Reasoning About Evolution’s Grand Patterns: College Students’ Understanding of the Tree of Life

Laura R. Novick
Vanderbilt University
Kefyn M. Catley
Western Carolina University

Tree thinking involves using cladograms, hierarchical diagrams depicting the evolutionary history of a set of taxa, to reason about evolutionary relationships and support inferences. Tree thinking is indispensable in modern science. College students’ tree-thinking skills were investigated using tree (much more common in professional biology) and ladder (somewhat more common in textbooks) cladogram formats. Students’ responses to questions assessing five tree-thinking skills provided evidence for several perceptual and conceptual factors that impact reasoning (e.g., the Gestalt principles of good continuation and spatial proximity, prior knowledge). Instructional implications of the results include using the tree format for initial instruction and clarifying that most recent common ancestry determines evolutionary relatedness. Broader implications for designing scientific diagrams and promoting diagrammatic literacy are considered.

KEYWORDS: Tree of Life, tree thinking, cladograms, evolution education, scientific reasoning

Diagrammatic Literacy in Science Education

Diagrams have a long history of use in science, dating back at least to the 15th century (Hegarty, Carpenter, & Just, 1991). Schematic (or relational) diagrams in particular, which depict the underlying structure of abstract...
concepts and rely on conventions for their use, are increasingly important tools for theory development, problem solving, and communication in virtually all fields of contemporary science (e.g., Hegarty et al., 1991; Hegarty & Stull, in press; Kindfield, 1993/1994; Lynch, 1990; Maienschein, 1991; Novick, 2006a). Consider, for example, introductory biology at the college level, as represented by the popular textbook, *Campbell Biology* (Reece et al., 2010). Schematic diagrams in this textbook include (but are not limited to) flow charts, food webs, gene maps, genetic similarity maps, process system diagrams, protein maps, and phylogenetic trees. Given the ubiquity of schematic diagrams throughout the sciences, it is clear that diagrammatic literacy underpins much of conceptual development in science education.

Because schematic diagrams rely on learned conventions, they do not transparently convey meaning (e.g., Cromley et al., 2010; Hegarty & Stull, in press; Yeh & McTigue, 2009). Thus, students must be taught the mapping between diagrammatic elements and conceptual meaning if they are to view these diagrams as more than just abstract designs composed of various graphic elements (Novick, 2006a). A key implication is that diagrammatic literacy cannot be taught in the abstract. Rather, it must be embedded within each area of science at all levels of the curriculum (e.g., Yeh & McTigue, 2009). In this article, we focus on students’ understanding of a schematic diagram from biology—a type of phylogenetic tree about which it has become increasingly clear that general citizens need a working knowledge as a key component of being scientifically literate.

The Tree of Life

Science is currently striving toward one of its grandest goals—assembling the *Tree of Life* (ToL), a schematic diagram depicting evolutionary relationships among all extant and extinct living things. The patterns of relationships produced by millions of years of evolution both inform theorizing regarding the processes that produced those patterns and support inferences concerning that which is unknown. Even in its current, incomplete state, the ToL is yielding considerable benefits to humankind in areas such as health, agriculture, and biotechnology (e.g., American Museum of Natural History [AMNH], 2002; Futuyma, 2004; Yates, Salazar-Bravo, & Dragoo, 2004). We briefly cite just two of many possible examples. A new placement of unicellular microsporidians within the fungi branch is spurring development of more effective control of these pathogens in honeybee colonies (Freeman, 2011). These infections are a significant problem for world agriculture, and thus for combating malnutrition, as every third bite we eat is a direct result of honeybee pollination (Morse & Calderone, 2000). With respect to health, knowing the evolutionary relationships of species of Yew trees from the genus *Taxus* directed the harvesting and synthesis of the widely used cancer drug Taxol. Taking a broader perspective,
understanding the life support systems of our planet, which are provided by its biodiversity, depends on the ToL.

Given the numerous current global crises, assembling the ToL is arguably the most important undertaking of contemporary science. Clearly, a scientifically literate citizenry prepared to understand and tackle 21st-century issues will need at least a basic understanding of the science behind the ToL. Thus, it should come as no surprise that science educators advocate including tree thinking, a component of diagrammatic literacy, in college and high school biology curricula (e.g., Baum, Smith, & Donovan, 2005; Catley, 2006; Catley, Lehrer, & Reiser, 2005; Goldsmith, 2003; Gregory, 2008; O’Hara, 1988). Tree thinking refers to the skills needed to interpret, reason about, and apply information about evolutionary relationships depicted in ToL diagrams. Because these diagrams use spatial relations to convey evolutionary meaning, they are schematic diagrams. The present study provides the first in-depth look at college students’ baseline level of competence at five tree-thinking skills, each representing an authentic skill used by evolutionary biologists.

An Introduction to Macroevolution and Cladograms

Microevolution Versus Macroevolution

Biologists commonly distinguish micro- and macroevolution. Microevolution concerns processes, such as natural selection, that occur at the level of the organism (i.e., genome, individual, and population). Microevolution explains, for example, the proliferation of antibiotic-resistant bacteria. Macroevolution, in contrast, concerns processes that occur at the level of species (e.g., blue whale) and above (e.g., cetaceans, mammals, vertebrates), resulting in the formation, radiation, and extinction of higher groups of taxa. A taxon (plural, taxa) is any taxonomic category ranging from species to higher-order groups. Macroevolution explains, for example, the origin and radiation of mammalian diversity. The ToL is a macroevolutionary concept.

Representing the Tree of Life

The most common diagrammatic depiction of the evolutionary relationships underpinning the ToL is the cladogram, a hierarchical diagram showing nested levels of common ancestry (e.g., see Figure 1). Cladograms are hypotheses about the evolutionary history of nested sets of taxa that are supported by special evolutionary characters called synapomorphies (Hennig, 1966; Thanukos, 2009). Synapomorphies are derived (i.e., newly evolved), invariant (therefore informative) physical, molecular, or behavioral characteristics shared by all members of a group (e.g., a multichambered stomach, certain DNA sequence, or specific social behavior). For ease of comprehension, we will generally use the less precise but more familiar term character,
by which we will mean synapomorphy as defined here. One character shared by whales and giraffes (and other taxa) is a multichambered stomach (horses and manatees instead have a single-chambered stomach; see Figure 1). That is, whales and giraffes have a most recent common ancestor (MRCA) that evolved a multichambered stomach, a novel character not found in any other taxon. Thus, contrary to outward appearance, whales are more closely related to giraffes and manatees are more closely related to horses than whales and manatees and giraffes and horses are to each other.

A group of taxa consisting of an MRCA and all its descendants is called a clade. In Figure 1, cows and whales comprise a clade. Because of the

---

**P1.** What character was possessed by the most recent common ancestor of cows and camels?  →  *answer = even number of toes*

**P2.** List all taxa (might be one or more) that evolved from an ancestor that had a single-chambered stomach.  →  *answer = horse, tapir, aardvark, rock hyrax, manatee, elephant*

**P3.** Do the bracketed taxa labeled “A” constitute a clade?  →  *yes, no → answer = yes*

What evidence supports your answer?

If you answered no, which taxa need to be removed and/or added to the group to make it a clade? (Make sure you indicate whether the taxa you list should be added or removed.)  →  *not applicable*

---

**Figure 1.** The tree format cladogram and associated questions discussed in this article (with the answers printed in italics) for 10 placental mammal taxa.
nesting of sets of taxa in a cladogram, giraffes, cows, and whales also are a clade in this figure. The molecular character supporting the cow plus whale clade distinguishes the MRCA of cows and whales from the earlier ancestor of giraffes, cows, and whales, supported by the character multi-chambered stomach in Figure 1. Because a clade comprises all and only those taxa with a shared evolutionary history, clades are the building blocks for reconstructing and understanding the history of life on Earth.

An Introduction to Five Core Components of Tree Thinking

The second author identified five core tree-thinking skills, presented in Table 1, based on his expertise in this area (e.g., Catley, 1994).\(^1\) Two skills assessed basic comprehension in terms of reading information off a cladogram. Skill I involved identifying a character shared by taxa due to inheritance from their MRCA (i.e., the synapomorphy that marks the MRCA of these taxa)

<table>
<thead>
<tr>
<th>Tree-Thinking Skill Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Identify the character shared by two or more taxa due to inheritance from their MRCA (i.e., the synapomorphy that marks the MRCA of these taxa)</td>
</tr>
<tr>
<td>A. Identify the synapomorphy for two taxa that comprise a clade</td>
</tr>
<tr>
<td>B. Identify the synapomorphy that supports the MRCA of more distantly related taxa</td>
</tr>
<tr>
<td>C. Map a verbal representation of nested structure onto the cladogram to identify the synapomorphy that supports the MRCA of 3-4 distantly-related taxa</td>
</tr>
<tr>
<td>II. Identify a set of taxa based on character information (i.e., a synapomorphy) provided</td>
</tr>
<tr>
<td>A. Identify taxa that evolved from a taxon with a specified character</td>
</tr>
<tr>
<td>B. Identify taxa that did not evolve from a taxon with a specified character</td>
</tr>
<tr>
<td>III. Understand the concept of a clade</td>
</tr>
<tr>
<td>A. Recognize that bracketed taxa do not comprise a clade</td>
</tr>
<tr>
<td>1. Taxa located at different levels</td>
</tr>
<tr>
<td>2. Taxa are a subset of a polytomy</td>
</tr>
<tr>
<td>B. Recognize that bracketed taxa comprise a clade</td>
</tr>
<tr>
<td>C. Identify a subset of taxa that comprise a clade</td>
</tr>
<tr>
<td>IV. Evaluate relative evolutionary relatedness</td>
</tr>
<tr>
<td>A. Reference taxon is the most derived of the three taxa</td>
</tr>
<tr>
<td>B. Reference taxon is at an intermediate level between the remaining two taxa</td>
</tr>
<tr>
<td>V. Use evidence of most recent common ancestry to support an inference regarding a shared character</td>
</tr>
</tbody>
</table>

Note. MRCA = most recent common ancestor.
Skill II involved identifying a set of taxa given a certain character. Each such character divides the taxa into two groups—those that share versus do not share the MRCA for which that character is a marker. Although in everyday situations two sets defined in this complementary manner are equally good groups (e.g., children in a class who are girls vs. not), this is not true with respect to macroevolution. The taxa that share an MRCA form a clade and are thus considered a natural group (i.e., one that shared a common history). Recovering the history of life on Earth, a key goal of evolutionary biology, involves identifying such natural groups. Before one can investigate whether students understand the lack of equivalence between sets of taxa identified by descent or not from a certain ancestor, it is necessary to determine whether they can identify the relevant sets of taxa. For example, for the cladogram in Figure 1, students were asked to list the taxa that evolved from an ancestor with a single-chambered stomach.

Two additional skills assessed understanding of the critical relational information depicted in cladograms. Skill III involved understanding the concept of clades. As noted earlier, clades are the fundamental building blocks of the ToL. For example, students were given a definition of a clade and were then asked whether the bracketed taxa labeled “A” (see Figure 1) comprise a clade. Skill IV involved evaluating relative evolutionary relatedness, which is arguably the quintessential tree-thinking skill as the purpose of cladograms is to depict evolutionary relationships. Again using Figure 1 as an example, students could be asked whether camels or whales are more closely related to giraffes. The correct answer is whales because they share a more recent common ancestor with giraffes than do camels.

Finally, skill V required students to make an inference about which taxon is most likely to share a character possessed by a certain other taxon. Students with a high level of diagrammatic literacy are able to apply the given information to new situations. Consider the following inference question for Figure 1: Given that manatees have a circumferential placenta, are whales or elephants more likely to also have this character? Without the benefit of the ToL, one would likely guess whales because whales and manatees are both aquatic mammals, sharing a similar body shape and habitat. The portion of the Tree depicted in Figure 1, however, shows that manatees are more closely related to elephants than to whales. Thus, it is actually much more likely that elephants rather than whales share this character with manatees.

Cladograms in Educational and Professional Practice

We are not aware of any empirical studies examining how cladograms and tree thinking are typically taught in introductory biology classes. However, Catley and Novick (2008) have documented that cladograms are common in textbooks, especially at the college level. Of the 111 evolutionary diagrams found in 9 high school biology texts they reviewed, 53% are
cladograms. Of the 313 such diagrams in 12 college introductory biology texts (6 each for nonmajors and majors), 77% are cladograms. On average, the high school, college nonmajors, and college majors texts have 7, 11, and 29 cladograms per textbook, respectively. It seems reasonable to assume that many (perhaps most) college instructors use the diagrams provided digitally with their textbooks in their lectures. In the current edition of *Campbell Biology* (Reece et al., 2010), a widely used textbook for majors, several cladograms appear prior to the chapter that provides information on phylogenetics and explains how to read trees (Chapter 26 of 56, which of course might or might not be assigned in a particular class). Cladograms then appear more extensively in subsequent chapters to explore and illustrate biodiversity (evolutionary relationships among groups) and to situate each new group of taxa, as it is introduced, in evolutionary and taxonomic space.

Cladograms are typically presented in two formats that we have referred to previously as trees and ladders (Catley & Novick, 2008; Novick & Catley, 2007). Figures 1 through 4 provide examples of the tree format; Figure 5 provides examples of the ladder format. The high school and college textbooks surveyed by Catley and Novick (2008) include cladograms in both formats, with slightly more ladders than trees at both levels. In contrast, the professional evolutionary biology literature is skewed in the opposite direction, with a strong preference for the tree format over the ladder. Novick and Catley (2007) found this in their analysis of the articles in the 2005 issues of *Systematic Biology*. K. Quillin (personal communication, May 27, 2008) found the same result from her broader analysis of articles in 11 biology journals from late 2007 through mid 2008.

The inclusion of cladograms in textbooks is consistent with the science education standard that students should learn to use the methodologies and tools from professional practice (American Association for the Advancement of Science [AAAS], 2001; National Research Council [NRC], 1996). The preference for the ladder over the tree format, however, is not. Given the prevalence of both formats in high school and college textbooks, as well as the discrepancy in their relative frequencies in textbooks versus professional practice, it is critical to determine whether students are equally successful at reasoning from information provided in both diagrammatic formats.

**Research on Tree Thinking**

Despite the ubiquity of cladograms in textbooks, we know of only three research articles evaluating the diagrammatic literacy skill of tree thinking. Those studies all have limitations for the goal of understanding student competence in this area. Novick and Catley (2007) used a diagram translation task; although translation is an important aspect of diagrammatic literacy generally (e.g., Kozma & Russell, 1997; Novick, 2004), it is not at the core of tree thinking. Meir, Perry, Herron, and Kingsolver (2007) and Sandvik
College Students’ Understanding of the Tree of Life

(2008) studied more central tree-thinking skills but (a) tested only relatively knowledgeable students who had taken several college biology courses or had received instruction in the scientific principles underlying cladograms, (b) used only ladder cladograms, and (c) provided primarily descriptive results. Sandvik’s results are further limited by his very small sample.

The most important finding from these studies, consistent across all three articles, is that students (even those with stronger biology backgrounds) had considerable difficulty understanding and reasoning about evolutionary relationships depicted in the ladder format. Novick and Catley’s (2007) results indicated that college students understood much better the nested structure of the tree than the ladder format. Because ladders are somewhat more prevalent in high school and college biology textbooks (Catley & Novick, 2008), these results suggest that evolutionary relationships are preferentially depicted in the format that is harder for students to understand!

One cause of students’ difficulty in understanding the structure of the ladder format is the Gestalt principle of good continuation (e.g., Hochberg, 1978), which states that a continuous (straight or curved) line represents a single entity. As discussed by Novick and Catley (2007), this general perceptual principal suggests an interpretation of ladders that conflicts with the correct segregation of the taxa into nested levels of common ancestry. In particular, it implies that the long sloped line at the base of a ladder has a single interpretation—namely, represents a single level—whereas in fact it invariably has multiple interpretations (i.e., represents multiple levels).

Meir et al. (2007) suggested two inappropriate reasoning strategies as additional sources of students’ errors in evaluating the degree of evolutionary relatedness among taxa depicted in a ladder: reliance on (a) the relative distance in horizontal space between the taxa (also see Baum et al., 2005) or (b) the number of nodes encountered in tracing a path between two taxa.

Overview of the Present Research

Goldstone and Barsalou (1998) have argued that cognition and perception are highly intertwined. This is especially likely to be the case with regard to students’ interpretations of diagrammatic representations (i.e., their diagrammatic literacy), as diagrams present both perceptual and conceptual information, which then must be integrated with the conceptual information provided in the questions. We predicted, therefore, that both conceptual and perceptual factors would influence students’ assessed competence at tree thinking.

Conceptual Factors That May Affect Tree Thinking

Prior Knowledge

One conceptual factor that should be critically important is prior knowledge. We examined the influence of prior knowledge in two senses. First, we
split our sample into two groups based on students' biology knowledge, operationalized in terms of the number of (primarily organismal) college biology courses they had taken. Obviously, we predicted better performance for students with stronger as opposed to weaker backgrounds in biology. The more interesting question was whether this factor interacted with any of our other manipulations.

A second sense of prior knowledge concerns knowledge about the particular taxa shown in the cladograms. Although students were told to use the information in the diagrams to answer our questions, they may nevertheless bring to bear information learned previously and retrieved from memory. This may especially be the case for the inference questions (skill V) because the character about which students were asked to make an inference (e.g., circumferential placenta for our example question for Figure 1) was not depicted on the cladogram.

Positive Versus Negative Wording of the Question

The skill II questions asked students to identify sets of taxa that did or did not descend from an ancestor with a certain character. This required varying whether the question was worded positively (Which taxa evolved from an ancestor with character C?) versus negatively (Which taxa did not evolve from an ancestor with character C?). Psycholinguistic studies have found that negatively worded statements and questions are harder to understand than positively worded ones. This is because understanding a sentence with a negation (e.g., The book is not closed.) requires readers to represent (a) the negated state of affairs, (b) the fact that this state does not hold in the world under consideration, and (c) the actual state of affairs (e.g., Kaup, Lüdtke, & Zwaan, 2006; Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007). Understanding the corresponding sentence without negation (The book is open.), in contrast, only requires representing the actual state of affairs.

Perceptual Factors That May Affect Tree Thinking

Good Continuation

As discussed earlier, Novick and Catley (2007) found that college students' understanding of the nested structure of ladders, as assessed by their ability to translate cladograms from one format to another, is negatively impacted by the Gestalt perceptual principle of good continuation. Thus, cladogram format can be viewed as a perceptual factor. In the present study, we examined whether good continuation also affects students' reasoning about relationships by comparing success at tree thinking for cladograms drawn in the ladder versus the tree format. We predicted better diagrammatic understanding for the tree than the ladder format. How good continuation is
predicted to affect students’ responses to particular tree-thinking questions is discussed in the context of the specific questions.

**Spatial Proximity**

The Gestalt principle of spatial proximity may affect how nonexperts interpret both tree and ladder cladograms. According to this principle (e.g., Hochberg, 1978), items in a visual display that are near each other are grouped together and considered to belong to a coherent set, separate from other items that are less spatially proximate. However, spatial proximity is an imperfect cue to relatedness. For example, Landy and Goldstone (2007) found that college students used spatial proximity to group terms in arithmetic statements such as $t \ast o^{f} \ast d = o \ast t\ast d \ast f$, and when this perceptual information conflicted with the mathematical order of operations (as in the example), their ability to judge the validity of these statements was impaired.

Spatial proximity, in combination with students’ understanding of hierarchical relationships, leads to predictions for students’ success at answering our evolutionary relatedness questions (skill IV). These questions asked which of two taxa is the closest evolutionary relation to a third (reference) taxon. Consider koala, Tasmanian devil, and bandicoot in Figure 2: Koala is most derived (evolved most recently), Tasmanian devil is least derived (evolved least recently, is most ancient), and bandicoot is at an intermediate level. We manipulated whether the reference taxon was in the most derived or intermediate position. (The latter question type is similar to the questions asked by Meir et al., 2007, and Sandvik, 2008, but only about the ladder format.) In addition, as in the present example, the taxa were selected so that when the reference taxon was in the intermediate position (bandicoot), it was more spatially proximate to the less closely related taxon (Tasmanian devil) in terms of both horizontal and vertical distance.

We therefore predicted better performance when the reference taxon was most highly derived because the question can be answered simply by moving down levels in the cladogram until reaching the intersection with one of the queried taxa. The first such intersection indicates the answer because taxa that share a more recent common ancestor are more closely related than taxa whose MRCA is more ancient. College students should be able to leverage their basic diagrammatic literacy with respect to hierarchies (e.g., Novick, 2001, 2006b) to answer these questions. Relying on spatial proximity also yields the correct answer in this case. In contrast, when the reference taxon is at an intermediate level, students have to understand both how evolutionary relationship is mapped onto levels due to most recent common ancestry and the irrelevance of the relative number of steps up versus down in the hierarchy from the middle taxon (i.e., spatial proximity). Such knowledge requires an understanding of the conventions for conveying evolutionary information in cladograms.
M1. What character was possessed by the most recent common ancestor of honey possums and brush tailed possums? \( \rightarrow \text{answer} = \text{prehensile tail} \)

M2. Which taxon—koalas or bandicoots—is the closest evolutionary relation to brush tailed possums? \( \rightarrow \text{answer} = \text{koalas} \)
What evidence supports your answer?

M3. Do the bracketed taxa labeled “A” constitute a clade? \( \text{yes no} \rightarrow \text{answer} = \text{no} \)
What evidence supports your answer?
If you answered no, which taxa need to be removed and/or added to the group to make it a clade? (Make sure you indicate whether the taxa you list should be added or removed.) \( \rightarrow \text{answer} = \text{add koala, kangaroo, honey possum, and brush tailed possum} \)

Figure 2. The tree format cladogram and associated questions discussed in this article (with the answers printed in italics) for seven marsupial mammal taxa.

Research Aims

To summarize, this study assessed five core tree-thinking skills, an aspect of diagrammatic literacy in science, in college students. We sought to determine the extent to which tree-thinking ability is affected by three conceptual factors—biology background, prior knowledge of specific taxa, and positive versus negative question wording—and two perceptual
factors—good continuation (operationalized in terms of cladogram format) and spatial proximity.

Method

Participants

The participants were 107 students at a private, Research I university in the Southeastern United States. Of these students, 94 voluntarily participated in partial fulfillment of requirements for introductory psychology or for extra credit in a psychology, education, or upper-level biology class. The remaining 13 students completed our task as a requirement for their zoology class.

Participants completed a background questionnaire on which they were asked whether they had taken any of 10 college classes in (primarily organismal) biology. As expected, students who had taken more of these courses did better on our questions than those who had taken fewer courses. However, the relationship was dichotomous rather than continuous. For most variables, the dividing point was between two and three classes—namely, between those who had gone beyond the year-long introductory class for biology and pre-med majors and those who had not. Students who had taken zero to two classes ($M = .59; n = 78$; 61 females, 16 males, 1 undisclosed sex) formed a relatively homogenous group, as did those who had taken three to seven classes ($M = 4.52; n = 29$; 18 females, 10 males, 1 undisclosed sex). We therefore divided our sample into those two groups for statistical analysis. For simplicity, we refer to these groups as having weaker and stronger backgrounds in biology, respectively. The academic majors of the weaker background group primarily were in education (42%) and psychology (26%). The stronger background students primarily had biology-related majors (93%). We did not ask about students’ acceptance of evolution, so determining whether that factor influences tree thinking must await future research.

Design and Materials

Overview

Biology background was a between-subjects factor. Cladogram format was a within-subjects factor as all subjects received four cladograms in each of the tree and ladder formats. Each cladogram was accompanied by four questions, with comparable questions being asked for each cladogram format. Because no existing data provided guidance as to the best way to assess these skills, we wrote several types of questions to assess many of the skills. All question manipulations were implemented within subjects for both cladogram formats.

The eight cladograms with their questions were printed on separate sheets of paper, which were collated in four orders. Several constraints
were placed on these orders to help ensure that students' responses to later questions were relatively independent of their earlier responses. An instruction page told participants to do their best even if the material was unfamiliar, to complete the pages in order, and not to return to previously completed pages. This page also gave definitions of two terms needed to understand the questions: (a) “a taxon (plural taxa) refers to any taxonomic category” and (b) “a clade is a group that includes the most recent common ancestor of the group and all descendants of that ancestor.” These definitions were reprinted on a half sheet of paper that participants had in view at all times because they were not allowed to turn back in the booklet.

**Cladograms**

The four tree-format cladograms depicted evolutionary relationships among marsupials, arthropods, insects, and placental mammals. The ladders showed relationships among vertebrates, spiders, dinosaurs, and fish. For each format, these cladograms included, respectively, 7 taxa and 10 characters, 9 taxa and 11 characters, 10 taxa and 11 characters, and 10 taxa and 12 characters. The taxa were chosen to represent animal diversity and for their relative familiarity. All depicted relationships are supported by scientific evidence. Where possible, the topologies were based on information provided in the Tree of Life Web project (http://tolweb.org/tree/).

**Tree-Thinking Skills**

In addition to the five tree-thinking skills outlined earlier, we tried to assess two other skills: identifying the sequential order of appearance of characters on a certain evolutionary path and recognizing and explaining an instance of convergent evolution. We will not discuss these questions because they turned out to be poorly written. Even the stronger background students failed to understand the evolutionary sequence questions (e.g., there was no correlation between number of biology courses taken and accuracy on these questions, in contrast to the results for the other questions). The convergent evolution questions were poorly worded for the present purpose because students could answer them without referring to the cladograms. Thus, as written, these questions did not assess tree thinking, which is the topic of this article.

Each of the five tree-thinking skills discussed here was represented by one to four questions, some with multiple parts. These skills and associated question types are listed in Table 1 and described in the sections that follow. The tree cladograms and questions are shown in Figures 1 through 4. Structurally identical questions were asked about the ladder cladograms. A portion of each ladder is shown in Figure 5.

**Identifying characters.** There were three types of identify character questions. One question asked for the character that supports a two-taxon
clade (skill I.A; see question M1 in Figure 2). A second question asked for the character supporting the MRCA of two more distantly related taxa (skill I.B; question P1 in Figure 1). The final question type involved mapping a verbal representation of nested levels onto a cladogram (skill I.C) to identify the character that supports the MRCA of three or four distantly related taxa (questions A1 in Figure 3 and I1 in Figure 4).

Identifying taxa. There were two types of identify taxa questions. These questions asked students to identify the taxa that share or do not share a particular character (skills II.A and II.B, respectively; questions P2 in Figure 1 and A2 in Figure 3, respectively).

Determining clade status. There were four types of clade questions. One question asked students to identify a subset of taxa that comprise a clade (skill I.A; see question M1 in Figure 2). A second question asked for the character supporting the MRCA of two more distantly related taxa (skill I.B; question P1 in Figure 1). The final question type involved mapping a verbal representation of nested levels onto a cladogram (skill I.C) to identify the character that supports the MRCA of three or four distantly related taxa (questions A1 in Figure 3 and I1 in Figure 4).

A1. Which character allows us to infer that ((stink bugs + beetles) + (lobsters)) is a clade? → \textit{answer = compound eyes}

A2. List all taxa (might be one or more) that did not evolve from an ancestor that had 1 pair of antennae. → \textit{answer = water bears, scorpions, spiders, ticks, and mites}

A3. On the diagram, circle the smallest group of taxa that includes spiders and comprises a clade. → \textit{answer = spider, ticks, and mites}

Figure 3. The tree format cladogram and associated questions discussed in this article (with the answers printed in italics) for nine arthropod taxa.
I1. Which character allows us to infer that \((\text{fly} + \text{flea}) + (\text{caddisfly} + \text{butterfly})\) is a clade? → answer = silk glands

I2. Which taxon—dragonfly or butterfly—is the closest evolutionary relation to the walking stick? → answer = butterfly
What evidence supports your answer?

I3. Do the bracketed taxa labeled “A” constitute a clade? yes no → answer = no
What evidence supports your answer?
If you answered no, which taxa need to be removed and/or added to the group to make it a clade? (Make sure you indicate whether the taxa you list should be added or removed.) → answer = add fly, flea, caddisfly, and butterfly

I4. Given that termites digest cellulose, which taxon is most likely to share this character? → answer = hissing cockroach
What evidence supports your answer?

Figure 4. The tree format cladogram and associated questions (with the answers printed in italics) for 10 insect taxa.

(skill III.C; question A3 in Figure 3). Requesting the smallest clade that includes a certain taxon yields a unique answer. The remaining three questions were phrased as “Do the bracketed taxa labeled ‘A’ constitute a clade?” Students had to circle yes or no and provide written evidence to support that response. If they responded no, they also had to indicate how to make the group into a clade. One bracketed-taxa question required students
to recognize that a set of taxa do comprise a clade (skill III.B; question P3 in Figure 1).

The remaining two questions required students to recognize that a given set of taxa are not a clade because not all the descendants of the MRCA are included. We manipulated the structural relationship of the taxa enclosed by the bracket. In one case, the taxa were at different levels in the cladogram (skill III.A.1; question M3 in Figure 2). In the other case, they were a subset of taxa in a polytomy (skill III.A.2; question I3 in Figure 4). A polytomy is a set of three or more taxa that are not resolved into a structure of nested levels. In the insect cladogram in Figure 4, the beetle, the lacewing, and the clade defined by the character *silk glands* constitute a polytomy.

We predicted based on our earlier research (Novick & Catley, 2007) that students would do worse when the bracketed taxa were part of a polytomy, particularly for the ladder format. Such an interaction between format and not-a-clade reason is supported by perceptual aspects of the tree and ladder cladograms, due to good continuation. We discuss this prediction more fully later.

**Evaluating evolutionary relatedness.** There were two types of evolutionary relatedness questions, each asking which of two taxa is the closest evolutionary relation to a third (reference) taxon and then probing for supporting evidence. As discussed earlier, the questions differed in whether the reference taxon was the most derived of the three taxa (skill IV.A; question M2 in Figure 2) or was at an intermediate level compared with the other two taxa (skill IV.B; question I2 in Figure 4). For all questions, the reference taxon’s physical position in horizontal space was between the two comparison taxa. Thus, we can attribute any differences in students’ responses to differences in structural understanding rather than to superficial differences in adjacency relations among the taxa. In addition, the reference taxon was physically closer (fewer intervening branches and/or levels) to the comparison taxon to which it is less closely related. This is important when the reference taxon is situated at an intermediate level to distinguish whether students understand the underlying structure of cladograms or are responding based on spatial proximity. Diagrammatic literacy requires structural understanding within the domain of application.

**Making character inferences.** There was one inference question (skill V) for each cladogram format, which was followed by the supporting evidence question. The inference questions were worded as follows: “Given that Taxon T has Character C, which taxon is most likely to share this character?” (question I4 in Figure 4). Answering these questions requires making an inference because the named character was not depicted on the cladogram.

**Procedure**

This task was the first part of a multipart booklet that involved data collection for several studies addressing distinct conceptual issues. A
questionnaire at the end of the booklet included the coursework question used to create the biology background groups. The results of the other studies are published elsewhere (Catley & Novick, 2009; Morabito, Catley, & Novick, 2010; Novick & Catley, 2007). Experimental sessions included one or more participants, depending on how many students signed up for a particular time. All participants completed the booklet on their own without using outside resources. They worked at their own pace, taking about 45 to 70 minutes.

Results

Coding Students’ Written Justifications

The codes given to the evidence responses were determined based on what constitutes (components of) a scientifically valid response, expected naïve types of reasoning, and common explanations found in the data. The responses of a randomly selected subset of participants were used to develop the coding scheme and train the coders (the first author and a graduate student). Some codes overlapped across questions; others were specific to a particular question. The coding scheme for each question contained two to six different codes, including OTHER for responses that did not fit any of the named codes. For each question, the codes were ranked in quality from most to least sophisticated, with OTHER at the bottom.

After training, the coders independently coded the remaining 81 students’ written responses. One coder was blind to the biology background of the participants; the other coder knew that stronger background students tended to have higher subject numbers. Each response received a single code. In the few cases where multiple codes applied, the one highest in terms of quality was assigned. Each question that received a content code was an opportunity for the coders to agree or not. Across the 1,051 written responses for the questions and codes reported here, there were 975 agreements, for an agreement rate of 93%. Discrepancies were resolved by discussion. The coding categories for each question are presented with the analyses that rely on that coding.

Identifying Shared Characters Based on Most Recent Common Ancestry (Skill I)

These questions assessed students’ understanding that the character possessed by the MRCA of two or more taxa is the one encountered first in tracing back from each taxon to their branching point. The four questions were alternative ways of probing students’ understanding of this concept. Each question received a score of 0 (incorrect) or 1 (correct; i.e., scientifically accepted). Because we had no basis for predicting differing levels of accuracy as a function of question type, students received a mean score across the four questions for each cladogram format.
These data were analyzed with a 2 (biology background) x 2 (cladogram format) analysis of variance (ANOVA). For all ANOVAs, $\alpha = .05$ was the criterion for statistical significance. We report $\eta^2_p$ as a measure of effect size. Following Cohen's (1988) guidelines for proportion of variance accounted for, .01 is the minimum value taken to indicate a small effect, .09 is the minimum value for a medium-sized effect, and .25 is the minimum value for a large effect.

The ANOVA yielded three significant effects. The main effect of format, $F(1, 105) = 15.17, p < .001, MSE = .03, \eta^2_p = .13$, indicated that students did worse with the ladders than the trees. The main effect of biology background, $F(1, 105) = 23.29, p < .001, MSE = .07, \eta^2_p = .18$, indicated lower accuracy for weaker than stronger background students. The interaction, $F(1, 105) = 8.21, p < .01, MSE = .03, \eta^2_p = .07$, clarified that only weaker background students did worse with the ladder format. The means for stronger background students were near ceiling for these questions that simply required reading information directly off the cladograms: $M = .97$ for trees and $M = .95$ for ladders. Weaker background students had means of .86 and .69, respectively. Although weaker background students did fairly well at extracting appropriate character information from trees, they were much less successful at this seemingly simple task with ladders.

Given the large error rate for weaker background students answering questions about a ladder ($M = 31\%$), we examined the nature of these errors. The four questions fall into two groups, one containing the question that asked for the character possessed by the MRCA of a clade containing only two taxa and the other containing the three questions that asked for the character possessed by the MRCA of two to four more distantly related taxa.

The first type of question was asked about the vertebrate cladogram. Figure 5a reveals that the group in question (crocodiles + birds) joins with a third taxon (lizards) to form a more complex clade with taxa at two levels. Novick and Catley's (2007) cladogram translation results indicated that students often interpret such a nested structure on a ladder (technically, a resolved structure) as if all three taxa are at the same level (technically, a polytomy or unresolved group). If students interpret the clade in question in this manner, they should identify $\text{see UV light}$ as the appropriate supporting character. Indeed, of the 20 weaker background students who got this question wrong, 14 (70\%) gave that particular incorrect answer.

The remaining three questions could be misinterpreted as asking for the character that links taxa in two separate parts of the cladogram. This is especially clear for the fish and spider questions (see Figures 5b and 5c), for which the most common error was to identify the character that seemingly connects those two parts of the cladogram: $\text{jaws}$ for the fish question (22 of 32 errors = 69\%) and $\text{tangle web}$ for the spider question (15 of 26 errors = 58\%). For the dinosaur question (see Figure 5d), the most consistent error was $\text{peg-like teeth}$, the character that is between the $\text{Vulcanodon + Brachiosaurus}$ clade and $\text{Sellosaurus}$ (5 of 20 errors = 25\%).
Identifying Taxa Based on Descent or Not From a Specified Ancestor (Skill II)

These questions probed students’ ability to divide taxa into two groups at points of divergence supported by particular characters. This required varying whether the question was worded positively (skill II.A) or negatively (skill II.B). The questions were designed to test students’ understanding of evolutionary relationships and their ability to identify common ancestry based on shared characteristics.

Figure 5. Relevant subsets of the four ladder format cladograms for the character identification questions (a-d) and the dinosaur inference question (e).

Identifying Taxa Based on Descent or Not From a Specified Ancestor (Skill II)

These questions probed students’ ability to divide taxa into two groups at points of divergence supported by particular characters. This required varying whether the question was worded positively (skill II.A) or negatively (skill II.B).
(skill II.B). Students received separate accuracy scores (0 or 1) for the positively and negatively worded questions for each cladogram format (i.e., for identifying taxa that do vs. do not possess the character in question).

With only one exception, all stronger background students got all four questions correct (i.e., one participant missed one question). Obviously, neither cladogram format nor question wording affected these students’ ability to divide the taxa into two groups based on descent or not from a certain ancestor. We therefore restricted the statistical analysis of the effects of these factors to weaker background students because the lack of variability in the stronger background group means that the homogeneity of variances assumption would be violated if biology background were included as an additional factor in the ANOVA.

The 2 (cladogram format) × 2 (question wording) ANOVA on the data for weaker background students revealed two significant effects. The main effect of format, $F(1, 77) = 20.27, p < .001, MSE = .17, \eta_p^2 = .21$, indicated that these students did worse with the ladder ($M = .65$) than the tree ($M = .86$) format. The main effect of question wording, $F(1, 77) = 11.66, p < .01, MSE = .12, \eta_p^2 = .13$, indicated that they did worse when the question was worded negatively ($M = .69$) rather than positively ($M = .82$). The interaction was not significant, $F(1, 77) = .71, MSE = .11, \eta_p^2 = .01$. Only for the positively worded question for the tree cladogram did weaker background students perform comparably to stronger background students (means of .91 and .97, respectively). For the negatively worded question about the ladder, weaker background students showed particularly poor diagrammatic understanding ($M = .56$).

Understanding Whether and Why a Set of Taxa Constitute a Clade (Skill III)

Overview

The clade questions probed students’ understanding of whether a set of taxa comprise a clade and their ability to justify conclusions about the clade status of a set of taxa given what the cladogram shows concerning the descendants of the ancestral taxon with the relevant character. The data for questions examining recognition of what is not versus what is a clade are reported separately. Due to space considerations, we generally consider students’ explanations for their clade responses in terms of numerical quality scores assigned to the qualitative codes. However, we do discuss one qualitative response for the is-a-clade questions because it bears directly on an important misunderstanding of clades. Two other types of explanations given for both the not-a-clade and is-a-clade questions that suggest a confirmation bias in students’ reasoning are discussed in a companion paper (Catley & Novick, 2012).
Recognizing That a Group of Taxa Do Not Comprise a Clade

To assess students’ ability to recognize that a subset of taxa are not a clade, we computed a mean score across the three parts of the question (accuracy—scored 0 or 1; evidence quality—0, .5, or 1; modify group to create a clade—0 or 1) for each cladogram format and reason why the bracketed taxa are not a clade. The two structural reasons, again, were that the taxa were situated at different levels (skill III.A.1) or they were a subset of taxa comprising a polytomy (skill III.A.2). Students received an evidence score of 1, indicating good understanding of cladograms and the macroevolutionary processes they depict, if they wrote that not all the descendants of the MRCA are included in the bracket: for example, “The common ancestor of the marsupial mole & the badicoot [sic] is where I placed a * [drawn on the diagram at three pair lower teeth] and the clade would be from koala to bandicoot” (stronger background, marsupials). The next best response, indicating some understanding of why the bracketed taxa are not a clade and earning a score of .5, was to write that not all the descendants are included, a particular taxon not in the bracket also is a descendant, or other taxa also have character C (the character possessed by the MRCA): for example, “There is anther [sic] taxa that shares the combining character” (weaker, spiders) and “not all species that have 2 claws are included” (stronger, spiders). All other responses received an evidence quality score of 0.

Based on perceptual aspects of the tree and ladder cladograms at the relevant parts of those diagrams, we predicted students would do worse when the bracketed taxa were part of a polytomy than when they were at different levels, particularly for the ladder format. Frogs and salamanders, the taxa that comprise the bracketed subset of a polytomy on the vertebrate ladder (see Figure 5a), appear to be a coherent unit (thus a clade) because they branch off from the “main line” of the ladder at the same point (Novick & Catley, 2007). The fact that the section of this line that leads from that same point to mammals also is part of the clade is masked by the Gestalt principle of good continuation, which leads viewers to group that segment with the rest of the “main line.” In contrast, it seems clear in the insect tree (see Figure 4), which also contains a polytomy, that not only do beetle and lacewing (the bracketed taxa) extend from the same horizontal line, but so does the branch that leads to the descendants of the MRCA denoted by the character silk glands.

The results of the ANOVA support our predictions (see Table 2). As shown in Figure 6, the format by reason interaction reflects the fact that with the tree format, both groups of students showed comparable understanding that the bracketed taxa do not comprise a clade regardless of the reason; but with the ladder format, students in both groups did much worse when the bracketed taxa were a subset of a polytomy than when they were situated at different levels. The figure also shows the large difference in
understanding between stronger and weaker background students. Note that both groups of students had particular difficulty with the question that required recognizing a subset of a polytomy on a ladder: 56% of weaker and 34% of stronger background students incorrectly said these bracketed taxa (see Figure 5a) comprise a clade.

Recognizing and Identifying Taxa That Do Comprise a Clade

To assess students’ understanding of what is a clade (skills III.B and III.C), we computed a mean score for each cladogram format across three questions concerning recognition (is it a clade—scored 0 or 1; evidence quality—scored 0, .5, or 1) and production (circle the taxa—scored 0 or 1) of a clade. Students who wrote that the bracket included the most recent common ancestor and all its descendants received an evidence quality score of 1. Two types of responses that reflected partial understanding of this concept—focusing either on all the descendants or on the recent common ancestor—earned a score of .5. The descendants partial-credit response involved noting that all descendants are included or that the bracketed taxa are all those with the relevant character (i.e., the one possessed by the MRCA): for example, “every taxa that walks on palm of foot” (weaker, placentals) and “they are all the descendants of a common ancestor” (stronger, fish). The ancestor partial-credit response involved noting that all the bracketed taxa have the same recent common ancestor or all evolved from an ancestor that had character C (that possessed by the MRCA): for

<table>
<thead>
<tr>
<th>Source</th>
<th>$F(1, 105)$</th>
<th>$p$</th>
<th>$\eta_p^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology background (B)</td>
<td>33.06</td>
<td>&lt; .001</td>
<td>.24</td>
</tr>
<tr>
<td>Error (0.17)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format (F)</td>
<td>17.63</td>
<td>&lt; .001</td>
<td>.14</td>
</tr>
<tr>
<td>$F \times B$</td>
<td>2.42</td>
<td>ns</td>
<td>.02</td>
</tr>
<tr>
<td>Error (0.05)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reason (R)</td>
<td>40.33</td>
<td>&lt; .001</td>
<td>.28</td>
</tr>
<tr>
<td>$R \times B$</td>
<td>0.30</td>
<td>ns</td>
<td>.00</td>
</tr>
<tr>
<td>Error (0.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F \times R$</td>
<td>21.41</td>
<td>&lt; .001</td>
<td>.17</td>
</tr>
<tr>
<td>$F \times R \times B$</td>
<td>0.34</td>
<td>ns</td>
<td>.00</td>
</tr>
<tr>
<td>Error (0.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. MSE values are given in parentheses.
example, “Share the most recent common ancestor” (weaker, fish) and “all from ancestor who walks on palm of foot” (stronger, placentals). All other responses earned a quality score of 0.

The ANOVA yielded only a main effect of biology background, $F(1, 105) = 38.14, p < .001, MSE = .07, \eta_p^2 = .27$, with higher scores for stronger than weaker background students ($M = .74$ vs. $M = .49$, respectively). The main effect of format, $F(1, 105) = .94, MSE = .04, \eta_p^2 = .01$, and the format by background interaction, $F(1, 105) = .00, \eta_p^2 = .00$, were not significant.

For students who incorrectly answered that bracketed taxa comprising a clade are not a clade, we briefly discuss one evidence response that provides a key insight into the nature of students’ difficulty understanding cladogram structure. Some students wrote that the bracketed group includes more than just two most-closely-related taxa. These responses indicated that there is a more recent common ancestor, or the clade includes more than just the descendants of the MRCA, or multiple clades are included. Consider two such responses from weaker background students for the

![Graph showing mean score for questions assessing skill at recognizing that bracketed taxa do not comprise a clade as a function of biology background, tree versus ladder cladogram format, and reason (different levels vs. subset of a polytomy) why the bracketed taxa do not comprise a clade.](image)

**Figure 6.** Mean score for the questions assessing skill at recognizing that bracketed taxa do not comprise a clade as a function of biology background, tree versus ladder cladogram format, and reason (different levels vs. subset of a polytomy) why the bracketed taxa do not comprise a clade.

*Note.* Error bars represent the standard error for each format and question type.
cladogram in Figure 1: (a) “The rock hyrax does not have subdermal fat stores as the manatees, and elephant do. They don’t share the most recent common ancestor.” and (b) “A also includes rock hyrax, which does not come from the most recent ancestor.” Collapsed across cladograms and biology background, 36% of those who incorrectly said the bracketed group is not a clade gave such an explanation. In 12 of these 16 cases, students said to remove the least related taxon (e.g., rock hyrax in Figure 1) to make the group a clade. Because clades are nested, however, no taxon has to be removed. These students seemed to think a clade may contain only two taxa that are most derived, so the least derived of the three taxa must be removed to create a clade. That is, they appear not to understand that cladograms depict multiple, nested clades.

**Determining Relative Evolutionary Relatedness (Skill IV)**

To evaluate students’ understanding of evolutionary relatedness, we computed a mean across accuracy (scored 0 or 1) and evidence quality (scored 0, .5, or 1) for each reference taxon location (skills IV.A and IV.B) and cladogram format. Students received an evidence quality score of 1 if their explanation appealed to the MRCA of the reference taxon and the taxon selected as its closer relative: for example, “Butterflies have a more recent common ancestor with the walking stick” (stronger, insects) and “koala’s [sic] and BT possums share a closer common ancestor then [sic] bandicoots and BT possums” (stronger, marsupials). The next best response (score of .5) was to state that the reference taxon and the selected taxon share an ancestor with the relevant character (i.e., the one possessed by the MRCA). For example, one student correctly chose butterfly for the insect question because “both have an ancestor with wings folded over the body” (stronger). Another student correctly selected lungfish for the fish question because the two taxa are “both descendants of ‘cycloid scales’” (weaker). Although these students mentioned common ancestry and the relevant shared character, they failed to indicate that this is the appropriate character because it was possessed by the MRCA of the taxa in question.

All other explanations received a quality score of 0. These explanations included responses like those described by Meir et al. (2007) in which students focused on the number of steps or levels separating the taxa or on the horizontal distance between the taxa across the top of the cladogram. Such responses reflect a reliance on spatial proximity for evaluating evolutionary relatedness. Example steps and physical distance responses supporting incorrect answers are, respectively: (a) “Less character steps between the two, more in common” (weaker, insects) and (b) “pirate spider comes right after brown recluse” (weaker, spiders).

Because Meir et al. (2007) previously reported on students’ explanations for their evolutionary relatedness responses, albeit only for the ladder
format, we focus on the quantitative results for our study. We predicted
greater diagrammatic understanding when the reference taxon was most
highly derived than when it was at an intermediate level. The ANOVA results
are given in Table 3. As shown in Figure 7, students did much better when
the evolutionary relationships were depicted on a tree than a ladder ($M = .43$
vs. $M = .27$, respectively) and when the reference taxon was most highly
derived than when it was at an intermediate level ($M = .47$ vs. $M = .23$,
respectively). In addition, stronger background students did better than
weaker background students ($M = .47$ vs. $M = .30$, respectively). The
extremely low scores for the intermediate-position question for the ladder
($M = .18$) replicates the findings of Meir et al. (2007) and Sandvik (2008).
The comparisons between this condition and the other three conditions
are new to this study, as is the assessment of understanding based on
both accuracy and evidence quality.

Using Tree Thinking to Support Inferences (Skill V)

*Quantitative Analysis*

Although two taxa can share a character due to convergent evolution
from separate ancestors, it is more parsimonious (absent contradictory evi-
dence) to infer that common ancestry is the cause. To assess students’ ability
to use cladograms to make and support character inferences, we computed
a mean across inference accuracy (scored 0 or 1) and evidence quality
(scored 0, .5, or 1) for each question. The evidence codes are discussed in

---

*Table 3*

Results of the ANOVA on Performance on the Questions Assessing Skill at
Evaluating Evolutionary Relatedness

<table>
<thead>
<tr>
<th>Source</th>
<th>$F(1, 105)$</th>
<th>$p$</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology background (B)</td>
<td>17.97</td>
<td>&lt; .001</td>
<td>.15</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>(0.17)</td>
<td></td>
</tr>
<tr>
<td>Within subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format (F)</td>
<td>53.29</td>
<td>&lt; .001</td>
<td>.34</td>
</tr>
<tr>
<td>F × B</td>
<td>0.09</td>
<td>ns</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>(0.04)</td>
<td></td>
</tr>
<tr>
<td>Location (L)</td>
<td>58.56</td>
<td>&lt; .001</td>
<td>.36</td>
</tr>
<tr>
<td>L × B</td>
<td>0.05</td>
<td>ns</td>
<td>.00</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>(0.09)</td>
<td></td>
</tr>
<tr>
<td>F × L</td>
<td>2.50</td>
<td>ns</td>
<td>.02</td>
</tr>
<tr>
<td>F × L × B</td>
<td>0.81</td>
<td>ns</td>
<td>.01</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>(0.05)</td>
<td></td>
</tr>
</tbody>
</table>

*Note. MSE values are given in parentheses.*
the following subsections. The ANOVA yielded two significant effects: Students did much better with the tree (insect) question (M = .52) than the ladder (dinosaur) question (M = .25), F(1, 105) = 39.58, p < .001, MSE = .08, η² = .27; and stronger background students (M = .56) did better than weaker background students (M = .32), F(1, 105) = 12.34, p < .01, MSE = .20, η² = .11. The interaction was not significant, F(1, 105) = .04, MSE = .08, η² = .00.

Use of Cladogram-Based Evidence

The best evidence (score of 1; MRCA) was to appeal to the MRCA of the relevant taxa. The next best response (score of .5; REL) was to note that termites and roaches or hawks and T. rex are most closely related or closest evolutionarily. For both questions, students provided such justifications only to support correct inferences. For example, one student inferred that T. rex was most likely to be warm-blooded like birds because “the hawk
and T-rex share a more recent common ancestor that the other taxa do not share” (MRCA; stronger). Another student inferred that hissing cockroaches are most likely to digest cellulose like termites because “it is the closest relative of the harvester termite, linked by asymmetrical male genitalia” (REL; stronger). Considering MRCA and REL responses together, stronger background students did fairly well for the insect question, with 72% providing an appropriately justified correct inference. But for the dinosaur question, this number decreased to 45%. For weaker background students, these numbers are only 37% and 15%, respectively.

The top zero-credit code was shared character (SC), which was given when students appealed to a character shared by the given and selected taxa. This was the modal evidence cited by stronger background students to support incorrect inferences for the dinosaur question: 69% of stronger but only 19% of weaker background students. The most common shared character cited was hollow bones (13 of 20 students who earned this code). For example, one student (stronger) answered hawk because “it descended from an ancestor w/ hollow bones, a trait of birds.” Another student (weaker) answered Ischisaurus because “Ischisaurus has hollow bones, as do birds, humans, and other warm-blooded mammals.” The principle of good continuation is likely to lead students to interpret Ischisaurus, Velociraptor, T. rex, and hawk as a valid group (clade) supported by the hollow bones character (see Figure 5e). Note that both of these students also brought information from memory to bear on answering this question.

**The Influence of Prior Knowledge**

The final two content-specific codes provide evidence that students’ incorrect inferences were based on prior knowledge rather than on the information given in the cladograms. The category (CATEG) code was used when students chose the specific example depicted in the cladogram of the more inclusive taxon named in the question because it is in the same category (harvester termite because it is a termite or hawk because it is a bird). The morphological similarity (MORPH) code was used when students wrote that their selected taxon has a certain character, which is not shared with the relevant taxon from the question. The most interesting code was MORPH, which accounted for roughly half of both stronger (60%) and weaker (50%) background students’ evidence for their incorrect inferences for the insect question. The modal incorrect answer was to state that walking sticks are most likely to also digest cellulose because they have sucking mouthparts (some students picked butterflies because they have coiled mouthparts). Although most students simply named the relevant mouthparts as their entire answer, a few provided more insight into their reasoning. For example, referring to the walking stick, a weaker background student wrote that “its sucking mouth makes it seem...
the most capable of digesting food.” A stronger background student wrote that “walking sticks are the only taxon w/ ability to suck w/ sucking mouthparts. Sucking mouthparts would be helpful in sucking up cellulose.” Students who received the MORPH code seem to have reasoned that because the question concerns the digestive system, the most appropriate inference was to a taxon for which information relevant to digestion was provided. The mouth is part of the digestive system, so the presence of a character pertaining to mouthparts was judged an appropriate basis for inference.

We suspect MORPH justifications were less common for the dinosaur question (4% for both biology groups) because none of the characters on the cladogram seem relevant to warm-bloodedness, the character in question. In fact, some students said exactly this. Seven students (6 weaker, 1 stronger background) refused to make an inference to any of the taxa on the cladogram. Five of these students gave supporting explanations such as “none of these characteristics have to do with blood—all bones” (weaker) and “this evolutionary tree has no information on ‘blood’ characteristics” (stronger); the remaining two students left the evidence question blank as well.

The modal evidence cited by weaker background students to support an incorrect inference for the dinosaur question was CATEG: 40% said hawks are most likely to be warm-blooded (given that birds are) because they are birds (23 of the 78 students in this group overall). In contrast, 23% of stronger background students who made an incorrect inference responded this way. For the insect question, 26% of justifications for an incorrect inference, collapsed across the two groups, received this code. We suspect the CATEG code was especially prevalent for weaker background students for the dinosaur question because of their unwillingness to accept that birds are dinosaurs. Thus, an inference from birds to canonical dinosaurs, especially T. rex, which does not look like a bird, likely seemed unwarranted. Students’ understanding of the relation between modern birds and extinct non-avian dinosaurs would be an interesting topic for future research.

We believe the differences in the explanations provided for incorrect inferences for the insect and dinosaur questions are primarily due to students’ prior knowledge about the taxa and characters depicted rather than to the cladogram format. In research with high school students, the same types of incorrect explanations were found for the dinosaur cladogram drawn in the tree format as are reported here for the ladder format (Catley, Phillips, & Novick, 2012).

An Overall Assessment of Tree-Thinking Competence

Finally, we provide a more global view of students’ tree-thinking competence. Because students may be able to provide correct answers before
they understand the macroevolutionary concepts that support those answers, Table 4 shows the mean accuracy score for the non-evidence questions for each of the five tree-thinking skills (averaged across question types) as a function of biology background and cladogram format.

To get an overall idea of the depth of students’ understanding of the evidential basis for cladograms, we computed for each format the mean evidence quality score across six questions: bracketed-taxa clade \( (n = 3) \), evolutionary relatedness \( (n = 2) \), and inference \( (n = 1) \). The best explanation for each question was to cite evidence concerning most recent common ancestry. An ANOVA\(^6\) revealed significant main effects of cladogram format, \( F(1, 104) = 35.09, p < .001, \) \( \text{MSE} = .01, \) \( \eta_p^2 = .25 \), and biology background, \( F(1, 104) = 107.99, p < .001, \) \( \text{MSE} = .03, \) \( \eta_p^2 = .51 \). Students showed greater competence when reasoning about information depicted in a tree than a ladder, with means of .24 and .17, respectively. In addition, stronger background students \( (M = .39) \) had higher scores than weaker background students \( (M = .13) \). A significant interaction, \( F(1, 104) = 5.24, p < .03, \) \( \text{MSE} = .01, \) \( \eta_p^2 = .05 \), indicated that the increase in reasoning ability for trees compared with ladders was larger for stronger (means of .45 and .34, respectively) than weaker (means of .16 and .11, respectively) background students.

Note that even the stronger background students reasoning about information depicted in a tree did quite poorly in an absolute sense, with a mean evidence quality score of only .45 \( (SD = .15) \). That is, roughly speaking, even the 29 most biologically educated students in our sample—those with three to seven semesters \( (M = 4.5) \) of coursework in (primarily) organismal biology—cited evidence concerning most recent common ancestry when reasoning about cladograms in the tree format slightly less than half the time such justifications were appropriate.
Discussion

On Students’ Competence at Tree Thinking

Cladograms, the tool of choice for studying macroevolution, organize information about taxa according to recency of common ancestry rather than overall similarity. The strong basis for inference afforded by this method of systematizing the 3.5 billion year history of life on Earth has had, and will continue to have, important benefits for humanity (e.g., AMNH, 2002; Futuyma, 2004; Yates et al., 2004). Addressing recent calls to include cladogram-based tree thinking in biology curricula from middle school through college (e.g., Baum et al., 2005; Catley, 2006; Catley et al., 2005; Goldsmith, 2003; Gregory, 2008; O’Hara, 1988), the present study provides information on college students’ competence at five core tree-thinking skills, an aspect of diagrammatic literacy, given existing instruction. In addition, the results support our predictions that students’ success at tree thinking is influenced by a variety of perceptual and conceptual factors. We investigated two Gestalt perceptual principles—good continuation and spatial proximity. The conceptual factors investigated were positive versus negative question wording, biology background, and prior knowledge of the taxa depicted in the cladograms. We consider each factor in turn.

Good Continuation

Using diagram segmentation and translation tasks, Novick and Catley (2007) found that the Gestalt perceptual principle of good continuation hampers students’ ability to correctly extract the nested relationships depicted in the ladder format (e.g., see Figure 5): Because this principle leads people to interpret a continuous line as a single entity, many taxa in ladder format cladograms that are situated at different hierarchical levels are instead interpreted as being at the same level. Thus, our manipulation of cladogram format (tree vs. ladder) constituted a manipulation of the perceptual principle of good continuation. We found that students were more accurate at answering tree-thinking questions and used more sophisticated reasoning when relationships were depicted in a tree rather than a ladder, thereby providing the first evidence that this perceptual principle adversely affects students’ reasoning about taxa, characters, and relationships. For diagrammatic literacy skills involving simple identification of characters and taxa depicted in cladograms (skills I and II in Table 1), a decline in accuracy for the ladder format was only observed for students with weaker biology backgrounds. For skills requiring a good understanding of how evolutionary relationship is mapped onto cladogram structure (skills III and IV), however, both weaker and stronger background students had greater difficulty with the ladders. These effects were generally medium to large in size. Unfortunately,
high school and college biology textbooks preferentially use the ladder format (Catley & Novick, 2008).

Spatial Proximity

Cladograms depict nested levels of relationship indicating that two taxa are more closely related to each other than either is to a third taxon. Any cladogram, no matter how complex, can be decomposed into sets of such three-taxon statements. Because evolutionary relatedness questions (skill IV) directly probe students’ understanding of this core structure, they are arguably the quintessential tree-thinking question. We manipulated whether the reference taxon was the most highly derived of the three taxa or was at an intermediate level between the other two taxa. In the latter case, the reference taxon was spatially closer (horizontally and vertically) to the taxon to which it is less closely related. We found a significant effect of reference taxon location that was consistent across biology background groups and cladogram formats: Students were much less accurate when the reference taxon occupied a hierarchically intermediate level and spatial proximity provided a conflicting cue to evolutionary relationship than when that taxon was most derived and both nesting and proximity provided converging cues to evolutionary relationship. This was a large effect.

Positive Versus Negative Question Wording

Assessing students’ ability to distinguish taxa that have a certain character due to descent from a MRCA from those that do not (skill II) requires asking positively and negatively worded questions. We asked the negative question as simply as possible: for example, “List all taxa that did not [emphasis in the original] evolve from an ancestor that had 1 pair of antennae.” A general aspect of human cognition (e.g., Matlin, 2009), however, is that our cognitive processes cope better (more easily) with positive than negative information. We found that question wording interacted with biology background such that the negatively worded question was more difficult than the positively worded question for students with weaker, but not stronger, backgrounds in biology (for both cladogram formats).

We hypothesize that this interaction may be explained by differential working memory loads across conditions. The extra processing needed to understand negatively worded sentences (Kaup et al., 2006, 2007) presumably increases the load on working memory. Because students find it more difficult to interpret the nested structure of the ladder than the tree format, the ladder format also likely places a greater burden on working memory. Finally, the tree-thinking task itself (identifying an appropriate subset of taxa) may place a greater load on working memory for those who are less familiar with cladograms. Considered together, these hypotheses could yield the observed pattern in which question wording and cladogram format
affected weaker background students’ performance but not that of stronger background students. This explanation of our results could be tested using a dual-task methodology, which has been done previously with diagrammatic materials (e.g., Gyselinck, Cornoldi, Dubois, De Beni, & Ehrlich, 2002). More generally, we would predict a similar interaction between positive versus negative question wording and amount of prior knowledge for other kinds of reasoning tasks that call upon this knowledge and vary in their working memory requirements.

**Biology Background**

As predicted, students with stronger backgrounds in biology (i.e., who had taken more, and more advanced, biology courses) did better than those with weaker backgrounds. Statistically, the biology background effect was generally medium to large. This is encouraging because it suggests biology students receive some exposure to macroevolutionary concepts and cladograms in their classes. On the other hand, these students’ absolute level of performance on many questions indicates that current instruction is far from yielding students who are highly competent in this critical area. For example, these students’ average tree-thinking evidence quality score, which reflects their understanding of most recent common ancestry as the basis for interpreting and making inferences from cladograms, was only .39 on a 0 to 1 scale.

**Prior Knowledge**

When prior knowledge of the taxa under consideration and cladistic depictions conflict, students should accept the evidence-based cladogram to override faulty or incomplete knowledge acquired from other sources. Our results for the dinosaur inference question, however, suggest that prior knowledge of the taxa can interfere with successful tree thinking. Research on other topics in science education has also found interfering effects of prior knowledge (e.g., Chinn & Brewer, 1993; Kendeou & van den Broek, 2005). Consistent with accepted scientific knowledge, our cladogram shows that birds are dinosaurs. Given the relationships depicted, the most likely taxon to share a character (warm-bloodedness in our question) possessed by birds (represented by hawks in the cladogram) is *T. rex*. However, only about half (55%) of stronger background students were willing to use the evidence shown to make this inference, with weaker background students doing even worse (27% correct).

Students’ justifications underscore the conflict they had reasoning with the information that birds are dinosaurs. Of the 70 students who got the inference question wrong, 41 (59%) wrote that hawks are most likely to be warm-blooded. Because the question states that birds are warm-blooded, this response is an example of what one has been told rather than an
inference. In fact, this is how students tended to justify these responses (33 of the 41): 24 responded that hawks are birds, another 5 included this category response as part of their answer (e.g., “it is a bird and has hollow bones”), and 4 pointed in some other way to the relationship between hawks and birds in general (e.g., saying that a hawk “is most similar to the bird”).

Implications for Teaching Tree Thinking

Cladogram Format to Use for Initial Instruction

Professional biologists strongly prefer to depict evolutionary histories in the tree format (Novick & Catley, 2007; K. Quillin, personal communication, May 27, 2008). In contrast, ladders are more common in contemporary high school and college textbooks (Catley & Novick, 2008). Although the two formats are isomorphic representations, providing access to the same information about evolutionary relationships, they rely on different conventions for their construction and interpretation. We found for a wide variety of tree-thinking skills that students have greater difficulty answering questions about information depicted in ladders than in trees. We urge instructors, therefore, to use the tree format in initial instruction at both the high school and college levels. Teachers, of course, will need to be able to interpret the ladder format as long as it continues to be found in introductory textbooks. The ladder format does need to be introduced in upper-level classes for college biology majors, however, because that format is seen in the primary literature in evolutionary biology. When this format is introduced, we suspect students will need explicit instruction concerning the structural isomorphism of the two formats along with instruction in the conventions for mapping evolutionary meaning onto the ladder format.

Understanding Most Recent Common Ancestry

Fundamentally, the diagrammatic literacy skill of tree thinking is the ability to use the concept of most recent common ancestry to comprehend and reason from the evidence (characters) and relationships (topology) depicted in cladograms. Two aspects of our results indicate that college students, even those who have taken at least three to seven ($M = 4.5$) semesters of coursework for biology majors, do not have a good understanding of this concept that is the foundation for mapping evolutionary meaning onto and extracting such meaning from an abstract hierarchical diagram.

First, consider students’ justifications for their answers to the bracketed-taxa clade, evolutionary relatedness, and inference questions, for which the best response is to cite evidence of most recent common ancestry. Our composite tree-thinking measure, computed as the average of the evidence quality scores for these questions, suggests poor student understanding (also see
Evans et al., 2010). Even the best quality evidence, a mean of .45 on a 0 to 1 scale for our stronger biology background group when reasoning about cladograms in the tree format, falls far short of what would be expected of students with a good understanding of most recent common ancestry.

A second window into students’ understanding of most recent common ancestry comes from the evolutionary relatedness questions. Focusing on the questions with the reference taxon at the intermediate level, for which a correct response requires an understanding of most recent common ancestry rather than just a more general understanding of hierarchical structure, we found an accuracy rate of only 46% for the tree and 24% for the ladder (chance is 50%). The percentage of students who both got the question correct and provided MRCA evidence to support that answer was 29% and 3%, respectively, for stronger and weaker background students (with no difference across cladogram formats).

Understanding the Nesting of Clades

Another source of difficulty for students discovered by our study is understanding that cladograms are composed of multiple clades involving nested sets of taxa. When students had to add taxa to a bracketed group to form a clade, they provided some evidence for understanding that clades may include taxa at multiple levels. But when shown a clade composed of taxa at multiple levels and asked whether the taxa comprise a clade, they often said no. These contradictory responses were even observed within the same students.

Combining the information in this section and the previous one, our results suggest that initial instruction should focus on two core concepts that represent key conventions for constructing and interpreting cladograms. The first critical concept is that it is most recent common ancestry (as opposed to common ancestry)—namely, the relative level of nesting—that defines evolutionary relationships. Second, the three-taxon statement, as the basic building block of all cladograms, yields the nesting of clades that defines the structure (topology) of a given cladogram.

Use of Evidence

A critical aspect of cladograms is that they provide scientific evidence, in the form of synapomorphies (novel shared characters) supporting clades and patterns of nesting, to support reasoning about taxa and their evolutionary history. Our results indicate, however, that our participants did not always view this evidence as more reliable or more valid than their own naive reasoning strategies or prior knowledge. This was particularly evident for the insect inference question, for which the modal incorrect response for both biology background groups was to write that walking sticks are most likely to digest cellulose like termites because they have a certain kind of
mouthparts (noted on the cladogram), even though that character is not shared with termites. Another important aspect of instruction, therefore, concerns clearly communicating the evidential basis of cladograms and why these diagrams provide a stronger basis for inference than other types of knowledge. Understanding and interpreting scientific evidence is a critical aspect of science literacy.

Directions for Future Research

The present study, although large and multifaceted, is just the tip of the iceberg for understanding students’ competence at tree thinking and drawing implications for teaching this important component of diagrammatic literacy. Our recommendations for instruction are based on our results for only five skills. We identified two additional tree-thinking skills but were unsuccessful in writing questions to assess them: recognizing and understanding the evidential basis for convergent evolution and identifying the evolutionary sequence of characters supporting the historical relationship for a set of taxa. Evaluating students’ understanding of these (and other) concepts, and determining associated instructional implications, awaits future research. Another important direction for future research is to extend these inquiries to younger students—certainly high school and ideally K-8 as well. We suspect that our college student participants would not have had such a hard time with our dinosaur inference question if they had been presented with evidence early on that not all dinosaurs are extinct. Imagine if preschool children learned that the national symbol of the United States is a dinosaur. The benefit of research on and associated instruction in tree thinking will be increased science literacy in a domain that is revolutionizing applied research in fields as disparate as ecology, conservation, epidemiology, and pharmacology.

Promoting Diagrammatic Literacy

Diagrams are critically important in most, if not all, fields of science as tools for learning, reasoning about, and communicating structures, processes, and relationships (e.g., Hegarty et al., 1991; Hegarty & Stull, in press; Lynch, 1990; Novick, 2006a). Diagrammatic literacy, therefore, must be a primary goal of K-16 education. The cladograms studied here are a type of schematic diagram, in which spatial relations between elements metaphorically convey meaning (Hegarty et al., 1991). Other types of schematic diagrams that are important in STEM (science, technology, engineering, and mathematics) fields include Euler circles, line and bar graphs, circuit diagrams, matrices (e.g., the periodic table of the elements), and networks (e.g., food webs, the Krebs cycle). Schematic diagrams facilitate learning because they simplify complex situations, make abstract concepts more concrete, and allow learners to substitute easier perceptual inferences for more
difficult search and verbal deduction (e.g., Larkin & Simon, 1987; Winn, 1989).

Because of the metaphoricity of schematic diagrams, however, their meanings are not transparent. Students must be taught both the conventions underlying the construction of each type of diagram (e.g., Cromley et al., 2010; Hegarty & Stull, in press; Yeh & McTigue, 2009) and how the diagram should be scanned for appropriate interpretation of its structure (e.g., Corter, Nickerson, Tversky, Zahner, & Rho, 2008). Without sufficient instruction, students’ understanding of and learning from schematic diagrams is impaired, as we found in our study.

Moreover, although perceptual inferences are easier than verbal deductive inferences (Larkin & Simon, 1987), the perceptual inferences afforded by a particular diagram are not necessarily the most appropriate ones in the given situation. We saw evidence for this in our data with respect to both incorrect interpretations of hierarchical structure due to the Gestalt principle of good continuation and misinterpretations of evolutionary relatedness due to the Gestalt principle of spatial proximity. These are general perceptual processing principles that can be predicted to affect students’ interpretations of diagrams in domains other than the one we investigated, including those outside the sciences. For example, Landy and Goldstone (2007) argued that mathematical equations are like diagrams in some ways, and, consistent with this theorizing, they found evidence for effects of spatial proximity on students’ interpretations of equations.

The present results support an emerging pattern of findings indicating the need for a two-pronged approach to promoting diagrammatic literacy in students. One prong is that people who design diagrams need to consider the role of perceptual and conceptual processing factors in students’ interpretations of diagrams. The second prong is that people who develop curricula need to include instruction in how to interpret the types of diagrams students will encounter in that field (e.g., Cromley et al., 2010).

Acknowledgments

We thank Courtney Shade for her help collecting and scoring the data and Nancy Morabito for her help coding students’ written explanations.

Notes

1 At the time we developed our questions, there were no published studies in this area, although short reports by Meir, Perry, Herron, and Kingsolver (2007) and Sandvik (2008) now exist.

2 Of the 498 cladograms found in 27 high school and college textbooks surveyed by Catley and Novick (2008), 94% were depicted in the tree or ladder format.

3 There was no difference in overall performance between the captive group of zoology students and the volunteers from other biology classes who had taken comparable numbers of classes.
It was not feasible to try to determine the actual coverage of macroevolutionary concepts in each of the 10 classes across the several years in which our participants might have taken them. However, these are classes for which it is reasonable to expect that evolution concepts might have been considered and cladograms might occasionally have been presented. Indeed, the fact that stronger background students generally did better than weaker background students on our task supports our supposition that the former students received some exposure to macroevolutionary concepts and cladograms in (some of) these classes. Regardless, it is reasonable to think that students who have taken these classes know more about biology than those who have not, which is the only claim we make when stating that these students have a stronger background in biology.

The simplest way to modify each group of taxa to make it a clade is to add particular taxa to the group. When the bracketed taxa are at different levels, one could also add the relevant taxa and delete the bracketed taxon that is located at a lower level of nesting. Participants were clearly told in the question (see Figures 1, 2, and 4) to indicate whether the taxa they listed were to be added or removed. In a few cases, participants listed the correct taxa to be added but did not indicate what was to be done with those taxa. Those few responses received half credit (.5) for accuracy.

Examination of the individual participant data revealed that one weaker background student was an outlier. She received a mean score for all 12 questions of .96, whereas the next largest score by a weaker background student was only .42 (i.e., 77 students in this group had scores from 0 to .42 and 1 student had a score of .96). For statistical reasons, therefore, this student was excluded from the analyses.

References


Novick, Catley


Manuscript received October 23, 2010
Final revision received January 25, 2012
Accepted April 3, 2012