

Clas Olander is lecturer at the Department of Education, University of Gothenburg. Drawing on a background in teaching at several levels in public schools his current interest is the cross section of 'practice and theory'; especially students' use of language when making meaning of school science.

## CLAS OLANDER

Department of Education, University of Gothenburg, Gothenburg, Sweden.

clas.olander@gu.se

# Teaching biological evolution - internal and external evaluation of learning outcomes

## Abstract

*This paper reports from a study where teachers and researchers collaborate on designing and validating topic-oriented teaching-learning sequences. In an iterative process, data about learning and teaching biological evolution are generated through continuous cycles of design, teaching, evaluation, and redesign. The study involved 180 Swedish students aged 11 – 16, and the overall learning aim was that the students should be able to use the theory of evolution as a tool when explaining the development of life on earth. The aim of this paper is to validate the students' learning outcome, estimated as appropriation of scientific ways of reasoning in written answers. The students' answers of questions are analysed before and after interventions (internal evaluation), and compared with the answers from a national sample (external evaluation). The students in the experimental group did develop their reasoning, and they attained the aim, to a greater extent than a national sample.*

## INTRODUCTION

Biological evolution is an area in the science curriculum that has been identified as challenging and in which traditional teaching often fails, when it comes to students' lasting conceptual understanding (Asterhan & Schwarz, 2007; Bishop & Anderson, 1990; Bizzo, 1994; Kampourakis & Zogza, 2008; Sandoval & Millwood, 2005; Shtulman, 2006; Thomas, 2000; Wallin, 2004). Of the few intervention studies that show effective learning, some use activities among peers that focus reasoning; for example, paired problem-solving (Jensen & Finley, 1996), peer group discussions about different explanations (Jiménez- Aleixandre, 1992; Wallin, 2004), and dialectical argumentation (Asterhan & Schwarz, 2007). However, the successful examples of evidence-based teaching do not often reach practice, and this dividing line between research and practice is an issue that concerns governments as well as communities of educational research (Millar, Leach, Osborne & Ratcliffe, 2006).

The notion of *design-based research* is one strategy to bridge the supposed gap between research in science education and practice. In the US, design research is discussed in thematic issues of Educational Researcher (Kelly, 2003), the Journal of the Learning Sciences (Barab & Squire, 2004), and Educational Psychologist (Sandoval & Bell, 2004). In Europe, design research has been presented in a thematic issue of the International Journal of Science Education (Meheut & Psillos, 2004). Several research groups work in line with this approach using different headings, for example, "development research" (Linjse, 1995), "educational reconstruction" (Kattmann, Duit &

Gropengießer, 1998), “didactical transposition” (Tiberghien, 2000), “teaching-learning sequences” (Leach & Scott, 2002), “learning studies” (Lo, Marton, Pang & Pong, 2004), “design-based research” (Design-based research collective, 2003), and “pragmatic approach to design-based research” (Juuti & Lavonen, 2006). Albeit the differences among the approaches, they share, at least, two commonalities (Lindwall, 2008). They aim at doing real work in practical educational settings and the working process includes separable parts of “design, enactment, analysis and redesign” (ibid, p. 26). Sandoval and Bell (2004) summarize the design-based approach claiming that “it is theoretically framed empirical research of learning and teaching based on particular designs for instruction” (ibid, pp. 199-200).

The approach in the study presented in this paper share the mutual characteristics that are pointed out by Sandoval and Bell (2004) and Lindwall (2008). In an iterative process, empirical data about learning and teaching biological evolution is generated through continuous cycles of design, teaching, evaluation, and redesign. This is done as part of a developmental project, labelled *design and validation of topic-oriented teaching-learning sequences* (Andersson, Bach, Hagman, Olander & Wallin, 2005) where teachers and researchers collaborate. The aim of this paper is to describe the process of designing a teaching-learning sequence and validating versus the students’ learning outcome.

### **Explaining biological change**

In science classes, students are often asked to explain phenomena and are then expected to use a causal explanation. This is not a self-evident task. First of all, students have to distinguish descriptions from explanations. Ogborn, Kress, Martins and McGillicuddy (1996) conclude that describing and labelling is common in science classrooms, because it provides “material for explanations. The entities which are to be used in explanations have to be ‘talked into existence’ for students” (ibid, p. 14). When explaining biological phenomena, students often use spontaneous and situated explanations. These explanations could be anthropomorphic, attributing human characteristics to non-human organisms, or teleological, where events have a goal, purpose, or even a design (Zohar & Ginossar, 1998; Kattmann, 2008). According to Keleman (1999), reasoning in a teleological way is common, in the sense that we frequently ask “why?” and “what’s that for?” when confronted with a biological issue. For example, when we see the webbed foot of the mallard we spontaneously explain that the mallard have this foot in order to swim.

Causality in biology could be regarded as either proximate or ultimate according to Mayr (1961) or, as Ariew (2003) rephrased it, proximate or evolutionary explanations. Answers to questions that start with “what is the cause” deals with either short-term (proximate) or long-term (evolutionary/ultimate) perspectives. Responses with short time scales are due to immediate previous events and they are appropriate in physiology and medicine, while evolutionary (ultimate) explanations involve longer time, several generations, and selection. Students often don’t distinguish proximate from evolutionary explanations, or don’t recognise what kind of time perspective their answer is supposed to deal with. Abrams, Southerland and Cummins (2001) developed the idea that there are both proximate and ultimate answers to “how and why” questions. Ariew (2003) suggests that answers to questions about “how” should refer to proximate causes; while towards “why-questions” evolutionary explanations are more fruitful.

### **Aim and research focus**

Evaluation of interventions in school could be either internal or external. An example of internal evaluation is comparisons of pre- and post-tests within the experimental group. In external evaluation the aim is to compare with other teaching practices, for example experimental versus control groups. Irrespective of the approach, evaluation is performed in relation to specific goals. The goals of the Swedish compulsory school are evaluated by the Swedish National Agency who initiated a national evaluation in 2003 with the aim of *providing an overall picture of goal*

*attainment in the compulsory school, by subject and from an overall perspective* (National Agency of Education, 2004, p. 8). In the same document the different actors', the state and the individual teacher, responsibilities are formulated as:

The structure of the syllabuses reflects the division of responsibility between the state and the professionals in the school. By means of setting up the goals, as well as the results to be expected, the state imposes demands on the quality and equivalence of the education. How the goals are to be attained, namely choice of content and method, is determined by the teacher (ibid, p.16).

The Swedish grade 1 – 9 curriculum for compulsory school, states that “the school in its teaching of biology should aim to ensure [...] that pupils develop their knowledge of the conditions and development of life and are able to see themselves and other forms of life from an evolutionary perspective” (National Agency of Education, 2000). In the teaching-learning sequence developed in the present study, this national aim was interpreted into a specified learning goal, namely that the students should be able to use the theory of evolution as a *tool* when explaining the development of life on earth. In this context it is crucial if a model/ theory/concept are seen as *goal* (product) or *means* (process). To “learn” the theory of evolution could be a *goal* to attain, but then you open up for the possibility to learn more or less by heart; students repeat the right words. A model/theory/ concept could also be put into use as *means* or theoretical lever in the process of sense-making. It was in this latter way we regarded the theory of evolution in this teaching intervention, hence expressed as *theory as a tool*.

The specific research focus in this paper is to assess the teaching-learning sequence, with respect to students' learning outcome. Data are generated through students' written answers to a pre-test and a delayed post-test (three months after teaching), and a comparison is made with the answers from a comparable national sample. Doing this includes a validation of the test instrument in itself.

## METHODOLOGY

### Sample and project design

The intervention study could be described as a cyclic knowledge-building process, in which both teachers and researchers contribute. After an announcement-letter to schools, four teachers volunteered to participate. The teachers were all qualified to teach biology at this school level; beside that, they represent various teaching experience, gender, and age. The students were all part of the Swedish compulsory school system and came from different environments; one school was situated in the centre of town, two in multicultural suburbs, and one a bit outside town but still at commuting distance. All four teachers taught the sequence twice with different groups during one school year (Figure 1). This means that about 180 students in eight groups, 11 – 16 years old, were involved (these groups are henceforth referred to as ‘the experimental group’). For the sake of comparison the experimental group is sometimes discussed as two groups, divided according to the age of the students.

During the first design phase, teachers and researcher had four meetings where we made a didactical analysis of the content and formulated a teaching strategy. This analysis was grounded in literature and our own teaching experience, and both informed our planning of the teaching intervention. Early in this period the teachers gave some written questions about evolution to their own students, which they knew had participated in teaching about evolution. When we later discussed the students' answers, the teachers were at first dumbfounded about, “how on earth could my students write like this”. However, soon we started a rewarding conversation about reasons for students' reasoning. Retrospectively, this is seen as a turning point. From now on the teachers' engagement and ownership of the process increased and they made increasingly valuable contributions.

May/Aug/Sep/Oct	Nov/Dec	January	Feb/March	May/June
Design meetings	Teaching 1	Evaluation 1 / new design	Teaching 2	Evaluation 2

Figure 1. Working process during the two cycles.

The work reported in this paper emanates from a developmental project, in which both teachers and researchers contributed. However, all involved in a project like this do not contribute equally throughout the process. Earlier in this paper the process is described as “design, teaching, evaluation, and redesign”; these phases are interlinked but the different actors’ – the teachers’ and researchers’ – contribution varies. In the design phase the researchers were more active in the beginning, for example in choosing the literature and preparing the diagnostic instrument. The result of the design phase (the learning goal and the teaching strategy) was achieved through collaborative work and could be regarded as guidelines for the intended teaching. The teaching was solely the teachers’ responsibility, but in evaluations and redesign both teachers and researchers were involved.

### Formulating the learning goal

As an overall learning goal, we decided that students should be able to use the theory of evolution by means of natural selection as a tool when explaining the development of life on earth. It was a conscious decision, mainly due to time limits and the age of the students, not to include, for example, neutral evolution, or drift, and to restrict teaching to adaptive evolution (natural selection). One peculiarity with the theory of evolution, as with all theories, is that it is construed by different parts. Following advice from, among others, Bishop and Anderson (1990), and Wallin (2004), we decided to separate the explanation of evolution into two parts when staging the scientific story of evolution. The first process to be dealt with was the origin of variation, which is a random process, mainly due to mutation and recombination. Selection, which is the second part, is not a random process, but an effect of the variation meeting the environment. Selection is often perceived as differences in survival and we wanted it to (also) include different reproduction rates.

This means that we reformulated the modern synthesis of the theory of evolution (Stearns & Hoekstra, 2000) to school science settings. In Sweden, it is the responsibility of the teachers to attain these kinds of reformulations as well as the formulation of specific learning goals. The aim of this paper is to evaluate the students’ attainment of the learning goal, which is summarised as follows: *after the teaching intervention the students should be able to explain the evolution of life on earth using the meaning of the terms heredity, variation, and selection.*

### Formulating a teaching strategy

One of the mutual conclusions from the literature about students’ reasoning about evolution (see the references in the introduction) is that students often explain biological change referring to terms like “*need, wish, and/or effort*”. These terms along with terms that are scientifically central (we labelled them key terms), namely *heredity, variation, and selection* were to be elaborated and made meaning of in the teaching intervention. The intended strategy for the teaching was to present the theory of evolution as a scientific story and to engage students in different communicative settings; a meaning making process in relation to the key terms.

In that respect we reconsidered the staging of whole class and peer group discussions. In whole class discussion this meant that the ambition was to generally replace the frequent triadic communicative pattern of I-R-E with I-R-F-R-F-R- (Mortimer & Scott, 2003). The triadic pattern: Initiation (teacher’s question) – Response (student) – Evaluation (teacher) halts the conversation

and reduces the possibilities of true dialogue. If the teacher instead gives a feedback (F), the chance that the student will offer a new and elaborated response increases. If this is followed by another feedback, there is real communication going on. The teacher indicates by this that the students' contributions are valuable and worth attention.

Peer group discussions have successfully been used by Jiménez-Aleixandre (1992) in a teaching intervention about biological evolution. In order to link talks in peer groups and whole class we decided to try what Hoel (1996) calls "structured talks", which often emanate from a teacher initiated question or claim. Students then start with a short individual pondering; suggestively a minute or three, and preferably including some short writing. Then groups are formed where the issue is elaborated, and after some minutes the teacher enquires public response from the groups. This strategy give students the opportunity to ponder for themselves, test their opinions with others, and thus be better prepared for a public utterance in front of the whole class.

We agreed on a first set of student activities that would point to the use of key terms and the theory of evolution as a "two-part" process. On the basis of this, the teachers could plan their own teaching, guided by local circumstances, for example in relation to their students' age and previous experience. The scope of this article does not admit a close description of different activities, but a few that were supposed to contribute to the scientific story of "two separate processes" will be discussed.

We wanted to establish the notion of the existing variation in all populations. Our assumption was that this in a colloquial manner was frequently articulated as "everybody is unique". We wanted to extend this articulation to include all populations and start a discussion about the origin of these differences, and its consequences for future generations. One of the activities that were developed during the first teaching was a "game of recognition", using flowers, shells, leaves, peas, seeds etc. The description that follows concerns an activity in which dry chick peas were used.

Every student got a dry chick pea, and their assignment was to "learn to recognise it" without making any marks or the like on the pea. Then students gathered in groups of three/four, mingled their chick peas and tried to pick out their own. Nearly always this succeeded, and bigger and bigger groups were formed, and mostly students managed to recognise "their own pea" among the others. Many students started to care for their own peas, gave them names, wanted to take the peas home etc. Apart from this emotional aspect of the activity, the discoveries that even dry chick peas are different were striking to the students. This activity opened up for a teacher led structured talk concerning the origin of the differences among chick peas. The summary afterwards usually brought up both environmental and genetic reasons for the differences of the peas. The teacher could chose to pick up, or leave as a "cliff-hanger", the question of which of the reasons (genetic or environmental) would be important for future "pea-kids".

Another suggested activity was to make use of the students' answers to the pre-test questions. The answers should be made anonymous and handed out to groups as a basis for discussion. With this strategy individual students would not have to defend their own writing, if they at all remembered it; instead they are free to pick, from the pool of answers, what they now discern as a good explanation. This way of making use of students' own articulated explanations could be a way of facilitating productive talk, and sharing of ideas.

### **Selection of items and mode of comparisons**

When the Swedish National Agency for Education performed the national evaluation in 2003, a random national sample of students in grade 9 was given written questions. The evaluation was performed in the latter part of the spring, approximately three months before the end of the students' compulsory schooling. In this study the national sample is regarded as control group

in relation to the experimental group, which is an approach previously used by Bach (2001). In the national evaluation of science 2003 (National Agency of Education, 2004), the students were given 37 tasks to solve (divided into three tests), and in the Biology part, three tasks dealt with evolution (one open question and one multiple choice question accompanied with a request to justify their choice). These three tasks were also given to the experimental group, but only at the delayed post-test. In that way they were unfamiliar to the students and could serve as a point of comparison with the national sample. However, the students' ambition to answer is probably lower in the national sample, for example, there are around 50 % who "don't answer" the open-ended tasks about evolution compared to less than 10 % in the experimental group. It should be noted that it is not missing values; the students had the opportunity to answer, but preferred not to write anything, or wrote something irrelevant. Furthermore, there is a possibility that some students in the national sample not yet had any teaching about evolution; when the national evaluation was made there were still three months left of the term, and Zetterqvist (1998) conclude that many teachers in Sweden choose to teach evolution late in school year nine.

With the differences in answering rates in mind, a conversion of the results was made in order to get a more fair comparison; especially not overrating the results of the experimental group. The percentages presented in findings are recalculated as proportions of students answering the actual question, and the statistical comparisons of potential differences between groups are calculated with the  $\chi^2$ -method.

The intercoder reliability was checked by giving the answers and category headings to two educational scientists who were familiar with the content area. They independently categorised answers; at first the three of us agreed in 77 % of the cases, and after a discussion about interpretation we reached 90 % agreement.

### **Analysing answers about the evolution of a new trait**

Here focus in analysis is a question that was first used by Bishop and Anderson (1990):

Cheetahs are able to run fast, around 100 km/h when chasing prey. How would a biologist explain how the ability to run fast evolved in cheetahs, assuming their ancestors could only run 30 km/h?

Students' written answers were grouped in order to reflect qualitatively different ways of reasoning. The system of categories that emerged had students' actual wording in the foreground but was influenced by previous educational research about students' ideas and ways of arguing as well as scientific views of the specific area. The students' answers to the "cheetah-problem" are categorised as follows, with examples from the students' answers in italics.

- a) The answer describes, but do not explain: *They developed and got longer legs, and they became more vigorous.*
- b) The answer explains in a teleological way; mainly with words like need, had to, strive: *Cheetahs have to run fast in order to catch their prey.*
- c) The answer explains only with some key terms: *A biologist would explain like this; it occurred mutations in the genes of the cheetah, which made it run faster.*
- d) The answer explains in terms of natural selection: *When one cheetah was born it had, for example longer legs, which made it run faster and therefore gets more food, survive longer and then spread its genes.*
- e) No answer or irrelevant answer, or repeats the question: *don't know etc.*

The answers in the category a) describe change, either changes in the environment or the anatomical changes an animal might have gone through when evolving the actual trait. Here it is also a matter of knowing what the acceptable school-scientific vocabulary is, especially the distinction between a description (category a) and an explanation (which is the basis for category b, c, and

d). Teleological or anthropomorphic explanations are put together in category b), and here the answers focus purpose (for example *in order to*). The explanations in category c) and d) rely on the sense that students make of the key terms heredity, variation, and selection. Explanations in category c) mainly deal with proximate causes and only make use of some of the key terms, often interspersed with some scientific terms (mostly “genetic words”). It is a mix where students’ growing understanding and mimicking of the scientific language is used when formulating answers. Natural selection is the basis of the fourth category (d), but in various steps, from only mentioning differential survival to differential reproduction and further to accumulation of a trait/gene.

### Analysing answers about the origin of a new trait

These questions (multiple choice and an open ended request for justification) dealt with the origin of a new hereditary trait:

In the future, it is most likely that entirely new hereditary traits will develop among living organisms – traits that never existed before. What is the origin of an entirely new hereditary trait? Choose the statement that you consider is the best. Justify your choice.

- The individual’s need of the trait
- Random changes in the genes
- The specie’s pursue to develop
- In nature balance is pursued

The question is similar to a version from Wallin, Hagman and Olander (2001); however in this study the item was given both as a multiple-choice question, and accompanied with a request to give a reason for the choice. The alternative that is most in line with the scientific explanation is “Random changes in the genes.” A system of categories was generated in relation to the open ended task of justification (see Table 3 for more details). Two main categories was discerned, one type dealt with descriptions or explanations of *development in general*, and the other was based on *ultimate causes of new traits*. The latter type of answer refers more thoroughly to the heredity part of the task. The first type of answers used words like need, wish and/or adaptation, often referring to individual organisms.

## FINDINGS

### Experimental group versus national sample; evolution of a trait

The trait mentioned in the headline is the cheetahs’ ability to run fast. As mentioned before, a recalculation of all groups was made, in order not to overestimate the results of the experimental group. The number of students that are compared then decreases; in the experimental group grades 5 – 7 from 80 to 73, which means that 7 students didn’t answer. In the same way, answers from the experimental group grade 9 alter from 83 to 82, and the national sample from 620 to 335.

Table 1. Students’ post-test explanations of how cheetahs’ ability to run fast has evolved. All differences between groups are significant ( $p < 0,01$ ).

Category of student response	Experimental group		National sample
	Grade 5 – 7 n = 73	Grade 9 n = 82	Grade 9 n = 335
a) describes but do not explain	14 %	9 %	38 %
b) explains in a teleological way	49 %	25 %	38 %
c) explains only with some key terms	21 %	16 %	10 %
d) explains in terms of natural selection	15 %	50 %	14 %

In the grade 9 experimental group, half of the students answered in terms of natural selection. In the Swedish national evaluation, 14 % answered in a similar way. The experimental groups in grades 5-7 gave answers in the category “natural selection” with an average of 15 % (see Table 1). In the national sample, 76 % of the students who gave answers use a type of reasoning that is not in line with the scientific view. In the experimental group of the same age, the proportion of non-scientific answers is 34 %. Even the younger students, in grade 5 – 7, answer in a more scientific way than grade 9 students in the national sample.

**Experimental group versus national sample: origin of a new trait**

The proportion of “no answer” in the multiple-choice question (Table 2) is the same in all groups. The alternative most in line with the scientific view is “Random changes in the genes”. This was also the alternative that was chosen most frequently in all groups, but the experimental group (all ages) chose the correct answer significantly more often than the students in the national sample.

*Table 2. Students’ post-test choice of the origin of a new trait. The difference between the national sample and the experimental grade 9 is significant (p < 0,01); no significant differences between the other groups.*

Multiple choice alternative	Experimental group		National sample
	Grade 5 – 7 n = 80	Grade 9 n = 83	Grade 9 n = 620
1) The individual’s need of the trait	29 %	27 %	24 %
2) Random changes in the genes	45 %	57 %	31 %
3) The specie´s pursue to develop	13 %	11 %	30 %
4) In nature balance is pursued	8 %	0 %	8 %
No answer	6 %	6 %	6 %

The students’ written statements when asked to justify their choice of alternatives are shown in Table 3. Note that a recalculation of answers from all groups has been made; in the same way as for the cheetah-question. The percentage of students who explained the origin of the new trait (category b in Table 3) differs from the percentage of students who correctly answered the multiple choice item (alternative 2 in Table 2). In experimental group 5 – 7 the numbers decrease, mainly because many of the students did not write any justification. On the other hand, the percentage rates for the groups in grade 9 are higher when students motivate their choice. This increase is due to the fact that some students wrote a justification in line with a scientific view, despite choosing an alternative other than number two in the multiple-choice item. For example the answer that one student writes in spite of choosing alternative 1 “need”:

*If for example there is a quick and a slow zebra in a flock, the quick zebra will survive if a lion chases it. Then the genes of the quick zebra will be passed to a new generation. Eventually there will be more and more quick zebras because it helps them when they are threatened.*

This student chooses alternative 1 “need”, but writes a justification that is rather well in line with the scientific view of natural selection.

It is noteworthy that there is only a slight difference between the answers from the national sample and the experimental group in grade 5 – 7; they give similar answers in spite of the age difference. The experimental group in grade 9 writes answers more in line with current views in science with a greater frequency (a difference with 27 percentage points compared with the national sample).

Table 3. Students' post-test justifications of the origin of a new trait. Differences are significant ( $p < 0,01$ ) between the national sample and the experimental grade 9 and between experimental groups; no significant difference between national sample and experimental grade 5-7. Note that the main categories (a, b and c) are printed in bold.

Category of student response	Example of response	Experimental group		National sample
		Grade 5 - 7 n = 59	Grade 9 n = 82	Grade 9 n = 335
<b>a) Explains or describes the development, but not the origin of a new trait</b>		<b>65 %</b>	<b>38 %</b>	<b>59 %</b>
a1) describes development	<i>At first humans were apes (animals) and then in the course of time we have developed to become humans.</i>	24 %	11 %	10 %
a2) focuses on "need" as rationale	<i>You develop things that you have a great need of in order to survive.</i>	22 %	18 %	15 %
a3) focuses on "wish/strive" as rationale	<i>Everybody wants to develop and become better. Then maybe you develop if you try.</i>	14 %	5 %	21 %
a4) focuses on adaptation as rationale	<i>Animals adapt to the environment you/ they live in. Something like the birds on Galapagos.</i>	5 %	4 %	13 %
<b>b) Explains the origin of a new trait</b>		<b>35 %</b>	<b>63 %</b>	<b>36 %</b>
b1) explains how "changes in gene pool" originate	<i>Sometimes mutations happen in the genes, the DNA. In that way new traits originate.</i>	27 %	31 %	31 %
b2) specifies also mechanism of natural selection	<i>If an animal happens to get a random change, which gives that individual an advantage, it makes it live longer and it is able to spread its genes, which means that more and more individuals get that trait.</i>	8 %	32 %	5 %
<b>c) Other</b>		<b>0 %</b>	<b>0 %</b>	<b>5 %</b>
	<i>E.g. a mole that lives on worms has to dig in soil; it was therefore made with feet that are easy to dig with.</i>			

### Experimental group versus national sample: consistency, aggregated findings

Presented here are the three questions, which were identically formulated and analysed for the national sample and the grade nine experimental group. Note that a recalculation of answers from both groups has been made in order to achieve fairer comparisons. A rough division into two parts of every question is made: answers in line or not in line with the theory of evolution. In line with theory are answers from Table 1 in category c and d, from Table 2 alternative 2, and from Table 3 category b. These answers have been aggregated as shown in Table 4. The experimental group's answers are significantly more in line with the theory of evolution than those of the national sample.

Table 4. Consistency among three written answers; national sample and grade nine experimental group. Differences between the groups are significant ( $p < 0,01$ ).

	No answer in line with theory	One or two answers in line with theory	Three answers in line with theory
National sample	59 %	32 %	9 %
Experimental group	16 %	41 %	43 %

### Experimental group before and after teaching: evolution of a trait

In the pre-test, the students were asked to explain how a trait had evolved:

Seals can remain underwater without breathing for nearly 45 minutes as they hunt for fish. How would a biologist explain how the ability to not breathe for long periods of time has evolved, assuming their ancestors could stay underwater for just a couple of minutes? (Settlage, 1994)

Answers to the seal-question were compared to the answers given at the post-test about the evolution of a trait in cheetahs. There may be a problem of comparison when you change the question. Students have different experience and knowledge about the subject in each question, in this case seals and cheetahs. However, the conclusion from pilot studies was that the change of species had little influence as long as you kept a typical trait of the species in question as invariant. The advantage of changing, for example by presenting a new question, would make the comparison with the national sample fairer, since the cheetah-question would be a novel problem to both groups.

The students did improve their explanations in terms of using more scientific language (see Table 5). This occurred more with the grade 9 students than with the younger ones. There are few differences between grades 5-7 and grade 9 concerning the pre-test results and the percentages are nearly the same for the categories c and d, which are the categories with at least some key terms.

Table 5. Students' explanations of how a trait evolved: findings within the experimental group. Differences between both groups' pre- and post-test results are significant ( $p < 0,01$ ); no significant differences between the groups' pre-test results.

Category of student response	Grade 5-7		Grade 9	
	Pre-test n = 81	Post-test n = 80	Pre-test n = 87	Post-test n = 83
a) describes but do not explain	43 %	12 %	36 %	8 %
b) explains in a teleological way	38 %	45 %	46 %	25 %
c) explains only with some key terms	4 %	19 %	5 %	16 %
d) explains in terms of natural selection	1 %	14 %	2 %	49 %
e) not answered	6 %	9 %	10 %	1 %
f) other	7 %	1 %	1 %	0 %

### DISCUSSION

In summary, this study's findings about learning outcomes show that students, who participated in the teaching intervention, did improve their answers to written questions. Students in the experimental group answered significantly more in line with set learning goals than a national sample. Within the experimental group, students' answers were significantly more in line with goals three months after teaching than before. Both these forms of evaluation, external and internal, are dependent of how the goals are interpreted and the quality of the assessment instrument. The fact

that students do improve indicates that the questions are in alignment with the goals and that the teaching enhances this improvement. This article has described and argued the steps taken from national aims to explicit learning goals and further to the choice of test questions. The questions used in this study have the quality to offer the students latitude, and possibility to improve the quality of reasoning, when answering them.

When researchers construct specific questions that are to be used for assessing conceptual learning, the context, and especially the change of context, is important to be aware of. Assessments try to find signs of learning and the students' ability to use similar reasoning in different contexts is such a sign. In this study the specie varied, for example seals and cheetahs, while the evolution of a typical trait of that specie was kept as invariant, namely the ability of prolonged diving, and fast running, respectively. A similar shift in what is variant/invariant is made by Asterhan and Schwarz (2007), who also found that the suggested approach did not alter the questions' validity. However, my point is that when constructing context shifts it is important to keep track of what is invariant and what is not. Another methodological reason for change of context was to get a more fair comparison between the experimental and the national sample. In the experimental design the question about cheetahs was not introduced until the delayed post test. The assumption was that by this the probability increases that at least the context would be equally new to both the experimental and the national sample.

Studies about teaching interventions often report improved learning outcomes estimated with external evaluations. Methodologically, external evaluations in educational settings are hard to perform. The usual procedure with control groups, which gets ordinary teaching, has (far too) many factors that are difficult to keep invariant (Bach, 2001; Juuti & Lavonen, 2006; Millar, Leach, Osborne & Ratcliffe, 2006). For example, what is ordinary teaching and what constitutes a comparable group of students; not to mention the teacher's influence, even if it is the same teacher in both groups. The external evaluation in this study was framed in relation to a national evaluation in Sweden. Partly this frame was facilitated since the researchers that performed the national evaluation also took part in the intervention study. This connection increases the reliability, when sorting students' written answers in different categories.

The purpose of the introduced systems of categories was to reflect different qualities in reasoning. In this way it is a specification of the zone of proximal development (Vygotsky, 1978) in relation to explaining different phenomena, thus pointing at what is likely to be the students' reasoning. Apart from being a concrete analytical tool when assessing written answers, such as the ones presented in this paper, the systems of categories can inform the planning of teaching; especially when it comes to teaching about the nature of science (NOS), and what is regarded as appropriate answers in science classrooms. The system of categories points at areas where teaching activities have to be designed. Such activities are, for example, those which elaborate differences between descriptions and explanations in science, which is a general issue for all science subjects. Two other distinctions in the systems of categories are more specific when it comes to explaining biology. Explanations could be either teleological or causal, and they could deal with proximate or ultimate causation. All these three considerations (description/explanation, teleology/causality, and proximate/ultimate) need some kind of activity and/or talk, if the quality of students' reasoning is to improve. Without explicit attention, these distinctions are not likely to be noticed by the majority of students. A straight forward suggestion is to let students answer a question, make answers anonymous and list them all; then let students do the categorisation with the help of the system of categories.

The systems of categories also have the potential of serving as facilitator when teachers perform assessment for learning (Black, Harrison, Lee, Marshall & Wiliam, 2003). An example is when the teachers lead whole class discussions and have to give appropriate feedback that makes progress in the scientific story. All oral contributions from the students are to be taken seriously. However, the teacher should be sensitive to comments that touch upon a new step in the system of categories.

For instance, if suggestions about how selection works come up, it could be wise to ask students to elaborate about the outcome of different survival rates and reproductions rates. The didactical analysis hypothesises that the aspect of existing variation is essential to discern; thus feedback from the teacher about variation should be prominent. Then feedback comments have the potential of turning into real “feed forward” comments.

Concerning the actual teaching practice of evolution of life in Swedish compulsory schools, the national evaluation is a point of reference. Two striking features in the evaluation were the low attainment of national goals and the large proportion of students who chose not to answer. The researchers who wrote the report (National Agency of Education, 2004) speculate that although the curriculum since 1994 has emphasised evolution, seemingly teaching practice does not. This conclusion is supported by the present study since students, irrespective of their age, gave similar answers in the pre-test. If this is a reflection of current teaching practice, it shows few signs of the impact of previous teaching of evolution.

In the present study the ambition was to alter teaching practice in especially two ways. Firstly, to unfold the scientific story by using the key terms heredity, variation, and selection. These terms were supposed to be treated both separately and as part of a coherent theory. Secondly, to give plenty of opportunities for students to engage in productive talk using key terms as tools for reasoning. It would be interesting to explore the relationship between teaching approach and learning outcome. Further analysis of communication patterns in the actual classrooms could be an important way of deepening the understanding.

This study has contributed with a suggestion for a change of practice regarding teaching evolution, and it is shown that the majority of students made considerable improvements towards the set learning goals of the project. But after teaching, there is still one of six students (16 %) in the experimental group grade 9, who does not answer in line with the set learning goals. This is estimated by written answers three months after teaching. There are several possible reasons for why students do not answer as expected. Many of these reasons are common to all assessments that are free of the external motivation of grading. Students possibly articulate a doubt about the scientific account or simply do not make the effort of answering. However, this study is part of an iterative design based research effort, where continuous improvements in the design are to be made if more (all) students should reach the set learning goals. The change of practice has to be continued.

## ACKNOWLEDGEMENTS

The research reported in this paper was funded by the Swedish Research Council (VR). When preparing the paper I had valuable discussions with Anita Wallin, Ann Zetterqvist, Björn Andersson, Frank Bach and Åke Ingerman; and I also acknowledge the insightful comments from two anonymous reviewers.

## REFERENCES

- Abrams, E., Southerland, S., & Cummins, C. (2001). The how's and why's of biological change: how learners neglect physical mechanism in their search for meaning. *International Journal of Science Education*, 23(12), 1271-1281.
- Andersson, B., Bach, F., Hagman, M., Olander, C., & Wallin, A. (2005). Discussing a research programme for the improvement of science teaching. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhof (Eds.), *Research and the quality of science education* (pp. 221-230). Dordrecht: Springer.
- Ariew, A. (2003). Ernst Mayr's 'ultimate/proximate' distinction reconsidered and reconstructed. *Biology and Philosophy*, 18, 553-565.

- Asterhan, S., & Schwarz, B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology*, 99(3), 626-639.
- Bach, F. (2001). *Om ljuset i tillvaron, ett undervisningsexperiment inom optik*. Göteborg studies in educational sciences 162. Göteborg: Acta Universitatis Gothoburgensis. English summary retrieved June 28, 2009, from <http://gupea.ub.gu.se/dspace/handle/2077/10514>
- Barab, S., & Squire, K. (2004). Design-based research: putting a stake in the ground. *The Journal of the Learning Sciences*, 13(1), 1-14.
- Bishop, B., & Anderson, C. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415-427.
- Bizzo, N. (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.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning – putting it into practice*. Open University Press
- The Design-Based Research Collective (2003). Design-based research: an emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5-8.
- Hoel, T. L. (1996). Kjønn, interaksjon og elevsamarbeid. *Nordisk Pedagogik*, 1(16), 31-41.
- Jensen, M. S., & Finley, F. N. (1996). Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *Journal of Research in Science Teaching*, 33(8), 879-900.
- Juuti, K. & Lavonen, J. (2006). Design-based research in science education: one step towards methodology. *Nordic Studies in Science Education*, 4, 54-68.
- Jiménez-Alexandre, M.P. (1992). Thinking about theories or thinking with theories?: A classroom study with natural selection. *International Journal of Science Education*, 14, 51-61.
- Kampourakis, K., & Zogza, V. (2008). Students' intuitive explanations of the causes of homologies and adaptations. *Science & Education*, 17, 27-47.
- Kattmann, U. (2008). Learning biology by means of anthropomorphic conceptions? In M. Hammann, M. Reiss, C. Boulter & S. Tunnicliffe (Eds.), *Biology in Context, Learning and teaching for the twenty-first century*. (pp. 7-17). Institute of Education, London.
- Kattmann, U., Duit, R., & Gropengießer, H. (1998). The model of educational reconstruction. Bringing together issues of scientific clarification and students' conceptions. In H. Bayrhuber & F. Brinkman (Eds.), *What – Why – How? Research in Didaktik of Biology* (pp. 253-262). ERIDOB, Kiel: IPN.
- Kelemen, D. (1999). Function, goals and intention: children's teleological reasoning about objects. *Trends in Cognitive Sciences*, 3(12), 461-467.
- Kelly, A. (2003). Theme issue: the role of design in educational research. *Educational Researcher*, 32(1), 3-4.
- Leach, J., & Scott, P. (2002). Designing and evaluating science teaching sequences: An approach drawing upon the concept of learning demand and a social constructivist perspective on learning. *Studies in Science Education*, 38, 115-142.
- Lijnse, P. (1995). Developmental research' as a way to an empirically based 'didactical structure' of science. *Science Education*, 79(2), 189-199.
- Lindwall, O. (2008). *Lab work in science education: instruction, inscription, and the practical achievement of understanding*. Ph D thesis. Linköping Studies in Art and Science. No. 426. Department of Theme research, Linköping.
- Lo, M., Marton, F., Pang M., & Pong, W. (2004). Toward a pedagogy of learning. In F. Marton, & A. B. M. Tsui (Eds.) *Classroom discourse and the space of learning*. (pp. 189-225). NJ: Lawrence Erlbaum.
- Mayr, E. (1961). Cause and effect. *Science*, 134, 1501-1506.
- Méheut, M., & Psillos, D. (2004). Teaching– learning sequences. Aims and tools for science education. *International Journal of Science Education*, 26(5), 515-535.

- Millar, R., Leach, J., Osborne, J., & Ratcliffe, M. (2006). *Improving subject teaching - lessons from research in science education*. London and New York: Routledge.
- Mortimer, E. & Scott, P. (2003). *Meaning making in secondary science classrooms*. Maidenhead, Philadelphia: Open University Press.
- National Agency of Education (2004). *National evaluation of the compulsory school in 2003*. Retrieved June 28, 2009, from [www.skolverket.se/sb/d/663/a/2308](http://www.skolverket.se/sb/d/663/a/2308)
- National Agency of Education (2000). *Compulsory schools syllabuses 2000*. Retrieved June 28, 2009, from [www3.skolverket.se/ki/eng/comp.pdf](http://www3.skolverket.se/ki/eng/comp.pdf)
- Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining science in the classroom*. Buckingham: Open University Press
- Sandoval, W. A., & Bell, P. (2004). Design-Based Research Methods for Studying Learning in Context: Introduction, *Educational Psychologist*, 39(4), 199-201.
- Sandoval, W., & Millwood, K. (2005). The quality of students' use of evidence in written explanations. *Cognition and Instruction*, 23(1), 23-55.
- Settlage, J. (1994). Conceptions of natural selection: a snapshot of the sense-making process. *Journal of Research in Science Teaching*, 31(5), 449-457.
- Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, 52(2), 170-194
- Stearns, S., & Hoekstra, R. (2000). *Evolution: an introduction*. Oxford University Press.
- Tiberghien, A. (2000). Designing teaching situations in the secondary school. In R. Millar, J. Leach & J. Osborne (Eds), *Improving science education. The contribution of research*, (pp. 27 - 47). Buckingham: Open University press.
- Thomas, J. (2000). Learning about genes and evolution through formal and informal education. *Studies in Science Education*, 35. 59-92.
- Vygotsky, L. (1978). *Mind in Society*. Cambridge, Harvard University Press.
- Wallin, A. (2004). *Evolutionsteorin i klassrummet. På väg mot en ämnesdidaktisk teori för undervisning i biologisk evolution*. Göteborg studies in educational sciences 212. Göteborg: Acta Universitatis Gothoburgensis. English summary retrieved June 28, 2009, from <http://gupea.ub.gu.se/dspace/handle/2077/9494>
- Wallin, A., Hagman, M., & Olander, C. (2001). Teaching and learning about the biological evolution: conceptual understanding before, during and after teaching. In *Proceedings of the III Conference of European Researchers in Didactic of Biology (ERIDOB)*, (pp.127-139). Universidade de Santiago de Compostela, Spain.
- Zetterqvist, A. (1998). Teachers views on their teaching of evolution. In H. Bayrhuber & F. Brinkman (Eds.), *What - Why - How? Research in Didaktik of Biology* (pp. 11-20). ERIDOB, Kiel: IPN.
- Zohar, A., & Ginossar, S. (1998). Lifting the taboo regarding teleology and anthropomorphism in biology education - heretical suggestions. *Science Education*, 82, 679-697.