Modeling and Measuring Tree-Reading Skills in Undergraduate and Graduate Students

Thilo Schramm,* Anika Jose, and Philipp Schmiemann†
Department of Biology Education, University of Duisburg-Essen, 45141 Essen, Germany

ABSTRACT

Evolutionary trees are central to learning about evolutionary processes, yet students at all educational levels struggle to read and interpret them. The synthetic tree-reading model (STREAM), based on published and not yet empirically tested models, was tested to determine whether the assumed hierarchy of the model could be substantiated and how far students' skills could be distinguished empirically. We developed a tree-reading test instrument based on STREAM and assessed it with 592 undergraduate and graduate biology students. Following item response theory, we conducted a dimensional analysis and evaluated item difficulty. Investigating item difficulty and the resulting Wright map showed that skill levels displayed a broad scatter of overlapping item difficulty. Furthermore, the skill level assumed easiest was actually the third most difficult. No conclusive evidence of the hierarchical nature of the model was obtained. Dimensional analysis showed that a five-dimensional model outperformed all other reasonable models, corroborating that the skills could be arranged in empirically differentiable groups. Consequently, we revised the STREAM by discarding the hierarchical organization, using a five-dimensional organization instead. Comparison of the revised STREAM with another recently published approach showed that, although these two instruments have a different focus, they are supplemental approaches that show comparable results.

INTRODUCTION

The central claims of Charles Darwin's revolutionary work (1859)—the relatedness of living species and descent with modification from common ancestors—are directly represented in evolutionary tree diagrams. Presently, evolutionary trees are frequently used by biologists to examine patterns of evolutionary relatedness and to test evolutionary hypotheses (Baum *et al.*, 2005). Therefore, they are integral elements of modern evolutionary biology (Meisel, 2010; Catley *et al.*, 2012).

Evolutionary trees are diagrams consisting of lines and nodes based on the mathematical field of graph theory (Wiley and Lieberman, 2011). Typically, an evolutionary tree starts branching from a root, representing the earliest ancestor of all the presented species. From this point, the tree usually spreads in a dichotomous manner. Each node represents the most recent common ancestor (MRCA) of all groups branching from it. The inner nodes of an evolutionary tree are critically important for interpreting relative evolutionary relatedness (Baum and Smith, 2013; Blacquiere and Hoese, 2016).

Diagrams of evolutionary relatedness can also show information such as apomorphies, which are newly developed evolutionary traits (Baum and Smith, 2013). Specifying which traits developed in a group can explain and emphasize bifurcation events. Evolutionary trees that include apomorphies might be easier to read than those without (Catley *et al.*, 2010; Novick *et al.*, 2010).

Extensive research has been conducted on typical learners' conceptions of evolutionary trees (Gregory, 2008); teleological interpretations are especially common (Schramm and Schmiemann, 2019). A wide array of students' conceptions have been

Jennifer Knight, *Monitoring Editor*
Submitted Jun 30, 2020; Revised Apr 8, 2021; Accepted Apr 15, 2021
CBE Life Sci Educ September 1, 2021 20:ar32
DOI:10.1187/cbe.20-06-0131
ORCID-ID: 0000-0002-6290-246X.
†ORCID-ID: 0000-0001-7827-7008.
*Address correspondence to: T. Schramm (Thilo.Schramm@uni-due.de).

reported (Baum *et al.*, 2005; Meir *et al.*, 2007; Gregory, 2008; Omland *et al.*, 2008; Thanukos, 2009; Catley *et al.*, 2010; Halverson *et al.*, 2011; Kummer *et al.*, 2016), and high school students (Catley *et al.*, 2013; Bokor *et al.*, 2014), as well as graduate and undergraduate college students (Meir *et al.*, 2007; Omland *et al.*, 2008; Omland, 2014; Halverson, 2011; Catley *et al.*, 2012; Dees *et al.*, 2014; Blacquiere and Hoese, 2016; Leone, 2017), show major difficulties in understanding and working with evolutionary trees.

Tree-Thinking

The set of abilities needed to read and interpret evolutionary trees, but also to construct a tree hypothesis from the given data, is called tree-thinking (O'Hara, 1988). Tree-thinking can be further subdivided into tree-reading and tree-building (Halverson, 2011). Tree-reading describes all tasks and abilities linked to the extraction and interpretation of information given in an evolutionary tree. Tree-building involves phylogenetic inference, constructing a diagrammatic estimate, or creating hypotheses about evolutionary relatedness based on the given data (Halverson and Friedrichsen, 2013). Tree-thinking deals with the most direct form of presenting macroevolutionary patterns and is seen as an integral part of modern evolutionary biology (Meisel, 2010; Catley *et al.*, 2012).

A number of instruments have been published that investigate different aspects of tree-thinking (Baum *et al.*, 2005; Naegle, 2009; Halverson, 2011; Halverson *et al.*, 2011; Catley *et al.*, 2012, 2013; Gibson and Hoefnagels, 2015; Blacquiere and Hoese, 2016; Leone, 2017; Kummer *et al.*, 2019).

Tree-Thinking Skills

To effectively teach tree-reading at universities and high schools, educators need to be aware of the factors that make trees difficult to read and the subskills needed for tree-reading. Different authors have compiled and presented simple or elaborate tree-reading skill systems (Meir *et al.*, 2007; Halverson, 2011; Halverson and Friedrichsen, 2013; Novick and Catley, 2013, 2016; Blacquiere and Hoese, 2016). These works mostly do not reference each other and are mainly based on research findings or teaching experience. In addition, they typically lack empirical testing. The most extensive systems have been published by Halverson and Friedrichsen (2013; a seven-level model of representational competence of tree-reading and tree-building) and Novick and Catley (2013, 2016; a system of 11 different tree-reading skills). Halverson and Friedrichsen's model is described as a system of seven levels of competence, and students are described as advancing from one level to another. Hence, we see this approach as indication for a kind of hierarchical structure of tree-reading skills worthy of investigation.

It is reasonable to assume that students learn tree-thinking skills one by one, starting with knowledge about the general meaning of the elements and structure of the diagram. Consequently, students need to grasp lower-level competencies before advancing to the higher-level ones. However, a hierarchy within a skill system could also simply reflect the different challenges of somewhat independent skills. This could be demonstrated by investigating the difficulty of items linked to their respective skills.

As there is not much insight into the organization of skill hierarchy in the context of tree-reading, we investigated this topic by considering both aspects of hierarchical systems: learning progressions and different difficulties of differentiable skills.

We compiled all these published works on tree-reading skills and developed a synthetic tree-reading model that brings together all the published tree-reading systems, which we named the synthetic tree-reading model (STREAM; Schramm *et al.*, 2019). This six-level hierarchical model is intended to enable educators to structure their learning environments and to provide researchers with a theoretical basis for planning and using diagnostic instruments. Currently, we view this model as a skill system, not as a learning progression, as learning progressions include very strong claims about the structure of a set of skills or competencies (Gotwals and Alonzo, 2012), and these are not yet available. The six levels of STREAM are presented in Table 1 and outlined here.

The base level (0, *naïve handling*) describes the uninformed use of evolutionary trees. Learners at this level do not have a deep understanding of the symbolic meaning of diagrammatic elements and lack the ability to obtain meaningful information from a given tree. Students' understanding at this level is based on a number of misconceptions and overinterpretation of the uninformative elements of a tree diagram.

Skill level 1 (*identifying structures*) describes required knowledge about diagrammatic structures and their relevance, including the biological meaning of internal and terminal nodes, the direction of the flow of time, and more complex ideas such as

TABLE 1. The STREAM, a six-level hierarchical system of skills comprising tree-reading, based on previous findings in the field

Skill level	Skill description
0. Naïve handling	Students do not interpret the tree correctly. Uninformative features are overinterpreted, and critical misconceptions are applied.
1. Identifying structures	Students are able to identify and interpret the elements of the diagram (nodes, branches, labels, direction of time, etc.) and can answer questions about the structure of the tree.
2. Handling apomorphies	Students are able to answer questions about the meaning and implications of apomorphies. Taxa can be grouped based on apomorphies presented in the tree.
3. Identifying relationships	Students are able to state whether groups form clades and can evaluate the relative relatedness of a set of taxa. This includes simple and complex statements about the relationship of three taxa and about taxa and their MRCA(s).
4. Comparing trees	Students are able to reason about relationships when different trees (like rotations or subtrees) are presented.
5. Arguing and inferring	Students are able to use the depicted to form conclusions and predictions that go beyond the presented information

how MRCAs are graphically represented in tree diagrams. Although students at this level can describe the diagram and its elements, they lack the ability to interpret a given tree in a biologically meaningful way. Questions at this level typically ask for the meaning of diagrammatic elements and how to read and interpret them.

The second skill level (*handling apomorphies*) is about interpreting information regarding newly developed evolutionary traits called apomorphies. This information can be given in a textual, pictorial, or combined form and can be presented in the diagram in many ways, for example, directly along the branches of the tree, near the terminal nodes, or linked with arrows around the diagram. Tasks at this level typically require students to name the traits shown by a specific group or to identify if groups show a given list of characteristics.

The third level (*identifying relationships*) describes the skill to infer relative relationships of different groups and the formation of monophyletic groups in an evolutionary tree. This level encompasses the main use of evolutionary trees in modern biology, inferring evolutionary relationships from data. The most crucial aspect in tree-reading at this level is the ability to interpret the MRCAs of groups in a tree diagram, represented by internal nodes, which forms the basis for determining both monophyletic groups and relative relationships.

The fourth level (*comparing trees*) describes the ability to compare and contrast information from multiple evolutionary trees in order to identify similarities and differences. This can encompass understanding and identifying identical trees with rotated nodes but also the evaluation of whether different (sub-)trees agree with each other.

The fifth and final level (*arguing and inferring*) describes going beyond the information given in the diagram. Students at this level can formulate conclusions and predictions based on the presented phylogeny and even propose hypotheses about taxa or traits not presented. This level is based on Halverson and Friedrichsen's (2013) level 7 (expert use of representation) describing the skill level of experts in the field, typically not reached by novice students. Scientists at this level are able to quickly compare multiple representations and form mental models that bring these pieces of information together, creating a basis for complex inferences and arguments, and use information given in a tree to form predictions in complex problem-solving situations. As this skill level is typically not expected for student learners, it will not be investigated in this study.

Another approach to systemizing tree-reading was recently presented by Kummer *et al.* (2019). These authors developed and tested the Evolutionary Tree Concept Inventory (ETCI) based on the learning outcomes and organized them into five factors. In their work, they collected data on students' understanding of evolutionary trees through multiple-choice and free-response items, as well as student interviews. Based on these data, they developed learning outcomes that are similar to other published outcomes (Kummer *et al.*, 2019). The resulting 24-item concept inventory has five empirically distinguishable factors with a total of 11 learning outcomes (Table 2). As this work was published after the design of the STREAM and collection of data for the present study, it was not considered in the study design. Nevertheless, the implications and results of both models are discussed.

RESEARCH QUESTIONS

In this study, we used a newly developed testing instrument to assess students' ability to read evolutionary trees. By doing so, we aimed to assess whether we could empirically prove the assumed hierarchy of the STREAM by investigating the item difficulty. In particular, we asked the following:

RQ1: How far can the hierarchical nature of the STREAM be validated empirically?
RQ2: How far can the different skill levels of the STREAM be distinguished empirically?

METHODS

Research Design and Testing Instrument

A questionnaire survey based on a 28-item multiple-choice single-select instrument was administered in a test booklet design to undergraduate and graduate students at four universities in Germany. To prove validity evidence, we also conducted a think-aloud study with a small group of students.

Item Development

To develop items for the testing instrument, we analyzed existing tree-thinking diagnostic instruments regarding the compatibility of the item formats and item content, with the aim of focusing on different aspects of tree-reading based on the STREAM (Schramm *et al.*, 2019). Testing instrument items were then created that were similar to the existing items (Table 3), typically keeping the item stem as close to the original as possible and only changing the organisms to fit the item to the presented trees. When distractors included explanations (e.g., A1: What does the node L represent? A: The node represents the conjunction of the development lines of Kaluga and Chinese paddlefish.), we tried to keep the explanation as close to the original as possible.

As we were investigating the way students read evolutionary trees, we had to decide which trees the students should work on. We constructed a tree graph consisting of 13 terminal nodes and created a twin-tree with the same branching pattern by rotating multiple internal nodes, resulting in two superficially different tree diagrams with the same properties on a graph theoretical level. The trees were then filled with two different contexts (arthropods and fish). There were two reasons for this approach. First, we used trees with a relatively high number of nodes to enable generation of multiple complex items based on a single tree, particularly for the items about the higher skill levels, where a certain complexity of the presented tree was required. At the same time, a certain number of items for each skill level were needed to reliably investigate the hierarchy. To maintain the number of nodes and simplify the already complex structure of the task, we used the twin-tree approach. Second, we used fish and arthropod trees with the assumption that most students would not have an in-depth understanding of the relationships between the species within these groups. This was a tactic to prevent students from using prior knowledge of the presented trees. We refrained from using abstract trees (where terminal nodes are labeled with numbers or letters), as such trees are not representative of the trees biological scientists are typically expected to be able to read. Practicing biologists will typically read and work with concrete evolutionary trees that show explicitly represented species relationships. Therefore, the

TABLE 2. Learning outcomes of the ETCI (Kummer *et al.*, 2019)

Number	Learning outcome
1	Compare evolutionary relationships between taxa.
2	Distinguish between evolutionary trees with differing ordering of the species and evolutionary trees depicting differing evolutionary relationships.
3	Use an understanding of the theoretical aspects of evolutionary trees to evaluate group and character evolution based on common ancestry and parsimony.
A	Identify cases of homology and analogy when interpreting an evolutionary tree.
B	Analyze character information and evolutionary trees using parsimony.
C	Distinguish monophyletic, paraphyletic and polyphyletic groups.
D	Identify what the various components of an evolutionary tree represent.
4	Demonstrate an understanding of how characters are inherited from common ancestors by accurately interpreting an evolutionary tree with characters.
A	Identify synapomorphies for a group on a given evolutionary tree.
B	Identify character states as derived or ancestral on a given evolutionary tree.
D	Use an evolutionary tree to identify characters a given taxon would exhibit.
5	Demonstrate an understanding of evolution as a continuing and nonteleological process.
A	Identify why using simplicity and complexity to categorize organisms as primitive and advanced species is inappropriate from an evolutionary perspective.

choice of context was to enable investigation of the skills of the students in a manner most closely resembling the application of tree-reading in practice while reducing the impact of prior knowledge as much as possible.

Eight items were developed for skill levels 1 to 3, and four items for level 4. We developed fewer items for level 4, because comparing multiple trees is much more time-consuming than the tasks required for the other skill levels.

Fourteen items were designed for each of the two contexts (arthropods and fish). It is known that the order in which items are presented might influence the test performance based on fatigue effects or framing of the item (Lavrakas, 2008), or a different outcome may result based on which items have been solved before (Halverson *et al.*, 2013). To reduce the influence of item order on the results, we adopted two different approaches.

First, items in each context were arranged in two ascending series of assumed difficulty: two level 1 items were followed by two level 2 items, two level 3 items, and one item at level 4. This pattern was repeated once. In some cases, we deviated from this arrangement by swapping items to avoid large empty spaces. We did not choose a complete randomized approach; instead, we aimed to reduce the possibility of student demotivation by selectively preventing a difficult item from being the first one. Therefore, the instrument started with a presumably easy item. Second, two different test booklets were developed, in which the order of the two contexts was reversed; one student group received the A items about fish first, and the other student group received the B items about arthropods. This was implemented to reduce effects of sequence or fatigue. The final instrument can be found in Supplement 1.

Participants

Responses were obtained from $N = 455$ biology students (undergraduates and graduates) from four different universities in Germany. An additional 137 students participated in the pretest versions of the instrument but did not participate in the final survey. Students participated on a voluntary incentivized basis,

TABLE 3. Overview of item levels, contexts, and the use of similar items in other works

Item number[a]	Level[b]	Sources of similar items
A1	1	Naegle, 2009
A2	1	Gibson and Hoefnagels, 2015
A7, B1, B2, B8, B9	1	Naegle, 2009; Gibson and Hoefnagels, 2015
A3, A10, B3, B10	2	Baum *et al.*, 2005; Naegle, 2009; Halverson *et al.*, 2011; Gibson and Hoefnagels, 2015; Leone, 2017
A11, B11	2	Baum *et al.*, 2005; Halverson *et al.*, 2011; Catley *et al.*, 2013; Leone, 2017
A4, B4	2	Catley *et al.*, 2013
A5, A12, B5, B12	3	Catley *et al.*, 2013
A6, B7, B13	3	Naegle, 2009; Catley *et al.*, 2013; Gibson and Hoefnagels, 2015
A13	3	Baum *et al.*, 2005; Halverson *et al.*, 2011; Novick *et al.*, 2010; Leone, 2017
A9, B6	4	Baum *et al.*, 2005; Naegle, 2009; Halverson *et al.*, 2011; Catley *et al.*, 2013; Gibson and Hoefnagels, 2015; Blacquiere and Hoese, 2016; Leone, 2017
A14, B14	4	Baum *et al.*, 2005; Naegle, 2009; Catley *et al.*, 2013; Gibson and Hoefnagels, 2015; Blacquiere and Hoese, 2016

[a]Item numbers represent the order in which participants received the items. A items are about the context of fish, and B items are about arthropods.
[b]Skill levels 1 and 2 consist of seven items, as one was excluded on each level based on item performance. Level 3 consists of eight items and level 4 of only four items, as items at this level are much more time-consuming to answer.

with the prospect of winning a voucher for a major online shop. Furthermore, we conducted a think-aloud study with eight students who did not participate in the pretest or final survey to further investigate the validity of the evidence.

Data acquisition took place during life science courses. Most participants were studying biology for teaching purposes, and data were collected during their biology education courses. Additionally, data were collected during a first-year microbiology course, attended by biologists and aspiring teachers. The gender distribution of the sample population (66.8% female) is not unusual for biology and biology education courses (e.g., Becker *et al.*, 2010; Eddy *et al.*, 2014).

Item Analysis

We assessed 455 undergraduate and graduate students using the tree-reading instrument, and the responses were analyzed following classical test theory (CTT) and item response theory (IRT), using the TAM package (Robitzsch *et al.*, 2019) in R software (R Core Team, 2017). Item difficulty, student abilities, and item fits were calculated following IRT.

Testing following IRT explains the relationship between latent traits and their manifestations. It considers the relationship of an individual's performance on test items, as well as the properties of items on an instrument (Boone *et al.*, 2014). As it is a probabilistic approach, it enables the estimation of student ability and item difficulty on the same scale based on the pattern of the responses observed. These measures can be used to create a Wright map that shows the arrangement of item difficulty and participants' ability by converting raw scores into logarithmic units called logits. Wright maps can provide a good impression of how the modeled skills are organized (Wright and Masters, 1982; Boone *et al.*, 2014), and they can provide a strong argument for the interpretation of student abilities (Wilson, 2005). Within a Wright map, participants are represented by a function of their ability, whereas items represent a function of their difficulty along the same logit scale. More able persons and more difficult items are represented at the higher end of the logit scale, whereas less able persons and easier items are represented at the lower end of the scale (Wilson, 2005; Boone *et al.*, 2014). In IRT, reliability can be measured via expected a posteriori (EAP) reliability. The EAP reliability values can be interpreted much like Cronbach's alpha (Bond and Fox, 2015).

We chose to base our investigation on IRT, as it enabled us to obtain a good estimate of how difficult the test items were and to what extent they could be differentiated into different groups, in this case, by conducting a dimensional analysis. Dimensional analysis can provide insights into the factor structure and organization of an investigated instrument and is viewed as the IRT equivalent to confirmatory factor analysis (Wirth and Edwards, 2007; Immekus *et al.*, 2019).

We investigated item difficulty, because we wanted to determine whether the hierarchy of the skill levels could be represented (RQ1). If skills follow a linear scale, items corresponding to the skills should show differences in their individual difficulties. We used dimensional analysis, because we wanted to investigate the extent to which different skill levels could be differentiated empirically (RQ2).

Before statistically analyzing the data, we searched for missing values. Where all items on a page were not answered, these items were considered as skipped by accident, and therefore as missing values. If a single item was skipped, it was considered as wrongly answered. If students stopped at a point and did not answer all the remaining items, these items were also deemed missing values and thus were considered in the IRT analysis. We did this so as not to penalize possible test time constraints or overlooked pages due to double-sided printing.

Think-Aloud Study

To investigate the instrument's validity, we conducted an additional think-aloud study (Ericsson and Simon, 1984) involving eight students. Each participant was asked to work on a subsample of the STREAM test with a total of five items per participant: one item for each of skill levels 1, 2, and 4, and two items for skill level 3. We chose to present two items on level 3, as this skill level encompasses items about both monophyletic groups and inferring relative relationships. Two sets of five items each were created, and the participants were randomly assigned one of the sets.

While think-aloud, the participants verbalized everything that came to mind as they worked on the presented items. Their verbalizations were recorded, and a partial transcript of all topic-related passages was created for further analysis. The resulting protocols provided insights into their cognitive processes and rationales for responding to the presented items and enabled analysis of the extent to which the participants understood and worked on the items in the intended manner.

RESULTS

Think-Aloud Study

In this section, we briefly explain the think-aloud participants' rationales for the different aspects of the skill model, focusing on whether the responses were correct, incorrect and based on expected misconceptions, or incorrect for other reasons, possibly indicating the existence of difficulties with the instrument.

Of the 40 responses recorded, 23 were correct and based on scientifically accurate arguments. Three responses were correct, although based on one or more common misconceptions. Two responses were incorrect, despite being based on scientifically accurate arguments. Nine responses based on common misconceptions did not match the correct answer. One response was incorrect without referencing a common misconception. This student attributed an apomorphy to both sister groups in item B3, thus answering the item incorrectly. Finally, two responses were provided without any rationale or indication that the answers were chosen randomly. When first reading the items about apomorphies (A12 and B12), six out of the eight participants read the multiple-choice options in columns rather than rows. After thinking about these items, the participants realized that the options were arranged in rows and were then able to continue working on the items. However, through this, we realized that the arrangement of the response options was disadvantageous, even though all participants understood the correct way to read these items. A detailed description of the way students reasoned each item is provided in Appendix 2.

The results of the think-aloud study indicated that most students either responded correctly or followed expected misconceptions. Only one response was based on an unexpected mistake; however, this did not indicate a problem with the instrument, but more likely a mistake out of carelessness, as the student interpreted similar connections correctly.

Item Analysis

To prove the test quality, we considered item fit measures as a first step (Bond and Fox, 2015). Item infit and outfit provide information about how accurately the observed data fit the estimated IRT model. Items with good infit provide more information about students' abilities close to the item's difficulty, whereas items with a good outfit provide broader information about students' abilities at all levels. A perfect infit or outfit is achieved at a value of 1.0. Deviation from this value indicates that the students' abilities are either overestimated or underestimated. Depending on the context of the test situation, different thresholds for item fits are used; typically values between 0.7 and 1.3 are seen as acceptable (Wilson, 2005; Boone *et al.*, 2014).

Two items (A8 and B13) showed fit values beyond the acceptable range and therefore were removed from the item pool. Both items had the correct answer expressed as "all other options are equally correct," which might have confused the participants. A pattern regarding tree-reading skills was not apparent. As the results of IRT analysis are always based on all examined items, the calculation was repeated without these items. Hence, they were not factored into further analyses.

We calculated item difficulty, infits and outfits, and students' abilities using a unidimensional IRT analysis of the 26 items remaining after the initial analysis. An overall EAP reliability of 0.766 was obtained.

The unidimensional analysis showed item outfits (M = 0.995 ± 0.111), ranging from 0.749 (item B11) to 1.193 (item B9). Item infit was, on average, 0.997 ± 0.069, with a minimum of 0.874 (item A11) and a maximum of 1.159 (item B9). Except for one item (A11), all fit values lay within a reasonable range for high-stakes assessments (Boone *et al.*, 2014). As item A11 was still in the fit range for run-the-mill multiple-choice tests (Boone *et al.*, 2014), it was still included in the investigation. Furthermore, it showed a downward deviation, indicating a stronger deviation than expected. Item difficulty was, on average, –0.030 ± 0.908 logit, with a range of –1.555 (item A11) to 1.577 (item B8). For detailed values of all items, see Appendix 1.

We investigated the hierarchical nature of the skill system to answer RQ1. This required examination of the difficulty of each item. The results of the dimensional analysis were used to answer RQ2.

Item Difficulty and Item Levels

If one group of skills is on a lower level than another group, its items should be easier than those in the higher skill-level group. To investigate the hierarchical nature of the skill system (RQ1), we examined the item difficulty resulting from the unidimensional IRT analysis regarding the different assumed hierarchical skill levels.

Items at skill level 1 (identifying structures) showed an average difficulty of 0.091 ± 0.751 logit with a total range of 2.16. Skill level 2 (handling apomorphies) had an average difficulty of –0.673 ± 0.587 with a total range of 1.717 and was the easiest skill level. Skill level 3 (identifying relationships) showed a difficulty of 0.231 ± 0.733 with a range of 2.213, and level 4 (comparing trees) had an average difficulty of 1.01 ± 0.750 with a range of 0.491 and was the most difficult level (Appendix 1 in the Supplemental Material). The Wright map (Figure 1) provides a more detailed view of the distribution of item difficulty in relation to skill levels. Levels 1, 2, and 3 showed a span of roughly two logits, whereas level 4 (consisting of only four items) was spread over approximately 0.5 logits. While level 2 showed a rather steady distribution, levels 1 and 3 showed a lumping of five items, with two items being much more difficult at each level. On level 4, two pairs of almost equally difficult items were found.

Regarding the assumed hierarchy of the skill levels, the mean value of level 1 was higher than that of levels 2 and 3, and the mean value increased from level 2 to 4, reflecting an increase in item difficulty across these levels. Another important finding is that the three easiest items were at skill level 2. Five items (four on level 2 and one on level 3) were easier than the easiest item at level 1. Considering the difficult items, the most difficult item was at level 1, and the second most difficult item was at level 3. Furthermore, the 95% confidence intervals of the item difficulties overlapped when the different item levels were compared.

To further investigate whether items form a hierarchical organization, we conducted an analysis of variance (ANOVA) to examine the assumed skill hierarchy. Prerequisites for the ANOVA were checked and fulfilled. As the groups that were compared varied in the number of items, Gabriel's procedure was chosen as a post hoc test (Field, 2018). There was a significant effect of the item level on the item difficulty; $F(3, 22) = 6.855$; $p = 0.002$, $\eta^2 = 0.483$. Post hoc procedures revealed significant differences between levels 1 and 2 ($p = 0.029$) and between levels 2 and 4 ($p = 0.002$). The other levels did not differ significantly among the groups.

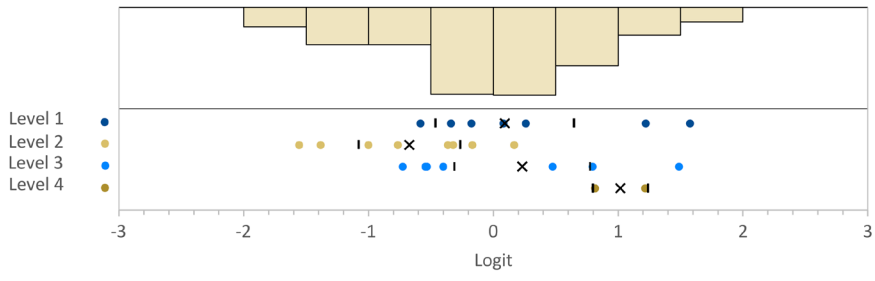

FIGURE 1. Wright map of the 26 investigated items, arranged according to their respective skill level (bottom) and an overview of students' abilities (top). An "X" indicates the average difficulty of the skill level; black bars represent the 95% confidence intervals of that level. Items marked in red are monophyletic groups.

Dimensional Analysis

To investigate the extent to which the different skill levels of the STREAM can be differentiated empirically, we conducted dimensional analyses. In total, six different models were tested: a one-dimensional (1D) model (all skill levels); a two-dimensional (2D) model (skill level 1 vs. 2, 3, and 4); a three-dimensional (3D) model (skill level 1 vs. 2, and 3 vs. 4); a four-dimensional (4D) model (skill level 1 vs. 2 vs. 3 vs. 4); a revised four-dimensional (4D)

TABLE 4. Comparison of different models of dimensionality

Model	Compared skill levels	Deviance	Npars[a]	AIC	BIC
1D		13,910.27	27	13,964.27	14,075.52
2D	1 vs. 2,3, and 4	13,857.53	29	13,915.53	14,035.02
3D	1 vs. 2 and 3 vs. 4	13,791.46	32	13,855.46	13,987.31
4D	1 vs. 2 vs. 3 vs. 4	13,916.98	36	13,988.98	14,137.32
4D rearranged	1 vs. 2r[b] vs. 3r [b] vs. 4	13,778.91	36	13,850.91	13,999.24
5D	1 vs. 2 vs. 3.1[c] vs. 3.2[c] vs. 4	13,674.90	41	13,756.90	13,925.83

[a]Npars, number of parameters.
[b]Four items of level 3 were moved to level 2.
[c]Level 3 was split into two levels: 3.1 and 3.2.

model (1 vs. revised 2 vs. revised 3 vs. 4); a five-dimensional (5D) model (skill level 1 vs. 2 vs. 3.1 vs. 3.2 vs. 4). In the 5D model, we investigated the following dimensions: 1) identifying structures; 2) handling apomorphies; 3.1) determining monophyletic groups; 3.2) identifying relationships; 4) comparing trees.

The 2D model was chosen because skill level 1 encompasses knowledge about the diagrammatic properties of evolutionary trees, whereas skill levels 2 to 4 have a stronger focus on explicit reading abilities. The 3D model was chosen because the model of Halverson and Friedrichsen (2013) combines two skill levels into one, whereas Novick and Catley (2016) viewed them as separate skills. The 4D model was chosen to investigate the extent to which all skill levels could be represented in empirically differentiable dimensions. The revised 4D and the 5D models were investigated because skill level 3 consists of skills for deducing relative relationships and investigating monophyletic groups, which Novick and Catley (2016) regarded as separate skills. We wanted to investigate whether these skills belonged to different dimensions (5D model) or whether the items about monophyletic groups were more appropriate as belonging to skill level 3 (revised 4D model).

Dimensional analysis was conducted using the TAM package (Robitzsch *et al.*, 2019) in R software (R Core Team, 2017). To investigate which model shows the best fit, we focused on the Bayesian information criterion (BIC), although the Akaike information criterion (AIC) showed the same result. Both AIC and BIC are penalty counters, and a lower score represents a better fit of the model (de Ayala and Kenny, 2009). A comparison of the BIC values of the six investigated models showed that the 5D model was superior to all other models (Table 4). Furthermore, it showed satisfactory EAP reliabilities of 0.604 (level 1), 0.711 (level 2), 0.703 (level 3.1), 0.536 (level 3.2), and 0.535 (level 4).

To further corroborate these findings, we conducted a deviance likelihood ratio test to compare the 5D model with the baseline 1D model (Wu and Vos, 2018). The test showed that the 5D model represented the data more significantly than the 1D model (LRT = 235.37; $df = 14$; $p < 0.001$).

After determining which model was the best fit for the data, we investigated the internal structure of this model. Even though a multidimensional model consists of different latent traits, the relationship between these traits can be very close or distant. Calculating the correlation between the latent traits permits deeper insight into how closely the traits are related to each other. The dimensions of the 5D model showed correlations between 0.322 (level 3.1 with level 4) and 0.659 (level 2 with level 3.1; Table 5). Levels 3.1 and 3.2 correlated weakly with each other and with level 4, whereas higher correlations were observed between both levels 1 and 2 and all the other skill levels. These correlations further support the multidimensional structure of the model.

DISCUSSION

Knowledge about evolution in general, and tree-reading in particular, is regarded as increasingly important for modern scientific literacy. Consequently, educators need to think about how to best teach tree-reading skills. As a basis for developing learning environments for any topic, it is helpful to understand the topic from an educational perspective. Knowledge about topic-related skills or the content that students need to grasp can greatly influence how a learning environment should be designed. Based on empirically untested tree-reading skill models in the published literature, we developed STREAM, a synthetic hierarchical tree-reading skill system (Schramm *et al.*, 2019). In the present study, we tested STREAM empirically by interviewing eight students and surveying 455 undergraduate and graduate students using a newly developed tree-reading measurement instrument.

Think-aloud results indicated that the participating students understood the survey items in the way that was expected by the researchers. The errors that occurred were typically linked to misconceptions considered during item generation. These results can be seen as indicators of validity evidence. Furthermore, the fit values of the IRT model also indicate validity. In addition, the created items were similar to the existing items, and all aspects of the STREAM were represented in the test items, further supporting content validity, which is also corroborated by the

TABLE 5. Correlation matrix of the 5D model

Dimensions	Level 1	Level 2	Level 3.1	Level 3.2	Level 4
Level 1	1	0.481	0.571	0.639	0.433
Level 2		1	0.659	0.503	0.498
Level 3.1			1	0.370	0.322
Level 3.2				1	0.333
Level 4					1

fact that all skills reported by other authors are represented in the STREAM.

Regarding the first research question (RQ1: How far can the hierarchical nature of the STREAM be validated empirically?), the results do not provide a clear answer. Skill levels 2 to 4 appeared to show an ascending order of difficulty based on their mean difficulty values; however, we could not definitively identify this as a hierarchy, because the different levels formed very broad ranges. Except for level 4, all other levels covered approximately half of the total difficulty range. Furthermore, the ranges of different item levels strongly overlapped, reflecting an unclear hierarchy. This was further corroborated by the overlap of the confidence intervals, which usually reflects nonhierarchical organization, although overlapping confidence intervals are more an indication than proof of significance (Cumming and Finch, 2005). Furthermore, the ANOVA showed that only some singular comparisons were significantly different. Another important factor contradicting the assumed hierarchy is that there were five items on levels 2 and 3 that were easier than the easiest item at the first skill level. Furthermore, the most difficult item was also at the first skill level. Based on the average difficulties, level 1 was the second most difficult skill level, strongly contradicting the assumed hierarchy.

An increase in difficulty was identifiable from levels 2 to 4. This increase did not need to evidence clear hierarchical organization of the levels to point to a learning progression. In fact, the increase in average difficulty may have originated from an increase in the complexity of the reading tasks. Items at level 2 required students to investigate apomorphies. These tasks could be solved by tracking the developmental lines of a single species until all distractors were rejected. At level 3, the students were required to investigate monophyletic groups, and by tracing the lines from a single internal node to the terminal nodes, they could determine which groups formed a monophylum. At level 3, students were also expected to determine relative relationships. This required them to compare different sets of relationships, taking multiple patterns of relationships into account at the same time, to find the correct answer. At level 4, they were expected to compare different trees. This required consideration of a large amount of information and comparison of multiple complex diagrams. Therefore, it could be argued that increase in difficulty does not necessarily reflect a hierarchy of tree-reading skills in the sense that one skill forms the basis for another one. Rather, it may merely reflect that skills and tasks tend to require more complex reading processes.

In summary, the results of the unidimensional IRT analysis showed that some skill levels could be distinguished empirically based on their item difficulty; however, most evidence did not corroborate the assumed hierarchy. Although skill level 2 (handling apomorphies) is significantly easier than level 1 (identifying structures) and level 5 (comparing trees), items on all skill levels showed a strong overlap of item difficulty, and evidence for the assumed hierarchy of skill levels was not convincing.

Considering the second research question (RQ2: How far can the different skill levels of STREAM be distinguished empirically?), the dimensional analysis showed that different skill levels form different dimensions. As the best-fit model was a 5D model, it seemed reasonable to subdivide level 3 (determining relationships) into two separate dimensions: "identifying relationships" and "identifying monophyletic groups."

The level 3 division is founded in theory: In Halverson and Friedrichsen's model (2013), different reading aspects are grouped into one skill level, together with others determining relative relationships and monophyletic groups. Novick and Catley (2013) regard these two as separate skills. They define the skill "understand the concept of a clade" as one of the five core components of tree-thinking, and "evaluate relative evolutionary relatedness" is defined as a separate core skill (Novick and Catley, 2013). In their revised model, they no longer use the term "core skills," but present the following skills: "identify/evaluate clades" and "identify nested clades," as well as "evolutionary relationship: resolved structure" and "evolutionary relationship: polytomy." Regarding the first two skills, students were asked to evaluate whether a set of taxa formed a clade and to mark all the nested clades in a cladogram. For the latter two skills, students were asked to assess the relative relationship of species in a three-taxon statement with resolved and polytomous data.

Even though both pairs of skills are heavily based on the interpretation of inner nodes as MRCAs, the approaches to solving tasks related to these skills show clear differences. While identifying monophyletic groups, students must look for the most recent common ancestor of all investigated taxa and then determine whether there are species that are part of the descendants of that node that are not part of the investigated taxa. While identifying relative relationships, students must solve multiple three-taxon statements (Novick and Catley, 2016). In the original construction of the STREAM, we followed Halverson and Friedrichsen's (2013) approach to group these skills. Based on the dimensional analysis, it seemed more reasonable to separate these skills.

Not only did we find evidence in the penalty scores (e.g., BIC) that the test we used represents different dimensions, but we also found that the correlations between these dimensions were not high enough to assume only one construct. The intercorrelations between the dimensions ranging from 0.322 (level 3.1 and level 4) to 0.659 (level 2 and level 3.1; Table 5) are, for example, lower than those among different subject areas in PISA, which typically have values greater than 0.8 (Bond and Fox, 2015). In PISA 2009, for example, correlations between mathematics and reading were at 0.82 and between mathematics and science at 0.88 (OECD, 2012) Different latent traits showing a low intercorrelation typically indicate that, to a degree, these traits represent distinct constructs without differing completely. Therefore, the results of the dimensional and correlation analyses strongly support an argument for viewing the different dimensions as distinct skills rather than part of a hierarchical organization. This argument is also supported by overlapping item difficulties and confidence intervals.

In conclusion, we found evidence that the assumed skills form different dimensions and therefore should not be seen as a cohesive construct. Beyond that, we could argue that splitting level 3 into two dimensions improves the model. However, we did not find strong supporting evidence for the assumed hierarchy.

Revised STREAM

The analyses conducted in this study showed that the data did not corroborate the assumed hierarchical organization of the

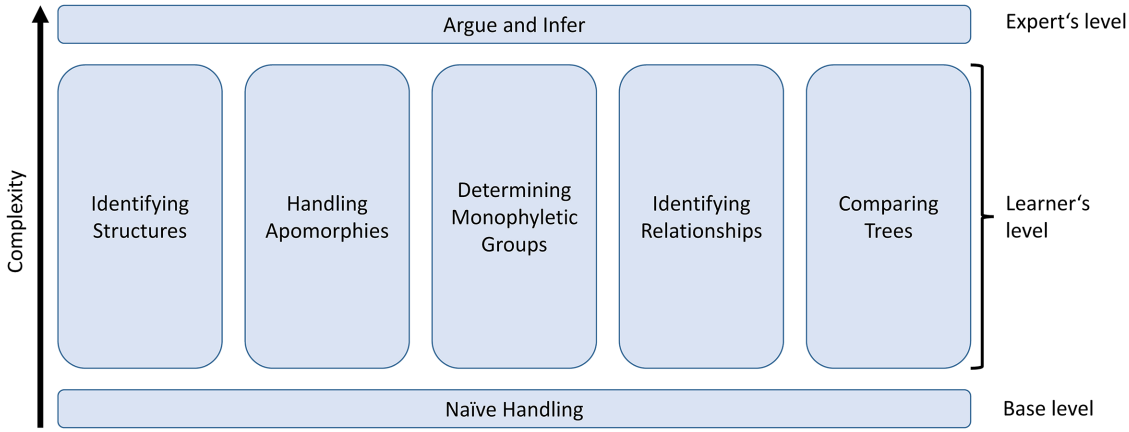

FIGURE 2. Revised STREAM. In this model, naïve handling is a baseline level, representing preconceptions about evolutionary trees. On the next level, students learn about different aspects of reading evolutionary trees, resulting in five distinct skills. Finally, the argue and infer level combines knowledge about all five skills. Experts use their knowledge of all skills and other concepts to interpret trees and infer information.

STREAM, indicating that the system needs modification. Although we found some significant differences among the item levels or dimensions in terms of item difficulty, we could not depict the assumed hierarchical nature of the model. This is because the required knowledge about the meaning of diagram elements is more complex at the first level than predicted from the theory. In addition, the other assumed item levels did not follow a strict hierarchical organization. Based on these findings, there was no evidence supporting the assumed hierarchy of the investigated skills, because item difficulty was spread across all levels, and the different levels showed strong overlaps.

Based on the dimensional analysis, we showed that different aspects of tree-reading could be differentiated from each other empirically. As the best model investigated was the 5D approach, skill level 3 should be split into two different skills. This is further supported by the fact that the correlations between the different dimensions are not high enough to conclude that the items form one cohesive construct; instead, the items are distinct latent traits.

Based on our findings, we concluded that a revised version of the STREAM is needed and that the hierarchical nature of the model should be discarded. Furthermore, skill level 3 should be separated into two distinct skills, resulting in a model of six skills in total, as shown in Figure 2.

In the revised STREAM, the base level (naïve handling of evolutionary trees) remains unchanged. Students at this level have no understanding or only very limited knowledge about evolutionary trees and how to read them. Above this level, students show varying degrees of skill in different skill dimensions. In the revised STREAM, the skills previously identified as skill levels are included as separate skill dimensions next to each other. There may be links between skill dimensions, which can vary in their difficulty, but no hierarchical organization of the skills is assumed. The third level represents the implementation and application of all skills in the model in overarching argumentation and complex problem-solving situations. This level corresponds to Halverson and Friedrichsen's seventh level called "expert use of representation," which describes "experts in the field of systematics" and the level being "not appropriate for beginning students" (Halverson and Friedrichsen, 2013, p. 196).

As the ETCI was not published when the STREAM was developed and tested, it could not be included in the theoretical foundation of the STREAM. To form a link between these two models, we aligned the learning outcomes of the ETCI with the skills of the STREAM (Table 6) and found that the two systems encompassed similar concepts. Nearly all aspects of the ETCI can also be found in STREAM, and vice versa. Nevertheless, the two systems have a different focus: the ETCI is designed as a concept inventory of tree-thinking, whereas the STREAM focuses on tree-reading. Kummer and colleagues (2019) showed that the ETCI is focused on conceptual understanding of evolutionary trees in biology, but that it does not necessarily reflect the tasks and actions students are expected to complete when working with evolutionary trees. Therefore, we view these two instruments as supplemental approaches with a different focus that show comparable results.

The findings of this study represent the first empirical investigations of the STREAM based on the system of Halverson and Friedrichsen (2013) and Novick and Carley (2016) and provide further insights into how tree-reading skills are systemized. However, this study had certain limitations, all of which provide scope for further research.

In contrast to most studies on tree-thinking, we tested German students rather than U.S. students. Although we do not expect tree-reading processes to work differently, depending on cultural or national backgrounds, differing educational systems could have an effect on tree-reading. To allow for an easier transfer of these results in future, we plan to perform a comparative study of skills and typical mistakes made by German and American students.

Based on the item difficulty analysis, we obtained no evidence that supports the assumed hierarchical nature of the investigated skills. However, these data were acquired through student responses to multiple-choice items on a single test date. Therefore, we were not able to investigate how skills develop or which conceptions students hold. Consequently, we cannot say

TABLE 6. Alignment of the ETCI learning outcomes with the skills of the STREAM

ETCI learning outcomes	Corresponding STREAM skills
1. Compare evolutionary relationships between taxa.	Identifying relationships
2. Distinguish between evolutionary trees with differing ordering of the species and evolutionary trees depicting differing evolutionary relationships.	Comparing trees
3. Use an understanding of the theoretical aspects of evolutionary trees to evaluate group and character evolution based on common ancestry and parsimony.	Argue and infer
a. Identify cases of homology and analogy when interpreting an evolutionary tree.	Argue and infer
b. Analyze character information and evolutionary trees using parsimony.	Argue and infer
c. Distinguish monophyletic, paraphyletic, and polyphyletic groups.	Determining monophyletic groups
d. Identify what the various components of an evolutionary tree represent.	Identifying structures
4. Demonstrate an understanding of how characters are inherited from common ancestors by accurately interpreting an evolutionary tree with characters.	Handling apomorphies
a. Identify synapomorphies for a group on a given evolutionary tree.	Handling apomorphies
b. Identify character states as derived or ancestral on a given evolutionary tree.	Handling apomorphies
c. Use an evolutionary tree to identify characters a given taxon would exhibit.	Handling apomorphies
5. Demonstrate an understanding of evolution as a continuing and nonteleological process	Argue and infer
a. Identify why using simplicity and complexity to categorize organisms as primitive and advanced species is inappropriate from an evolutionary perspective.	Argue and infer
b. Demonstrate an understanding that all extant populations continue to evolve and have evolved throughout their entire existence.	Argue and infer

for sure that the skill system does not follow a hierarchical nature. Additionally, further studies following a longitudinal approach that tracks students' learning processes could provide relevant clarifications. Additionally, it could be worthwhile to test novice students' ability to read evolutionary trees against the ability of those considered more advanced or experts in the field along a carefully chosen sample of differing expected skill levels. By investigating the full spectrum of assumed students' skills, more insight into how these skills develop could be gained.

Additional research is needed to further validate and investigate the proposed skill system and the corresponding measurement instruments. One important approach is to test the instrument with a larger sample, potentially with students from a different country. Further interviews should be conducted to acquire qualitative data on all test items. To further corroborate the findings, it would be worthwhile to test students using abstract trees or trees about which they could be expected to have prior knowledge.

To shed more light on how different items are interrelated, it would be worthwhile investigating how individual items are solved at each level and whether specific response patterns in one dimension are related to performance or misconceptions in other item dimensions. A different approach is to make the students the focus of the investigation. Cluster analyses would help identify groups of students answering in a similar way, potentially representing typical misconceptions or types of reasoning. Identifying these types of groups could be beneficial for constructing new learning environments.

IMPLICATIONS

The findings of this study are important for educators designing learning environments for teaching evolutionary biology and tree-reading. Previously, it was assumed that tree-reading skills follow a hierarchical organization (Halverson and Friedrichsen, 2013) and that students need to study one skill after another in succession to achieve a deeper understanding of tree-reading. The findings of this study do not support the hierarchical nature of tree-reading skills. Nevertheless, it was found that the different skills could be differentiated empirically, implying that students can be proficient in some aspects of tree-reading, independent of other aspects. Because different tree-reading skills appear to form different dimensions, it seems necessary to teach all aspects to beginner tree readers.

REFERENCES

Baum, D. A., & Smith, S. D. (2013). *Tree thinking: An introduction to phylogenetic biology*. Greenwood Village, CO: Roberts.

Baum, D. A., Smith, S. D., & Donovan, S. S. S. (2005). Evolution. The tree-thinking challenge.*Science (New York, N.Y.)*, *310*(5750), 979–980. https://doi.org/10.1126/science.1117727

Becker, R., Casprig, A., Kortendiek, B., Münst, S., & Schäfer, S. (2010). *Gender-Report 2010: Geschlechter(un)gerechtigkeit an nordrhein-westfälischen Hochschulen*. Essen, Germany: Netzwerk Frauen- und Geschlechterforschung NRW.

Blacquiere, L. D., & Hoese, W. J. (2016). A valid assessment of students' skill in determining relationships on evolutionary trees. *Evolution: Education and Outreach*, *9*(1), 979. https://doi.org/10.1186/s12052-016-0056-9

Bokor, J. R., Landis, J. B., & Crippen, K. J. (2014). High school students' learning and perceptions of phylogenetics of flowering plants. *CBE—Life Sciences Education*, *13*(4), 653–665. https://doi.org/10.1187/cbe.14-04-0074

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd edition). New York, NY: Routledge Taylor & Francis Group.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, Netherlands Springer.

Catley, K. M., Novick, L. R., & Funk, D. (2012). The Promise and Challenges of Introducing Tree Thinking into Evolution Education. In Rosengren, K. S., Brem, S. K., & Evans, E. M. (Eds.), *Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution* (pp. 93–101). New York, NY: Oxford University Press.

Catley, K. M., Novick, L. R., & Shade, C. K. (2010). Interpreting evolutionary diagrams: When topology and process conflict. *Journal of Research in Science Teaching*, *47*(7), 861–882. https://doi.org/10.1002/tea.20384

Catley, K. M., Phillips, B. C., & Novick, L. R. (2013). Snakes and Eels and Dogs!: Oh, My! Evaluating High School Students' Tree-Thinking Skills: An Entry Point to Understanding Evolution. *Research in Science Education*, *43*(6), 2327–2348. https://doi.org/10.1007/s11165-013-9359-9

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist*, *60*(2), 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London, Great Britain: Murray.

de Ayala, R. J., & Kenny, D. A. (2009). *The theory and practice of item response theory. Methodology in the social sciences*. New York, NY: Guilford Press. http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10468533

Dees, J., Momsen, J. L., Niemi, J., & Montplaisir, L. (2014). Student interpretations of phylogenetic trees in an introductory biology course. *CBE—Life Sciences Education*, *13*(4), 666–676. https://doi.org/10.1187/cbe.14-01-0003

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*(3), 478–492. https://doi.org/10.1187/cbe.13-10-0204

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambrige, MA: The MIT Press.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th edition). Los Angeles, CA: Sage.

Gibson, J. P., & Hoefnagels, M. H. (2015). Correlations Between Tree Thinking and Acceptance of Evolution in Introductory Biology Students. *Evolution: Education and Outreach*, *8*(1), 1891. https://doi.org/10.1186/s12052-015-0042-7

Gotwals, A. W., & Alonzo, A. C. (2012). Introduction: Leaping into Learning Progressions in Science. In Alonzo, A. C. (Ed.), *Learning progressions in science. Learning progressions in science: Current challenges and future directions* (pp. 3–12). Rotterdam, Netherlands: Sense Publ.

Gregory, T. R. (2008). Understanding Evolutionary Trees. *Evolution: Education and Outreach*, *1*(2), 121–137. https://doi.org/10.1007/s12052-008-0035-x

Halverson, K. L. (2011). Improving Tree-Thinking One Learnable Skill at a Time. *Evolution: Education and Outreach*, *4*(1), 95–106. https://doi.org/10.1007/s12052-010-0307-0

Halverson, K. L., Boyce, C. J., & Maroo, J. D. (2013). Order matters: pre-assessments and student generated representations. *Evolution: Education and Outreach*, *6*(1). https://doi.org/10.1186/1936-6434-6-24

Halverson, K. L., & Friedrichsen, P. (2013). Learning Tree Thinking: Developing a New Framework of Representational Competence. In Treagust, D. F. & Tsui, C.-Y. (Eds.), *Models and Modeling in Science Education. Multiple Representations in Biological Education* (pp. 185–201). Dordrecht, Netherlands: Springer.

Halverson, K. L., Pires, C. J., & Abell, S. K. (2011). Exploring the complexity of tree thinking expertise in an undergraduate systematics course. *Science Education*, *95*(5), 794–823. https://doi.org/10.1002/sce.20436

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, *4*, Article 45. https://doi.org/10.3389/feduc.2019.00045

Kummer, T. A., Whipple, C. J., Bybee, S. M., Adams, B. J., & Jensen, J. L. (2019). Development of an Evolutionary Tree Concept Inventory. *Journal of Microbiology & Biology Education*, *20*(2). https://doi.org/10.1128/jmbe.v20i2.1700

Kummer, T. A., Whipple, C. J., & Jensen, J. L. (2016). Prevalence and Persistence of Misconceptions in Tree Thinking. *Journal of Microbiology & Biology Education*, *17*(3), 389–398. https://doi.org/10.1128/jmbe.v17i3.1156

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. SAGE Publications. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=474384

Leone, A. E. (2017). *An Investigation of Relationships between Student Acceptance of Evolution, Tree-Thinkin, and Eye Movement among different Instructioinal Interventions [Master Thesis]*. Texas State University, San Marcos.

Meir, E., Perry, J., Herron, J. C., & Kingsolver, J. (2007). College Students' Misconceptions About Evolutionary Trees. *The American Biology Teacher*, *69*(7), e71–e76. https://doi.org/10.1662/0002-7685(2007)69[71:CSMAET]2.0.CO;2

Meisel, R. P. (2010). Teaching Tree-Thinking to Undergraduate Biology Students. *Evolution*, *3*(4), 621–628. https://doi.org/10.1007/s12052-010-0254-9

Naegle, E. (2009). *Patterns of thinking about phylogenetic trees:: A study of students and the potential of tree thinking to improve comprehension of biological concepts [Dissertation]*. Idaho State University, Idaho.

Novick, L. R., & Catley, K. M. (2013). Reasoning About Evolution's Grand Patterns: College Students' Understanding of the Tree of Life. *American Educational Research Journal*, *50*(1), 138–177. https://doi.org/10.3102/0002831212448209

Novick, L. R., & Catley, K. M. (2016). Fostering 21st-Century Evolutionary Reasoning: Teaching Tree Thinking to Introductory Biology Students. *CBE—Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.15-06-0127

Novick, L. R., Catley, K. M., & Funk, D. (2010). Characters Are Key: The Effect of Synapomorphies on Cladogram Comprehension. *Evolution: Education and Outreach*, *3*(4), 539–547. https://doi.org/10.1007/s12052-010-0243-z

OECD. (2012). *Pisa 2009 Technical Report*. OECD Publishing. http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10595644

O'Hara, R. J. (1988). Homage to Clio, or, Toward an Historical Philosophy for Evolutionary Biology. *Systematic Zoology*, *37*(2), 142. https://doi.org/10.2307/2992272

Omland, K. E. (2014). Interpretation of Phylogenetic Trees. In Losos, J. B. (Ed.), *Princeton reference. The Princeton guide to evolution* (pp. 51–59). Princeton, NJ: Princeton University Press.

Omland, K. E., Cook, L. G., & Crisp, M. D. (2008). Tree thinking for all biology: The problem with reading phylogenies as ladders of progress. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *30*(9), 854–867. https://doi.org/10.1002/bies.20794

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM

Schramm, T., Schachtschneider, Y., & Schmiemann, P. (2019). Understanding the tree of life: an overview of tree-reading skill frameworks. *Evolution: Education and Outreach*, *12*(1). https://doi.org/10.1186/s12052-019-0104-3

Schramm, T., & Schmiemann, P. (2019). Teleological pitfalls in reading evolutionary trees and ways to avoid them. *Evolution: Education and Outreach*, *12*(1). https://doi.org/10.1186/s12052-019-0112-3

Thanukos, A. (2009). A Name by Any Other Tree. *Evolution: Education and Outreach*, *2*(2), 303–309. https://doi.org/10.1007/s12052-009-0122-7

Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics (2. Aufl.)*. Wiley-Blackwell. http://gbv.eblib.com/patron/FullRecord.aspx?p=4030456

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York, NY: Psychology Press.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch measurement*. Chicago, IL: Mesa Pr.

Wu, Q., & Vos, P. (2018). Inference and Prediction. In Gudivada, V. N. & Rao, C. R. (Eds.), *Handbook of Statistics: Vol. 38. Handbook of Statistics 38: Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications* (Vol. 38, pp. 111–172). Elsevier. https://doi.org/10.1016/bs.host.2018.06.004