Research
DOI: https://doi.org/10.1128/jmbe.v20i2.1700

JMBE Journal of Microbiology & Biology Education

# Development of an Evolutionary Tree Concept Inventory

Tyler A. Kummer, Clinton J. Whipple, Seth M. Bybee, Byron J. Adams, and Jamie L. Jensen
*Department of Biology, Brigham Young University, Provo, UT 84602*

**Despite the importance of tree-thinking and evolutionary trees to biology, no appropriately developed concept inventory exists to measure student understanding of these important concepts. To address this need, we developed a multiple-choice concept inventory consisting of 24 pairs of items, and we provide evidence to support its use among undergraduate students. A set of learning outcomes was developed to guide the creation of the concept inventory. The learning outcomes, student interviews, and student responses were used to develop and revise inventory items. Supporting evidence was gathered from traditional item analysis, exploratory factor analysis, confirmatory factor analysis, traditional reliability analyses, and comparisons to alternative assessments. Appropriate implementation and utility of the concept inventory are discussed.**

## INTRODUCTION

Assessment of student understanding is a critical tool for educators as they try to help students learn the important content of a given class or course (1). Educators and researchers are often required to develop their own assessments for a myriad of reasons (e.g., unique subject area or appropriate level of rigor). These unique assessments pose a problem for researchers when we attempt to evaluate the effectiveness of an intervention and put it in context with other studies. Without a widely used and accepted assessment, it becomes difficult to appropriately compare the results of one study with those of another.

To address these problems, educational communities have developed concept inventories. Concept inventories are typically multiple-choice assessments that focus on important concepts relating to a subject area. This focus on conceptual understanding allows the assessment items to prioritize assessing understanding and applying concepts over knowledge level information that can be memorized (2, 3). A concept-focused assessment that has been properly developed provides educators with confidence in what the assessment reveals about their students' understanding and provides researchers with a more objective and meaningful form of evidence as we study student learning of the given subject.

Evolutionary trees are crucial in modern biology (4). Biologists utilize evolutionary trees to study biological phenomena ranging from genes to biogeography (5). Evolutionary tree concepts have been dubbed *tree-thinking* by researchers who advocate the importance of these concepts (6). In the last decade, significant research has been conducted on the learning of tree-thinking. Researchers have provided great insight into how students think about evolutionary trees, common misconceptions students exhibit, and how best to teach evolutionary trees to students (e.g., 7–10)

While important and interesting insights have resulted from research on tree-thinking, the research into this area is inhibited by the lack of a concept inventory for tree-thinking that has been published and made available to researchers (11). Without a widely available concept inventory, researchers are left to create their own assessments or modify existing assessments that were not generated or published for use as a concept inventory. While a few published assessments do exist, they do not meet the characteristics and standards of a concept inventory because they lack proper development/evidence or they cover only one component of tree-thinking (e.g., evolutionary relatedness) (5, 12, 13). The Tree-thinking Concept Inventory is the most promising assessment that has been published, but it has since had some of its evidence of validity (inaccurate use of terms and a lack of factor analysis) called into question (11, 14). Due to the clear need for a concept inventory on tree-thinking, we developed a new concept inventory targeted at undergraduate students, and we offer evidence of validity and

*Corresponding author. Mailing address: Department of Biology, 4102 LSB, Brigham Young University, Provo, UT 84602.
Phone: 801-822-1975. E-mail: tkummer01@gmail.com.
Received: 4 December 2018, Accepted: 14 June 2019, Published: 30 August 2019.

reliability that gives researchers and educators confidence in the measurements it provides.

## METHODS

### Subjects

A total of 1,069 undergraduate students were recruited as subjects for this study. Participants were enrolled at a highly selective private religious institution. One hundred seventeen were used in our initial phase, where we sought to learn about student understanding of evolutionary tree concepts. An additional 421 subjects completed preliminary versions of the concept inventory over three rounds of revision (A, B, and C) but did not complete the final version, and no data from their responses appear in this text. Students who participated in these early stage of development were offered extra credit for their participation. Data from a total of 531 subjects from a variety of life science courses were used in this study, including students who have declared a life science as their major (majors) and those who have not declared a major or declared a major outside of life science (nonmajors). Participants who were declared life science majors were recruited from an introductory biology course (typically consisting of freshman), a plant diversity course (typically consisting of sophomores), and an evolution course (typically consisting of seniors). Nonmajor participants were recruited from a general education biology course and accounted for 54% of participants. We used three subsets of subjects from the total of 531 to conduct different analyses. The Final Group consisted of all 531 subjects, who completed the final version of the concept inventory. The Convergent–Discrimination Group consisted of 124 nonmajor subjects who completed the final version of the concept inventory as well as two other assessments. The Test–Retest Group consists of 120 nonmajor subjects who completed the final version of the concept inventory twice in a six-week period. Students were assigned to groups based on enrollment in a course section participating in the study. Each student enrolled in a participating section completed the concept inventory as part of the course. We sought and obtained permission from the students to use their responses in this study. The IRB at Brigham Young University (BYU) approved the study, and subject consent was obtained for their participation.

### Content validity

**Student understanding.** We began the concept inventory design process by trying to learn more about how students think and reason with evolutionary trees. In order to do this, we administered a set of multiple-choice items and free-response items to 117 subjects in several biology courses. We reviewed student responses and coded them as being correct or as demonstrating one of the common misconceptions described in the literature (8, 9, 15–17). This process helped us become familiar with patterns of student thinking, including concepts they understood well and concepts they did not. We also met with eight students who were declared life science majors to hold discussions about evolutionary trees and the items they had responded to. We asked the students to review the items and their answers and tell us about their thinking. Evaluating their responses and discussing the items with them aided us in creating new items and distractors that would target key concepts while having them worded in ways that were compatible with student thinking. For further discussion of our research on student understanding please see our published work on this topic (18).

**Learning outcomes.** We developed an initial list of learning outcomes (Los) with the focus on what would be appropriate for undergraduate students of biology to understand about evolutionary tree concepts. Determining a set of learning outcomes is critical to the development of a concept inventory (19). It allows us to develop items that directly address key concepts of tree-thinking and it ensures that each outcome is covered in the concept inventory. A four-person panel of experts (three evolutionary biologists and a biology education researcher) reviewed the initial list. Each expert participated in developing and revising learning outcomes. The tree-thinking learning outcomes we developed for this concept inventory are comparable to the learning outcomes defined by others (7, 20), see Table 1.

**Item development.** We used student responses and the learning outcomes described previously to develop multiple-choice items. We developed at least two items to address each learning outcome in the hopes of providing robust coverage of the outcomes. Each item was comprised of two questions. The first question was directly related to the content the item was designed to assess, and the second question addressed the reasoning used to answer the content question. This method is patterned after Lawson's Classroom Test of Scientific Reasoning (LCTSR) (21). Using paired questions is beneficial in two ways. First, it reduces the impact of guessing by requiring the subject to answer both questions correctly to receive credit for a correct response to the item. Second, it helps us better identify misconceptions that a student might be using when they answer an item. We found in our discussion of student responses that students could often answer a content question correctly but for an incorrect reason. The paired questions allowed us to account for this and more accurately differentiate students with accurate understanding from those with incorrect understanding.

Student responses and discussions were valuable in developing questions that were appropriately worded for students while still targeting our learning outcomes. Student responses also served as the primary source for the wording of distractors that were appropriate for the question but also represented common misconceptions (2).

TABLE 1.
Identified learning outcomes and the hypothesized constructs.

| Learning Outcomes |
|---|
| **1 – Accurately interpret information depicted in an evolutionary tree using an understanding of common ancestry** |
| a. Distinguish monophyletic, paraphyletic and polyphyletic groups |
| b. Compare evolutionary relationships between taxa |
| c. Identify what the various components of an evolutionary tree represent |
| d. Distinguish between evolutionary trees with differing ordering of the species and evolutionary trees depicting differing evolutionary relationships |
| **2 – Demonstrate an understanding of how characters are inherited from common ancestors by accurately interpreting an evolutionary tree** |
| a. Identify cases of homology and analogy when interpreting an evolutionary tree |
| b. Analyze character information and evolutionary trees using parsimony |
| c. Identify synapomorphies for a group on a given evolutionary tree |
| d. Identify character states as derived or ancestral on a given evolutionary tree |
| e. Use an evolutionary tree to identify characters a given taxon would exhibit |
| **3 – Demonstrate an understanding of evolution as a continuing and non-teleological process** |
| a. Identify why using simplicity and complexity to categorize organisms as primitive and advanced species is inappropriate from an evolutionary perspective |
| b. Demonstrate an understanding that all extant populations continue to evolve and have evolved throughout their entire existence |

Previous research has shown that some students, while capable of accurately interpreting evolutionary trees using abstract or unknown taxa, are incapable of correctly interpreting phylogenies of known taxa (22). This is likely due to a common misconception, defined previously as Similarity Equals Relatedness. This misconception results when students rely on a similarity of features to determine relatedness rather than what is depicted in the evolutionary tree. When the taxa are abstract or unknown to the student they cannot rely on similarity to interpret the evolutionary tree. We developed items that used both abstract taxa and well-known taxa with this finding in mind. We believe this will allow the concept inventory to distinguish between students with no understanding of how to interpret evolutionary trees, those who can only do so with abstract or unknown taxa, and those who can regardless of the taxa used.

We used three rounds of revision to refine our items. A total of 421 major and nonmajor students completed preliminary versions, 79 students in Round A, 196 students in Round B, and 146 students in Round C. In Round A, 79 students were asked to answer and review the questions, and then comment on anything that seemed out of place or confusing about the questions. We selected 20 students to interview and asked them to describe their thinking about the items and give us feedback after Round A. After revisions, we administered the next version to 196 students in Round B, followed by a group discussion with six students asking them to discuss the items and provide feedback on any aspect that may have been confusing. The concept inventory was administered to 146 students in Round C. We, along with instructors (an evolutionary biologist and a biology education researcher) at two other institutions, reviewed the 26 two-part multiple-choice items following this final piloting of our concept inventory. Only minor changes were made, and we felt the instrument was ready for empirical analyses. The items that target each learning outcome are shown in Table 2.

**Item analysis.** Subjects from the Final Group were assessed using the Evolutionary Tree Concept Inventory (ETCI). We used item difficulty and item discrimination to evaluate each item. Item Difficulty was determined by calculating the proportion of students who correctly answered each item. Item Discrimination was evaluated in two ways. First, we calculated discrimination by taking the top scoring 27% of subjects and comparing the number of correct responses with the number of correct responses in the lowest scoring 27% of subjects (23). Next, we calculated a point-biserial correlation for each item to the total score. We used these three values to evaluate each item and to decide whether to include it on the ETCI. Based on poor item performance in terms of difficulty and discrimination, we removed two items (R1 and R2) from the ETCI (high difficulty and/or poor discrimination). R1 and R2 had performed poorly in preliminary versions of the ETCI. We had hoped the results would improve as we refined the instrument and added to the sample size, but this was not the case.

**Convergent and discriminant validity.** Convergence is the degree to which scores on two assessments that purport to measure the same construct correlate with one another. Discrimination, in contrast, is the degree to

TABLE 2.
The items intended to address each learning outcome.

| Learning Objective | 1a | 1b | 1c | 1d | 2a | 2b | 2c | 2d | 2e | 3a | 3b |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item # | 1 | 6 | 2 | 21 | 3 | 5 | 8 | 7 | 9 | 11 | 12 |
| | 20 | 15 | 10 | 22 | 4 | R1 | R2 | 13 | 14 | 19 | 17 |
| | | 16 | | | | 23 | | | | | |
| | | 18 | | | | 24 | | | | | |

R1 and R2 were items removed from the final version of the concept inventory as a result of item analysis.

which scores on two assessments, which claim to measure two differing constructs, correlate to one another. We used two assessments developed for a nonmajors introductory biology course to provide evidence of convergence and discrimination by comparing the student scores from the Convergent–Discriminant Group. The convergent assessment was designed to assess tree-thinking (TT) and consisted of 11 multiple-choice items. These items were selected from Tree Thinking Quiz I and II (5). The discriminant assessment was used to assess the nature of science (NOS) and consisted of 12 multiple-choice items written and used by the authors for course exams. We then used a Pearson and Filon's *z* test to compare the correlations.

**Factor analyses.** Factor analyses can be used to explore the underlying structure of the concepts being measured by an assessment and which items appear to be measuring the same concepts by analyzing the variation of student responses to correlated items. This allows for potential factors to be identified by grouping items together in a way that explains the variance seen in the data. These factors should represent the concepts being assessed. We used two types of factor analysis to evaluate whether items we intended to measure the same concept did in fact group together and whether our proposed grouping of the concepts aligned with the overall grouping of the items. We hypothesized that the learning outcomes created fit into three distinct categories as outlined in Table 1: accurately interpret information depicted in an evolutionary tree using an understanding of common ancestry (LO 1a–1d), demonstrate an understanding of how characters are inherited from common ancestors by accurately interpreting an evolutionary tree (LO 2a–2e), and demonstrate an understanding of evolution as a continuing and non-teleological process (LO 3a–3b). To test this initial hypothesis, we used student responses from the Final Group to conduct an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA). The EFA allowed us to generate a statistically supported hypothesis about the relationships of the items and the underlying structure of evolutionary tree concepts, while the CFA was used to test how well the hypothesis fit the data (24). We randomly split the subjects from the Final Group with half of the data being used for the EFA and half being used for the CFA. The CFA included generating a modification index to see whether any theoretically sound improvements to the model were justified. We also calculated multiple fit indices that are robust to differing data patterns to better evaluate the fit of the model.

**Shortened version.** In our process of evaluating the ETCI, we recognized its length would be a potential issue that reduces its utility to instructors. The length allowed us to have multiple items for all but one learning outcome; however, instructors may not wish to have an exam entirely devoted to tree-thinking. To compensate for the length, we selected 10 items that covered most of the learning objectives and had excellent item characteristics (see Table 3). We then ran a Pearson product–moment correlation to examine the relationship of the total score for the 10 items to the ETCI as a whole.

### Reliability

We used responses from the Final Group and Test–Retest Group to gather two forms of evidence of reliability. We measured the internal consistency of student responses by calculating a Cronbach's alpha coefficient for the Final Group. The second method we used was test–retest. Test–retest allows us to estimate the stability of the scores over time. We assessed 120 subjects using the ETCI, and then six weeks later, we assessed them with the ETCI a second time. We then calculated a Pearson's product–moment correlation between the two scores for each student. Using these two methods allowed us to produce multiple forms of reliability evidence for the results reported in this study.

## RESULTS

### Validity

**Item analysis.** We calculated item difficulty, item discrimination, and a point-biserial correlation for each item with the total score; results for each item are shown in Table 4. We used widely accepted standards to evaluate the values produced in the item analysis (23). Average item difficulty was 0.61, ranging from 0.22 (Item 5 and Item 24) to 0.91 (Item 9). Ideally it is best for difficulty values to range from 0.3 to 0.9. Item 5, Item 9, and Item 24 were all outside of this range. Average item discrimination was 0.52, ranging from 0.17 (Item 9) to 0.68 (Item 21). The average point-biserial correlation was 0.45, ranging from 0.25 (Item 9) to 0.56 (Item 19). Item 9 and Item 14 also had low discrimination (above 0.3 is ideal), but this could be explained by the ease of the items. While traditional discrimination values were low for these two items, the point-biserial correlations (another common means of evaluating discrimination) for both were in the acceptable range (> 0.2). We decided to keep these two items as part of the ETCI because they were the only items that targeted learning outcome 3e. Item R1 was removed due to the difficulty of the item (0.12) and

the discrimination value (0.18). Item R2 was removed for having a low discrimination (0.13) despite having acceptable item difficulty (0.47).

**Convergent and discriminant validity.** We used a Pearson product–moment correlation to compare the relationship between scores from the Convergent–Discriminant Group on the TT assessment and the ETCI as a measure of convergent validity. We found a large positive correlation between the scores on the two, $r(124) = 0.616$, $p < 0.001$.

We also used a Pearson product–moment correlation to compare the relationship between scores from the Convergent–Discriminant Group on the NOS assessment and the ETCI. A smaller positive correlation was found, $r(124) = 0.362$, $p < 0.001$.

Correlation coefficients relating scores of two assessments between 0.4 and 0.7 are considered to be strongly correlated while values between 0.2 and 0.4 are considered to be weakly correlated (23). We compared the two correlations with the ETCI to determine whether the TT–ETCI correlation was significantly different from the NOS–ETCI correlation. A Pearson and Filon's z test conducted using the cocor package for R showed that the TT–ETCI correlation was significantly larger than the NOS–ETCI correlation, $p = 0.002$ (25).

TABLE 3.
The learning outcomes associated with the 10 items selected for the shortened version of the ETCI.

| Learning Objective | 1a | 1b | 1c | 1d | 2a | 2b | 2c | 2d | 2e | 3a | 3b |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item # | 1 | 15 | 2 | 21 | 4 | 24 | 8 | 7 | — | 19 | 12 |

ETCI = Evolutionary Tree Concept Inventory.

TABLE 4.
The difficulty (p), discrimination (D), and point-biserial correlation ($r_{pb}$) for each item on the final version of the ETCI.

| Item # | p | D | $r_{pb}$ |
|---|---|---|---|
| 1 | .65 | .47 | .39 |
| 2 | .40 | .65 | .50 |
| 3 | .47 | .60 | .44 |
| 4 | .45 | .44 | .32 |
| 5 | **.22** | .40 | .38 |
| 6 | .69 | .64 | .53 |
| 7 | .56 | .62 | .47 |
| 8 | .69 | .34 | .29 |
| 9 | **.91** | **.17** | .25 |
| 10 | .69 | .59 | .50 |
| 11 | .71 | .60 | .53 |
| 12 | .87 | .34 | .41 |
| 13 | .64 | .47 | .39 |
| 14 | .89 | **.24** | .34 |
| 15 | .50 | .71 | .55 |
| 16 | .55 | .66 | .51 |
| 17 | .65 | .59 | .46 |
| 18 | .77 | .59 | .54 |
| 19 | .72 | .64 | .56 |
| 20 | .67 | .54 | .47 |
| 21 | .64 | .68 | .55 |
| 22 | .72 | .57 | .52 |
| 23 | .38 | .55 | .43 |
| 24 | **.22** | .41 | .41 |

ETCI = Evolutionary Tree Concept Inventory.
Bolding indicates values of concern. Items 5, 9, and 24 are all outside the ideal difficulty range of 0.3 to 0.9. Items 9 and 14 had low discrimination (above 0.3 is ideal).

**Factor analyses.** We conducted a principal axis factor analysis (PAF) of the 24-item ETCI on the responses of 265 randomly sampled subjects from the Final Group. This random sampling allowed for a confirmatory factor analysis (CFA) on responses of the 266 remaining subjects. The Kaiser-Meyer-Olkin (KMO) value was 0.84, indicating that, overall, the data were likely to be factorable. In addition, all individual item KMO measures were greater than 0.76, indicating that each item was suitable to be included. A Bartlett's Test of Sphericity was statistically significant ($p < 0.001$), which also indicates the data are likely to be factorable.

We used a scree plot and the total variance explained to help evaluate the number of factors that were appropriate to extract. Visual inspection of the scree plot indicated two potential inflection points at either three or five factors. Total variance explained exceeded 5% in five factors. These five factors cumulatively explained 45.3% of the total variance as opposed to only 35.0% in the three-factor solution. Based on these initial results, we decided a five-factor solution was more appropriate than the three-factor solution that may have matched our initial hypothesis. To provide further evidence, we used a PAF parallel analysis on the 265-subject dataset with 1,000 randomly generated and normally distributed data sets. A parallel analysis uses randomly generated data to determine whether factors produced from actual data are larger than would be expected by chance. The results (see Fig. 1) of the parallel analysis supported our decision to extract five factors.

We used a Promax rotation on the data, and the Item loadings on the five factors are shown in Table 5. The Item factor relationships differ in a number of ways from our originally proposed three-factor grouping. Item 1 most strongly loaded on the same factor as items that asked students to compare evolutionary relatedness, when it theoretically should have loaded on Factor 5, with Item 20, which measured the same construct (clade type). Item 10 also loaded most strongly on

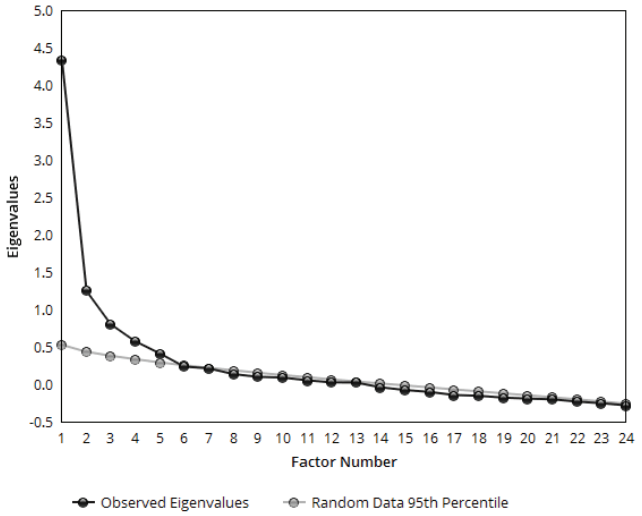


FIGURE 1. Scree plot of the observed eigenvalues and the 95th percentile of eigenvalues generated by random data

TABLE 5.
Largest factor loadings on the five extracted factors for each item and their corresponding learning outcome.

| Item # | LO | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | 1a | **.256*** | | | | **.021*** |
| 20 | 1a | **.053*** | | | | **.205*** |
| 6 | 1b | .713 | | | | |
| 15 | 1b | .594 | | | | |
| 16 | 1b | .803 | | | | |
| 18 | 1b | .639 | | | | |
| 2 | 1c | **.011*** | | | | .393 |
| 10 | 1c | **.229*** | | | | **.016*** |
| 21 | 1d | | | | .664 | |
| 22 | 1d | | | | .755 | |
| 3 | 2a | | | | | .415 |
| 4 | 2a | | | | | .292* |
| 5 | 2b | | | | | .471 |
| 23 | 2b | | | | | .356 |
| 24 | 2b | | | | | .429 |
| 7 | 2d | | | .485 | | |
| 13 | 2d | | | .415 | | |
| 8 | 2c | | | .173* | | |
| 9 | 2e | | | .227* | | |
| 14 | 2e | | | .552 | | |
| 11 | 3a | | .917 | | | |
| 19 | 3a | | .758 | | | |
| 12 | 3b | | .382 | | | |
| 17 | 3b | | .382 | | | |

Bolding indicates items that share the same learning outcome loading on separate factors.
* Indicates weak factor loadings.

the factor with evolutionary relatedness items rather than with Item 2 on Factor 5. While the factor loading of Items 1 and 10 differed from what we expected, their loading was relatively weak on either factor, meaning the evidence that they measure the same factor as the other items on Factor 1 is questionable. In addition to Items 1 and 10 Items 8, 9, 20, and to a lesser extent 4 all had low factor loadings (< 0.3). We also looked at the correlation between factors 1 and 5 because our weak loading items were on these factors. We found that these two factors had a correlation of 0.551.

We conducted a CFA based on the five factors and item loading patterns seen in the EFA using the lavaan package in R (26). The CFA allowed us to compare the fit of the five-factor model to the second half of the data we excluded from the EFA. After performing the CFA, we also computed a modification index that showed how the model might be altered to improve fit. Based on our theoretical reasoning, we adopted two of these suggestions that allowed for two covariance terms in the model: one between Items 6 and 16 and the other between Items 11 and 19. We accepted these suggestions because they significantly improved the model and each set of items loaded on the same factor, respectively. We used three fit indices to evaluate the fit of our model. Two indices indicated that our model was an acceptable fit to the data (Root Mean Square Error of Approximation [RMSEA] = 0.034 and Standardized Root Mean Square Residual [SRMR] = 0.043). The third index we used was the incremental fit index (IFI), which had a reported value of 0.94. This falls just below the conservative threshold of 0.95 and above the threshold of 0.90 that some recommend (27).

**Shortened version.** To evaluate the shortened 10-item version of the ETCI as a predictor of the full version, we used a Pearson product–moment correlation and treated the 10 items as a separate assessment and compared it with the ETCI as a whole. We found the scores of these 10 items to be strongly correlated with the scores of the ETCI, $r(531) = 0.918$, $p < 0.001$.

### Reliability

We calculated a Cronbach's alpha coefficient as a measure of internal reliability for the Final Group (531 subjects) who took the ETCI. The ETCI was shown to have a Cronbach's alpha of 0.845. We used a Pearson's product–moment correlation to assess the relationship between subject scores on the first test attempt and second attempt of our test–retest group of subjects. We found a large positive correlation between the two scores, which is to be expected, $r(120) = 0.828$, $p < 0.001$.

## DISCUSSION

We developed, reviewed, and revised items to directly address learning objectives that were developed for

undergraduate students. Student interviews, student open-answer responses, and literature defining common misconceptions were used to guide the development of each item. We conclude that the results of our analyses demonstrate that the process of item development produced appropriate items that measure what we intended them to measure and distinguish between students of differing ability.

The EFA results showed that our initial three-factor model, which was based on our theoretical grouping of the learning objectives, was not justified based on the pattern of student responses. We used the results of the EFA to propose a new five-factor model as seen in Figure 2. The five-factor solution differed from our proposed three-factor model in a number of ways but most importantly in the grouping of items targeting LO 1a and 1c (clade type and evolutionary tree components) with items targeting LO 2a and 2b (homology/analogy and using parsimony). All other factors consisted of a subset of items that fell in the same theoretical category in which they were initially placed in (e.g., Factor three consists of items targeting 2c, 2d, and 2e but none from groups one or three). The results of the CFA show this new model is an acceptable fit to the data. We created a new classification of our learning outcomes based on the results of the factor analyses (Table 6). We believe this new classification reflects sound theoretical groupings and is consistent with the underlying construct structure supported by the factor analyses.

While this new classification and the model used to create it were a good fit to our data, we did have a number of items that only weakly loaded on a factor. We believe the items that weakly loaded may have differed in the cognitive task students were asked to use to answer the question. Because of this, the factor analysis may be picking up on the shared cognitive task with items measuring other concepts. For example, Item 1 asks students to analyze character evidence to determine whether a group is monophyletic, while Item 20 asks students to identify how many monophyletic groups are in a given tree. Item 20 ended up weakly loading on Factor 1 (see Table 5), which had several items that required analyzing given evolutionary trees. We went to the EFA results to look at the item correlations between Item 1 and Item 20 and found a correlation value of -0.076. While these two items clearly both required students to use their understanding of monophyly, the scores for these two items were not correlated. The difference between analyzing and identifying might have caused the weak and opposite loadings we saw in the EFA results. The correlation between Factors 1 and 5 may have also played a role in the loading of Items 1, 2, 10, and 20. We saw a fairly strong correlation between these factors, and the loading of these four items was strongest (though weak overall) on Factors 1 and 5. Because these were weak results, we relied primarily on our theoretical understanding when building a new model to test for the CFA.

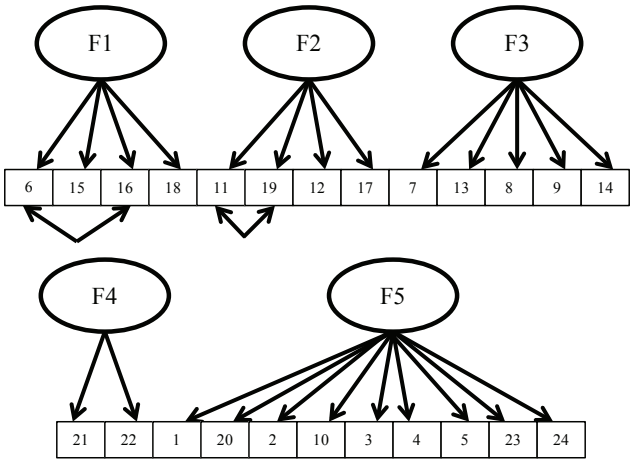


FIGURE 2. Five-factor model analyzed for fit in the CFA. Arrows between factors and items indicate loading. Arrows between two items indicate covariance. CFA = confirmatory factor analysis.

An exploratory factor analysis of the Force Concept Inventory (FCI) also showed several weak loading items (6 out of 26 items compared with 6 out of 24 items in this analysis). These researchers concluded that while their analysis produced novel and interesting results, further research was needed to understand the conceptual framework measured by the FCI (28). The weak loading of some items and the unexpected groupings we saw in our results likewise make us believe the conceptual framework of tree-thinking is in need of further research. While our new model is theoretically sound, further evidence is needed before we can conclusively say that it should be favored over our original model.

The significant difference found between the correlation coefficients (Tree-thinking–ETCI vs. Nature of Science–ETCI) and the differing classification of the correlation coefficients (strong and weak) serve as evidence that the ETCI measured the constructs we intended. While the correlation between the TT–ETCI was significantly higher, the correlation between NOS–ETCI was still significant. We believe that the higher-order cognitive skills required to answer the majority of items on both assessments can explain this significance. A previously published assessment that focused on evolutionary relatedness found a significant correlation with scientific reasoning (12). Scientific reasoning has been found to be highly correlated with performance on assessment items that require higher-order cognitive skill (29). We believe the significant correlation found between the NOS and the ETCI is likely due to both requiring higher-order cognitive skills. Students with higher scientific reasoning ability performed better on both assessments, leading to a significant correlation that was not due to similar constructs being assessed.

Our estimate of internal reliability produced a Cronbach's alpha coefficient well within the range of values expected for a concept inventory, strongly suggesting that student responses to the EFA were reliable (30). Additionally, the significant correlation found during the test–retest analysis also yields a strong estimate of reliability, providing

TABLE 6.
Learning outcomes aligned to the five-factor solution.

| | Learning Outcomes | Original |
|---|---|---|
| **1 –** | **Compare evolutionary relationships between taxa** | 1b |
| **2 –** | **Distinguish between evolutionary trees with differing ordering of the species and evolutionary trees depicting differing evolutionary relationships** | 1d |
| **3 –** | **Use an understanding of the theoretical aspects of evolutionary trees to evaluate group and character evolution based on common ancestry and parsimony** | N/A |
| a. | Identify cases of homology and analogy when interpreting an evolutionary tree | 2a |
| b. | Analyze character information and evolutionary trees using parsimony | 2b |
| c. | Distinguish monophyletic, paraphyletic and polyphyletic groups | 1a |
| d. | Identify what the various components of an evolutionary tree represent | 1c |
| **4 –** | **Demonstrate an understanding of how characters are inherited from common ancestors by accurately interpreting an evolutionary tree with characters** | 2 |
| a. | Identify synapomorphies for a group on a given evolutionary tree | 2c |
| b. | Identify character states as derived or ancestral on a given evolutionary tree | 2d |
| c. | Use an evolutionary tree to identify characters a given taxon would exhibit | 2e |
| **5 –** | **Demonstrate an understanding of evolution as a continuing and non-teleological process** | 3 |
| a. | Identify why using simplicity and complexity to categorize organisms as primitive and advanced species is inappropriate from an evolutionary perspective | 3a |
| b. | Demonstrate an understanding that all extant populations continue to evolve and have evolved throughout their entire existence | 3b |

The final column indicates their alignment in our original learning outcomes from Table 1.

evidence of stability over time. Given the amount of time between the two attempts, our correlation falls well above the acceptable cutoff ($r > 0.7$) (30). Thus, our results provide compelling evidence for the reliability of responses to the ETCI.

One use of evolutionary trees that was not covered by our concept inventory and learning outcomes was that of depicting the evolution of genes (17). While scientists commonly use gene trees in their research, these types of trees are rarely included in an introductory study of biology. We believe it would have been beyond the scope of this assessment to include gene trees and related concepts in our learning outcomes and concept inventory.

While we believe we have demonstrated that the ETCI is an adequate measure of tree-thinking, we recognize that its focus is on conceptual understanding and that it does not ask students to necessarily complete tasks that would be more authentic to how practicing scientists use evolutionary trees. We believe we appropriately focused on conceptual understanding given the goals and intended use of the ETCI, but it does not represent the entirety of ways in which an instructor may want to assess evolutionary tree concepts.

As we previously mentioned, the length of the ETCI is likely to be of concern for those who wish to use it in academic settings. The large correlation found between the shortened version and the full version indicates that the 10 items selected serve as a good predictor of student scores on the full version. The correlation between the shortened version and full version is higher than for a similarly shortened version of the Meiosis Concept Inventory and its full version (31). The evidence of reliability and validity of student responses to the ETCI outlined in this research only apply to the full version. Due to this, we would not recommend using the 10-item version for research purposes, but it may be useful to instructors as a pre-assessment, as a quiz, or as part of a unit assessment.

The ETCI has the potential to help researchers and instructors as a concept inventory. Researchers can use the ETCI in multiple ways. First, it can be used to better understand how tree-thinking concepts are related to each other. As we have shown, our own theoretical understanding differed from the pattern shown in our results. As we better understand the relationship tree-thinking concepts have to one another, we can design instruction to account for these patterns. For example one might traditionally teach about monophyly and paraphyly during a lesson that also covers evolutionary relatedness due to perceived theoretical connectivity. However, if, as our results indicate, these concepts are more closely tied to an understanding of homology and analogy, it may be better to include these concepts when teaching about homology and analogy. The ETCI can also be used by researchers to measure student understanding of tree-thinking. Doing this would allow researchers to make better comparisons between their own research and the research of others. Instructors, of course, can also use the ETCI as an assessment to determine how effective their

tree-thinking–related instruction has been in teaching tree-thinking concepts. Instructors can also use the ETCI as a formative assessment that would allow them to see what misconceptions a particular student holds or those most commonly held by their students. We used our work in this study as well as results from previous research to develop the alternative answer options in the items of the ETCI. Looking at the answers selected by an individual student or the class as a whole would allow an instructor to identify which misconceptions they might want to specifically address in future instruction. We invite any readers interested in using the ETCI for academic or research purposes to contact us for either the full or shortened version.

## ACKNOWLEDGMENTS

This project was supported by a National Science Foundation Grant (IOS-1253241) awarded to CJW. Any findings, opinions, recommendations, or conclusions expressed in this work are those of the authors and do not necessarily reflect the views of the NSF. We thank the participants and instructors who participated in this project, allowing us to complete this important work. The authors declare that there are no conflicts of interest.

## REFERENCES

1. Miller MD, Linn RL, Grunland NE. 2013. Measurement and assessment in teaching, 11th ed. Pearson, Boston.
2. Garvin-Doxas K, Klymkowsky M, Elrod S. 2007. Building, using, and maximizing the impact of concept inventories in the biological sciences: report on a National Science Foundation–sponsored Conference on the Construction of Concept Inventories in the Biological Sciences. CBE Life Sci Educ 6:277–282.
3. Krathwohl DR. 2002. A revision of Bloom's taxonomy: an overview. Theory Pract 41:212–218.
4. Thanukos A. 2010. Evolutionary trees from the tabloids and beyond. Evol Educ Outreach 3:563–572.
5. Baum DA, Smith SD, Donovan SSS. 2005. The tree-thinking challenge. Science 310:979–980.
6. O'Hara RJ. 1997. Population thinking and tree thinking in systematics. Zool Scr 26:323–329.
7. Eddy SL, Crowe AJ, Wenderoth MP, Freeman S. 2013. How should we teach tree-thinking? An experimental test of two hypotheses. Evol Educ Outreach 6:13.
8. Gregory TR. 2008. Understanding evolutionary trees. Evol Educ Outreach 1:121–137.
9. Halverson KL, Pires CJ, Abell SK. 2011. Exploring the complexity of tree thinking expertise in an undergraduate systematics course. Sci Educ 95:794–823.
10. Meir E, Perry J, Herron JC, Kingsolver J. 2007. College students' misconceptions about evolutionary trees. Am Biol Teach 69:e71–e76.
11. Walter EM, Halverson KM, Boyce CJ. 2013. Investigating the relationship between college students' acceptance of evolution and tree-thinking understanding. Evol Educ Outreach 6:26.
12. Blacquiere LD, Hoese WJ. 2016. A valid assessment of students' skill in determining relationships on evolutionary trees. Evol Educ Outreach 9:5.
13. Smith JJ, Cheruvelil KS, Auvenshine S. 2013. Assessment of student learning associated with tree-thinking in an undergraduate introductory organismal biology course. CBE Life Sci Educ 12:542–552.
14. Neagle E. 2009. Patterns of thinking about phylogenetic trees: a study of student learning and the potential of tree-thinking to improve comprehension of biological concepts. Idaho State University, Pocatello, ID.
15. Baum DA, Smith SD. 2013. Tree thinking: an introduction to phylogenetic biology. Roberts and Company Publishers, Greenwood Village, CO.
16. Meisel RP. 2010. Teaching tree-thinking to undergraduate biology students. Evolution 3:621–628.
17. Omland KE, Cook LG, Crisp MD. 2008. Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. BioEssays 30:854–867.
18. Kummer TA, Whipple CJ, Jensen JL. 2016. Prevalence and persistence of misconceptions in tree thinking. J Microbiol Biol Educ 17:389–398.
19. Wiggins GP, McTighe J. 2005. Understanding by design. ASCD, Alexandra, VA.
20. Novick LR, Catley KM. 2012. Reasoning about evolution's grand patterns: college students' understanding of the tree of life. Am Educ Res J 50(1):138–177.
21. Lawson AE. 1978. The development and validation of a classroom test of formal reasoning. J Res Sci Teach 15:11–24.
22. Novick LR, Catley KM, Funk DJ. 2011. Inference is bliss: using evolutionary relationship to guide categorical inferences. Cogn Sci 35:712–743.
23. Doran RL. 1980. Basic measurement and evaluation of science instruction. National Science Teachers Association, Washington, DC.
24. Matsunaga M. 2010. How to factor-analyze your data right: do's, don'ts, and how-to's. Int J Psychol Res 3:97–110.
25. Diedenhofen B, Musch J. 2015. Cocor: a comprehensive solution for the statistical comparison of correlations. PLOS One 10(4):e0121945.
26. Rosseel Y. 2012. lavaan: an R package for structural equation modeling. J Stat Softw 48(2).
27. Hu L, Bentler PM. 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model Multidiscip J 6:1–55.
28. Scott TF, Schumayer D, Gray AR. 2012. Exploratory factor analysis of a Force Concept Inventory data set. Phys Rev Spec Top Phys Educ Res 8(2):020105.
29. Lawson AE, Alkhoury S, Benford R, Clark BR, Falconer KA. 2000. What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. J Res Sci Teach 37:996–1018.
30. Kline P. 2000. The handbook of psychological testing. Routledge, London and New York.
31. Kalas P, O'Neill A, Pollock C, Birol G. 2013. Development of a meiosis concept inventory. CBE Life Sci Educ 12:655–664.