unicellular and prokaryotic taxa are clearly underexplored. Our results indicate that micro-organisms might have some affordances such as more mentions of lower organizational levels, links between organizational levels, and randomness. Also, bacteria are unicellular, and therefore students might be less tempted to explain evolutionary change with ideas such as willful changes in animals' organs. Thus, micro-organisms could be a fruitful context to use in evolution teaching besides the previously recommended animal and plant contexts (Nehm and Ha, 2011; Heredia *et al.*, 2016). In addition, we argue that bacteria have additional affordances, such as short generation times and (often) large numbers in small physical spaces. Also, using bacteria as biological model systems for evolution avoids the complexities of Mendelian genetics (e.g., dominant and recessive alleles), chromosomal crossover, and sexual reproduction.

The cheetah item also involves trait gain, but in a multicellular animal. Both cheetahs and the trait are likely familiar to learners. According to previous research, novices' reasoning about evolutionary processes is facilitated by consideration of familiar taxa and traits (Federer *et al.*, 2015). While this seemed to be the case for certain key concepts such as individual variation, differential survival, and reproductive success in this study, some of the threshold concepts were actually less frequent in responses to this item than in the bacteria item (i.e., *randomness and spatial scale*). However, probability appeared most frequently in responses to the cheetah item, in conjunction with high frequencies of individual variation and differential survival, because students connected running speed with survival chances. Thus, the combination of familiarity of the trait and its significance for survival is a likely explanation for the higher occurrence of probability in responses to this item, but not the lower frequency of organizational links (spatial scale) or randomness.

The salamander item involves trait loss and seems to be a poorer indicator of basic understanding of natural selection (in terms of key and threshold concepts) than the other items, eliciting the lowest frequencies of all the concepts, except for selection pressure and temporal scale. This is not surprising, considering students' documented difficulties with trait loss (Bizzo, 1994; Nehm and Reilly, 2007; Nehm and Ha, 2011; Ha and Nehm, 2013; Federer *et al.*, 2015). Many students in our sample provided simplistic explanations, suggesting that the absence of selection pressure or sometimes "lack of need" somehow brought about a reduction of eyes over time. This might explain why the frequencies of selection pressure and temporal scales were high in responses to this item. In addition, we suspect that this particular item is especially difficult, because students seem to have difficulties in imagining the selective advantage of impaired eye development. Many of the students actually devoted a large part of their answers to attempts to explain the selective advantage of reduced vision. Other examples of trait loss in which the loss of the trait is coupled to a clear selective advantage, like the loss of limbs in aquatic tetrapods, might be more suitable for novices.

LIMITATIONS

One limitation of the present study concerns the study design that we used to investigate how students use the threshold concepts randomness, probability, and temporal and spatial scales in their written explanations of the process of natural selection. The variation of surface features between the items was not completely systematic. This was a result of using the extant ORI instrument, but the advantages of using a well-established test instrument were considered to outweigh the limitations at this stage. If a more systematic variation and several items for each surface feature were used, the number of items required would have been substantial, thus risking test fatigue. Thus, our results should be considered a first indication that item features affect students' use of threshold concepts.

Although open-response items have several advantages over multiple-choice items, in that they measure higher cognitive abilities (e.g., Kuechler and Simkin, 2003; Nehm and Schonfeld, 2008) and reduce random guessing, unintended corrective feedback, and problem solving by working backward from the answers (Bridgeman, 1992), they also have several limitations. First, open-response items are less efficiently scored by individuals (Bennett, 1991; Nehm and Schonfeld, 2008), although this is changing with automated scoring (Moharreri *et al.*, 2014). Second, students' aversion to writing and/or poor writing skills may result in an inaccurate reflection of their knowledge (Nehm and Schonfeld, 2008). Still, the Nehm and Schonfeld (2008) study also indicated high correlations between interviews and open-response items. Finally, open-response items may capture what students view as most salient to answer the questions, but they may not capture students' implicit understanding of the respective key or threshold concepts. However, it has been shown that open-response items (the ORI) indeed produce similar magnitudes of key concepts as closed-response items (the CINS) or interviews in undergraduates (Nehm and Schonfeld, 2008). This suggests that the results obtained by using open-response items in our study are representative of what would have been uncovered in interviews. While the present study lacks such as interviews, they are planned for future studies in which we aim to also include teachers and experts.

However, our aim with the study was not to uncover the exact level of threshold concepts "awareness." Rather, we sought to investigate whether threshold concepts were present in students' explanations of typical natural selection items and whether the differing contexts of the items affected the frequencies of threshold concepts used.

Finally, students in the university sample included in this study were largely drawn from many different universities in Germany and, to a smaller extent, from different disciplines in

Sweden. Thus, results could differ in other populations. Despite that, we documented a wide range of explanations from virtually none to elaborate and scientifically acceptable explanations, and we are confident that our data represent some of the diversity likely to be found in broader samples.

CONCLUSIONS AND IMPLICATIONS FOR BIOLOGY EDUCATION

Our findings confirm previous findings regarding students' use of key concepts in explanations of natural selection (Nehm and Ha, 2011; Federer *et al.*, 2015; Heredia *et al.*, 2016), but at the same time, our results show that threshold concepts are inconsistently used by students in responses to items with various surface features. Our results also confirm previous findings that trait loss seems challenging for students to explain. More interestingly, in terms of our research questions, loss of traits seems to be a problematic context for eliciting not only key concepts but also threshold concepts. Overall, learners showed inconsistency in ability to consider organizational levels (spatial scale), randomness, and probability in responses to the three items with various surface features. Therefore, assessment of threshold concept acquisition should consider the items used. In addition, it should be considered whether items designed for probing natural selection understanding in terms of key concepts are in fact suitable for probing threshold concepts.

Briefly, our analysis revealed the following major findings: 1) Stochastic elements of natural selection (random origin of variation and probabilistic survival, reproduction, and inheritance) are rare and context dependent in students' explanations of evolutionary phenomena. 2) Frequencies of integration of organizational levels and linking between different spatial scales are generally low and context dependent in students' explanations (notably they were most frequent in responses to the bacteria item). 3) References to the origin of variation are also highly context dependent and were most prevalent in responses to the bacteria item. 4) Time aspects of natural selection were only mentioned in a minority of the explanations, and most of them were unspecific and disconnected from key concepts of natural selection.

In addition, our results indicate that, although Nehm and Reilly (2007) found minor differences in the total number of key concepts applied in responses to the three items, there can be significant differences in the specific key concepts an item elicits.

Our findings suggest a number of implications for educational research and practice. First, education should encourage students to compare and contrast different examples of evolution across taxa and trait polarity. Second, special attention should be paid to how learners understand and use threshold concepts in their explanations of natural selection. Third, the unity of life (such as cellularity and DNA) should be reinforced in connection with evolution and the relevance of organizational levels should be emphasized (in both teaching and textbooks); for example, it should be stressed that that causes of evolutionary phenomena arise on different (often imperceptible) scale levels from outcomes.

Fourth, the role of randomness in evolution should be a focus (in conjunction with help for students to transfer this principle to different examples of natural selection), as should the probabilistic rather than deterministic nature of natural selection (cf. "only the fittest survive").

Finally, learners might need training in how to interpret and solve biology problems within an evolutionary framework by establishing cognitive strategies and strengthening situational knowledge.

SUGGESTIONS FOR FUTURE RESEARCH

Future research should explore the extent to which threshold concepts elicited in open-response items are comparable to those obtained by other instruments such as interviews and items explicitly targeting threshold concept understanding. In addition, the context effect indicated our results need further confirmation and exploration. For example, potential effects of physiological/morphological changes and additional taxa, among other topics, could be explored. Also, interventions that strive to develop student understanding of threshold concepts and how they factor into evolutionary mechanisms such as natural selection could be used to assess the significance of threshold concepts for understanding of natural selection. In addition, comparisons could be made to assess whether grasping a threshold concept entails a change in understanding in several domains such as biology and chemistry (e.g., randomness and spatial scale are important concepts in both biological evolution and diffusion in chemistry). This would also yield evidence as to whether focusing on cross-cutting or threshold concepts in curricula is a feasible strategy.

ACKNOWLEDGMENTS

We thank the whole EvoVis group (Gustav Bohlin, Ute Harms, Gunnar Höst, Nalle Jonsson, Marta Koc-Januchta, Konrad Schönborn, and Jörgen Stenlund) for valuable support during work on the article. Special thanks are due to Gunnar Höst for valuable discussions on statistics and John Blackwell for language review. We also thank the Swedish Research Council for providing the funds for the research presented in the paper. This work was supported by the Swedish Research Council (VR 2012:5344, LT).

REFERENCES

Aleixandre, M. P. J. (1994). Teaching evolution and natural selection: A look at textbooks and teachers. *Journal of Research in Science Teaching*, *31*(5), 519–535.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*(10), 952–978. https://doi.org/10.1002/tea.10053

Batzli, J. M., Knight, J. K., Hartley, L. M., Maskiewicz, A. C., Desy, E. A., & Momsen, J. (2016). Crossing the threshold: Bringing biological variation to the foreground. *CBE—Life Sciences Education*, *15*(4), es9. https://doi.org/10.1187/cbe.15-10-0221

Bennett, R. E. (1991). On the meanings of constructed response. In Bennet, R. N., & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–28). Hillsdale, NJ: Erlbaum.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, *27*(5), 415–427. https://doi.org/10.1002/tea.3660270503

Bizzo, N. M. V. (1994). From down house Landlord to Brazilian high school students: What has happened to evolutionary knowledge on the way? *Journal of Research in Science Teaching*, *31*(5), 537–556. https://doi.org/10.1002/tea.3660310508

Bohlin, G., Göransson, A., Höst, G. E., & Tibell, L. A. E. (2017a). A conceptual characterization of online videos explaining natural selection. *Science & Education*, *26*(7–9), 975–999. https://doi.org/10.1007/s11191-017-9938-7

Bohlin, G., Göransson, A., Höst, G. E., & Tibell, L. A. E. (2017b). Insights from introducing natural selection to novices using animations of antibiotic resistance. *Journal of Biological Education 52*(3), 314–330. https://doi.org/10.1080/00219266.2017.1368687

Bray Speth, E., Long, T. M., Pennock, R. T., & Ebert-May, D. (2009). Using Avida-ED for teaching and learning about evolution in undergraduate introductory biology courses. *Evolution: Education and Outreach*, *2*(3), 415–428. https://doi.org/10.1007/s12052-009-0154-z

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*(3), 253–271.

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution* (Paper Commissioned by the National Academies Committee on Test Design for K–12 Science Achievement) (p. 67). Washington, DC: National Academy of Sciences.

Catley, K. M., & Novick, L. R. (2009). Digging deep: Exploring college students' knowledge of macroevolutionary time. *Journal of Research in Science Teaching*, *46*(3), 311–332. https://doi.org/10.1002/tea.20273

Cheek, K. A. (2012). Students' understanding of large numbers as a key factor in their understanding of geologic time. *International Journal of Science and Mathematics Education*, *10*(5), 1047–1069. https://doi.org/10.1007/s10763-011-9312-1

Cheek, K. A., LaDue, N. D., & Shipley, T. F. (2017). Learning about spatial and temporal scale: Current research, psychological processes, and classroom implications. *Journal of Geoscience Education*, *65*(4), 455–472. https://doi.org/10.5408/16-213.1

Demastes, S. S., Settlage, J., & Good, R. (1995). Students' conceptions of natural selection and its role in evolution: Cases of replication and comparison. *Journal of Research in Science Teaching*, *32*(5), 535–550. https://doi.org/10.1002/tea.3660320509

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, *35*(3), 125–129. https://doi.org/10.2307/4444260

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. https://doi.org/10.1002/tea.20186

Elmesky, R. (2013). Building capacity in understanding foundational biology concepts: A K–12 learning progression in genetics informed by research on children's thinking and learning. *Research in Science Education*, *43*(3), 1155–1175. https://doi.org/10.1007/s11165-012-9286-1

Federer, M. R., Nehm, R. H., Opfer, J. E., & Pearl, D. (2015). Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. *Research in Science Education*, *45*(4), 527–553. https://doi.org/10.1007/s11165-014-9435-9

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

Fiedler, D., Sbeglia, G. C., Nehm, R. H., & Harms, U. (2019). How strongly does statistical reasoning influence knowledge and acceptance of evolution? *Journal of Research in Science Teaching*, *56*(9), 1183–1206. https://doi.org/10.1002/tea.21547

Fiedler, D., Tröbst, S., & Harms, U. (2017). University students' conceptual knowledge of randomness and probability in the contexts of evolution and mathematics. *CBE—Life Sciences Education*, *16*(2), ar38. https://doi.org/10.1187/cbe.16-07-0230

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE—Life Sciences Education*, *7*(2), 227–233.

Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach*, *2*(2), 156–175. https://doi.org/10.1007/s12052-009-0128-1

Großschedl, J., Seredszus, F., & Harms, U. (2018). Angehende Biologielehrkräfte: Evolutionsbezogenes Wissen und Akzeptanz der Evolutionstheorie. *Zeitschrift für Didaktik der Naturwissenschaften*, *24*(1), 51–70. https://doi.org/10.1007/s40573-018-0072-0

Grotzer, T. A., Solis, S. L., Tutwiler, M. S., & Cuzzolino, M. P. (2017). A study of students' reasoning about probabilistic causality: Implications for understanding complex systems and for instructional design. *Instructional Science*, *45*(1), 25–52. https://doi.org/10.1007/s11251-016-9389-6

Ha, M., & Nehm, R. H. (2013). Darwin's difficulties and students' struggles with trait loss: Cognitive-historical parallelisms in evolutionary explanation. *Science & Education*, *23*(5), 1051–1074. https://doi.org/10.1007/s11191-013-9626-1

Heredia, S. C., Furtak, E. M., & Morrison, D. (2016). Exploring the influence of plant and animal item contexts on student response patterns to natural selection multiple choice items. *Evolution: Education and Outreach*, *9*(1), 10. https://doi.org/10.1186/s12052-016-0061-z

Hidalgo, A. J., & Otero, J. (2004). An analysis of the understanding of geological time by students at secondary and post-secondary level. *International Journal of Science Education*, *26*(7), 845–857. https://doi.org/10.1080/0950069032000119438

Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement*, *24*(4), 749–753.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Johnson, C. C., Middendorf, J., Rehrey, G., Dalkilic, M. M., & Cassidy, K. (2014). Geological time, biological events and the learning transfer problem. *Journal of the Scholarship of Teaching and Learning*, *14*(4), 115–129. https://doi.org/10.14434/v14i4.4667

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, *7*(2), 75–83.

Jördens, J., Asshoff, R., Kullmann, H., & Hammann, M. (2016). Providing vertical coherence in explanations and promoting reasoning across levels of biological organization when teaching evolution. *International Journal of Science Education*, *38*(6), 960–992. https://doi.org/10.1080/09500693.2016.1174790

Kampourakis, K., & Zogza, V. (2008). Preliminary evolutionary explanations: A basic framework for conceptual change and explanatory coherence in evolution. *Science & Education*, *18*(10), 1313–1340. https://doi.org/10.1007/s11191-008-9171-5

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* (PhD thesis). Utrecht, Netherlands: Centrum voor Didactiek van Wiskunde en Natuurwetenschappen, Universiteit Utrecht. Retrieved from https://dspace.library.uu.nl/handle/1874/219

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (p. 2013). Thousand Oaks, CA: Sage.

Kuechler, W. L., & Simkin, M. G. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education*, *14*(4), 389–400.

Larsson, C., & Tibell, L. A. E. (2014). Challenging students' intuitions—The influence of a tangible model of virus assembly on students' conceptual reasoning about the process of self-assembly. *Research in Science Education*, *45*(5), 663–690. https://doi.org/10.1007/s11165-014-9446-6

Lee, H.-S., Liu, O. L., Price, C. A., & Kendall, A. L. M. (2010). College students' temporal-magnitude recognition ability associated with durations of scientific changes. *Journal of Research in Science Teaching*, *48*(3), 317–335. https://doi.org/10.1002/tea.20401

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology & Systematics*, *1*, 1–18.

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

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction*, *16*(3), 285–265.

Meyer, J., & Land, R. (2003). Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines. In Rust, C. (Ed.), *Improving student learning—Theory and practice ten years on* (pp. 412–424). Oxford, UK: Oxford Centre for Staff and Learning Development. Retrieved January 8, 2019, from www.dkit.ie/ga/system/files/Threshold_Concepts__and_Troublesome_Knowledge_by_Professor_Ray_Land.pdf

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems.

Journal of Research in Science Teaching, *46*(6), 675–698. https://doi.org/10.1002/tea.20314

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: An online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1), 1–14.

Nadelson, L. S., & Southerland, S. A. (2009). Development and preliminary evaluation of the Measure of Understanding of Macroevolution: Introducing the MUM. *Journal of Experimental Education*, *78*(2), 151–190. https://doi.org/10.1080/00220970903292983

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*(2), 92–98. https://doi.org/10.1525/abt.2012.74.2.6

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, *48*(3), 237–256. https://doi.org/10.1002/tea.20400

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

Nehm, R. H., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? *Evolution: Education and Outreach*, *4*(4), 666–679. https://doi.org/10.1007/s12052-011-0369-7

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. https://doi.org/10.1002/tea.20251

Nieswandt, M., & Bellomo, K. (2009). Written extended-response questions as classroom assessment tools for meaningful understanding of evolutionary theory. *Journal of Research in Science Teaching*, *46*(3), 333–356. https://doi.org/10.1002/tea.20271

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching*, *49*(6), 744–777. https://doi.org/10.1002/tea.21028

Perkins, D. N., & Grotzer, T. A. (2005). Dimensions of causal understanding: The role of complex causal models in students' understanding of science. *Studies in Science Education*, *41*(1), 117–165. https://doi.org/10.1080/03057260508560216

Ross, P. M., Taylor, C. E., Hughes, C., Kofod, M., Whitaker, N., Lutze-Mann, L., & Tzioumis, V. (2010). Threshold concepts: Challenging the way we think, teach and learn in biology. In Meyer, J. H. F., Land, R., & Baillie, C. (Eds.), *Threshold concepts and transformational learning* (Vol. 1, pp. 165–178). Rotterdam, Netherlands: Sense Publishers.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science*, *21*(2), 147–177.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss (Jahrgangsstufe 10) [National educational standards for the subject of biology concerning the 10th grade]*, Munich, Germany: Luchterhand.

Settlage, J. (2007). Conceptions of natural selection: A snapshot of the sense-making process. *Journal of Research in Science Teaching*, *31*(5), 449–457. https://doi.org/10.1002/tea.3660310503

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Skolverket. (2011). *Biology curriculum for Swedish upper-secondary school*. Retrieved August 11, 2019, from www.skolverket.se/download/18.4fc05a3f164131a7418104a/1535372296309/Biology-swedish-school.pdf

Smith, M. U. (2009a). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. *Science & Education*, *19*(6–8), 523–538. https://doi.org/10.1007/s11191-009-9215-5

Smith, M. U. (2009b). Current status of research in teaching and learning evolution: II. Pedagogical issues. *Science & Education*, *19*(6–8), 539–571. https://doi.org/10.1007/s11191-009-9216-4

Swarat, S., Light, G., Park, E. J., & Drane, D. (2011). A typology of undergraduate students' conceptions of size and scale: Identifying and characterizing conceptual variation. *Journal of Research in Science Teaching*, *48*(5), 512–533. https://doi.org/10.1002/tea.20403

Tibell, L. A. E., & Harms, U. (2017). Biological principles and threshold concepts for understanding natural selection: Implications for developing visualizations as a pedagogic tool. *Science & Education*, *26*(7–9), 953–973. https://doi.org/10.1007/s11191-017-9935-x

Tsui, C.-Y., & Treagust, D. F. (2013). Introduction to multiple representations: Their importance in biology and biological education. In *Multiple representations in biological education* (pp. 3–18). Dordrecht, Netherlands: Springer. Retrieved March 22, 2017, from http://link.springer.com/10.1007/978-94-007-4192-8_1

Wallin, A. (2004). *Evolutionsteorin i klassrummet: På väg mot en ämnesdidaktisk teori för undervisning i biologisk evolution (PhD thesis)*. University of Gothenburg, Gothenburg, Sweden. Retrieved January 12, 2015, from https://gupea.ub.gu.se/handle/2077/9494

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. https://doi.org/10.1023/A:1009421303064

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. https://doi.org/10.1037/a0037489

Zohar, A., & Ginossar, S. (1998). Lifting the taboo regarding teleology and anthropomorphism in biology education—Heretical suggestions. *Science Education*, *82*(6), 679–697.