JOURNAL OF RESEARCH IN SCIENCE TEACHING

Research Article

Visualizing Biological Data in Museums: Visitor Learning With an Interactive Tree of Life Exhibit

Michael S. Horn,[1] Brenda C. Phillips,[2,3] Evelyn Margaret Evans,[4] Florian Block,[5,3] Judy Diamond,[6] and Chia Shen[3]

[1]*Learning Sciences and Computer Science, Northwestern University*
[2]*Department of Psychological and Brain Sciences, Boston University*
[3]*School of Engineering and Applied Sciences, Harvard University*
[4]*Center for Human Growth and Development, University of Michigan, Ann Arbor*
[5]*Department of Theatre, Film, and Television, University of York, United Kingdom*
[6]*University of Nebraska State Museum, University of Nebraska, Lincoln*

Received 9 January 2015; Accepted 1 February 2016

Abstract: In this study, we investigate museum visitor learning and engagement at an interactive visualization of an evolutionary tree of life consisting of over 70,000 species. The study was conducted at two natural history museums where visitors collaboratively explored the tree of life using direct touch gestures on a multi-touch tabletop display. In the study, 247 youth, aged 8–15 years, were randomly assigned in pairs to one of four conditions. In two of the conditions, pairs of youth interacted with different versions of the tree of life tabletop exhibit for a fixed duration of 10 minutes. In a third condition, pairs watched a 10 minute video on a similar topic. Individual responses on a 53-item exit interview were then compared to responses from a fourth, baseline condition. Contrasting with the baseline condition, visitors who interacted with the tabletop exhibits were significantly more likely to reason correctly about core evolutionary concepts, particularly common descent and shared ancestry. They were also more likely to correctly interpret phylogenetic tree diagrams. To investigate the factors influencing these learning outcomes, we used linear mixed models to analyze measures of dyads' verbal engagement and physical interaction with the exhibit. These models indicated that, while our verbal and physical measures were related, they accounted for significant portions of the variance on their own, independent of youth age, prior knowledge, and parental background. Our results provide evidence that multi-touch interactive exhibits that enable visitors to explore large scientific datasets can provide engaging and effective learning opportunities. © 2016 Wiley Periodicals, Inc. J Res Sci Teach
Keywords: interactive tabletops; informal science learning; museums; evolution; information visualization

The nature of scientific research has undergone a profound shift in recent decades. More than ever, scientists pursue lines of research that rely on massive data sets and computational methods of inquiry (Foster, 2006). As an example relevant to this paper, researchers around the globe are

Contract grant sponsor: National Science Foundation; Contract grant number: DRL-1010889.
Correspondence to: Michael S. Horn; E-mail: michael-horn@northwestern.edu
DOI 10.1002/tea.21318
Published online in Wiley Online Library (wileyonlinelibrary.com).

engaged in an ambitious effort to assemble the evolutionary relationships of millions of species into a unified tree of life (Cracraft & Donoghue, 2004). These transformational changes in the nature of scientific inquiry raise important questions about the nature of learning experiences provided by natural history museums, science centers, and other informal science institutions. Specifically, can computational tools and large datasets be used to create unique and meaningful learning experiences for visitors? And, can brief engagements with such learning experiences lead to improved understanding of complex science concepts such as evolution?

To address these questions, we investigated visitor learning at a natural history museum exhibit designed to convey concepts of evolution and biodiversity. The exhibit presents an evolutionary tree of life consisting of over 70,000 species that visitors explore using a deep zoom interaction interface on a multi-touch tabletop display. Visitors can move from the origins of life 3.5 billion years ago to present-day species representing a diversity of life on the planet (Figure 1). To develop this exhibit we combined several large scientific datasets and created a novel visualization technique (Block et al., 2012a).

Our exhibit design was informed by research on learning in museums and other informal environments (Crowley et al., 2001; Falk & Dierking, 2000). Researchers have previously identified design factors that promote *active prolonged engagement* (APE) with interactive exhibit elements in science museums (Humphrey & Gutwill, 2005). Prototypical APE exhibits support open-ended exploration, self-driven discovery, and collaborative engagement. From this perspective, social interaction is seen as critical to learning, and effective exhibits must be intentionally designed to foster collaboration while limiting confusion and disruption.

Large interactive displays such as multi-touch tabletops are popular in museums and other public spaces (Antle, Tanenbaum, Seaborn, Bevans, & Wang, 2011; Horn et al., 2012; Hinrichs & Carpendale, 2011; Hinrichs, Schmidt, & Carpendale, 2008; Hornecker, 2008; Snibbe & Raffle, 2009). These displays have made it possible for visitors to “touch” and explore visualizations of large scientific data sets (e.g., Hinrichs et al., 2008; Louw & Crowley, 2013; Ma, Liao, Ma, & Frazier, 2012; Roberts, Lyons, Cafaro, & Eydt, 2014). Despite these opportunities, however, supporting intuitive interaction for multiple users that goes beyond superficial levels of engagement is still deceptively challenging (Block, Wigdor, Phillips, Horn, & Shen 2012b; Hinrichs & Carpendale, 2011; Hornecker, 2008). Large multi-touch displays invite simultaneous use by multiple visitors, but they also invite confusion, conflict, and interference as visitors work

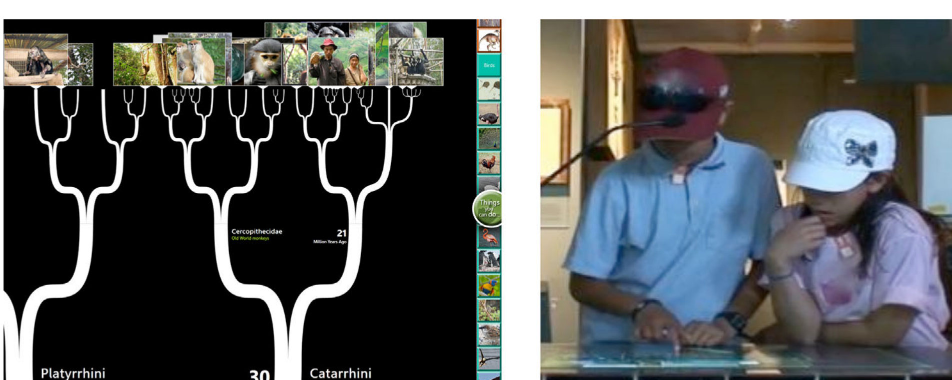

Figure 1. Screenshot from DeepTree (left). A dyad interacting with DeepTree on a multitouch tabletop display at a natural history museum (right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

at cross purposes without the guidance of established interaction techniques and social norms. And, while design frameworks have been proposed to promote social interaction and playful exploration (e.g., Block et al., 2012b; Snibbe & Raffle, 2009), few existing studies have demonstrated learning gains with tabletops or other large displays in museums especially around difficult topics like evolution. This study adds to the existing literature and makes three main contributions: First, we demonstrate through a controlled study with 247 museum visitors that people can learn complex science concepts from an interactive visualization of a large scientific dataset. Second, we contribute to an understanding of how social engagement can contribute to learning in informal environments. In particular, we use linear mixed models to understand the effect of several measures of verbal engagement and physical interaction on individual youth learning outcomes. Finally, our results provide insight into how to help learners make sense of large scientific datasets by building on their intuitive conceptual frameworks.

Evolution and the Tree of Life

Evolution is a central organizing principle for all of the life sciences. However, broad public understanding of evolutionary concepts remains elusive (Banet & Ayuso, 2003; Bishop & Anderson, 1990; Evans, 2001, 2013). Conveying concepts of evolution to museum visitors is especially challenging (Diamond & Evans, 2007; Diamond & Scotchmoor, 2006) due to short engagement times, difficulties of conveying complex, dynamic processes with static representations (Wilensky & Reisman, 2006), and widespread resistance to the idea of evolution among the general public in the United States (Gallup, 2014; Miller, Scott, & Okamoto, 2006). Even among natural history museum visitors, only about a third of adults grasp evolutionary concepts (Evans et al., 2010; MacFadden et al., 2007). Previous research has demonstrated that evolution exhibits with multiple components can be effective for learning (Spiegel et al., 2012; Tare, French, Frazier, Diamond, & Evans, 2011), especially if they are embedded in a meaningful narrative (Evans, Weiss, Lane, & Palmquist, 2015). A challenge in this project was to achieve learning through interactions with a visualization of a large scientific dataset. Museums have made use of video to convey some of the dynamic aspects of evolution (e.g. Prum, 2008). However, while video can be engaging and informative, visitors have limited control over the scope and flow of information.

Building on Visitors' Intuitive Understandings of Evolution

Given the challenges of conveying evolution in museums, one promising direction has been to offer learning experiences that appeal to or clarify visitors' intuitive or everyday reasoning. Prior research suggests that some intuitive concepts hinder understanding (Evans, Rosengren, Lane, & Price, 2012), while others may serve as a foundation for more sophisticated understandings (Evans et al., 2010; Legare, Lane, & Evans, 2013).

The idea that there can be dramatic changes in species over time runs counter to children's intuitive "essentialist" beliefs in the stability and immutability of kinds (Gelman & Rhodes, 2012; Samarapungavan & Wiers, 1997). Young children, in particular, are likely to argue that species remain unchanged over time or to use anthropomorphic (intentional) reasoning, stating that organisms *want* to change (Evans, 2000, 2001, 2013). By the middle elementary-school years, however, children often adopt *restricted* teleological explanations of species change. That is, they endorse the idea that organisms *need* to change in order to survive in a particular environment, while simultaneously rejecting anthropomorphic mechanisms of change. This developmental shift could potentially scaffold more scientifically accurate understandings by increasing children's receptivity to mechanisms of change that do not involve the intentions of individual organisms (Evans et al., 2012, 2015). Supportive evidence for this argument is found in studies conducted in museums among children, youth, and adults (Evans et al., 2010; Legare et al., 2013;

Spiegel et al., 2012). In these studies, restricted teleological reasoning (e.g., “the first fungus *needed* to be protected from the second fungus”) was positively correlated with an understanding of natural selection. On the other hand, anthropomorphic reasoning, the idea that adaptive change is intentionally caused (e.g., “[the finches] had to try and work harder, probably, to develop their beaks”), was uncorrelated or negatively correlated, depending on the measure.

Other studies have shown that the phylogenetic tree diagrams used to communicate macroevolutionary ideas can invite confusion. Phylogenetic trees are core representations in the life sciences and are used by biologists to derive lineages of species according to the characteristics that they share with a most recent common ancestor (Baum, Smith, & Donovan, 2005). Despite their importance, however, prior research has shown that tree diagrams are difficult for novices to understand even at the college level (MacDonald & Wiley, 2012; Meir, Perry, Herron, & Kingsolver, 2007; Novick & Catley, 2013; Phillips, Novick, Catley, & Funk, 2012). For example, high school and college students have more difficulty interpreting the relationships between species when their intuitive beliefs conflict with the information depicted (Novick, Catley, & Funk, 2011). Further, the results of a recent qualitative study revealed that high school students have considerable difficulty reasoning about the ancestors that humans share with other species even when these relationships are depicted diagrammatically (Seoh, Subramaniam, & Hoh, 2015).

One of the goals of this study was to help visitors make sense of phylogenetic trees by exploring relationships among diverse species. The notion of “relatedness” was therefore an important consideration for the study. Relatedness is a fundamental concept for understanding tree diagrams (Catley, Phillips, & Novick, 2013); however, it can elicit both intuitive (family relationships) and more expert (tree of life) reasoning. Our interactive tree of life frames the core idea that all living things on earth are related, with the aim of clarifying and reinforcing visitors’ intuitive concepts. Furthermore, by providing evidence of common ancestry, the tree of life counters the essentialist notion that each living kind has a unique essence, an acknowledged barrier to understanding common descent (Evans, 2000; Gelman & Rhodes, 2012; Shtulman & Schulz, 2008).

Study Overview

In this study, we were interested in the effect of our tabletop exhibit on visitor learning in natural history museums. We were also interested in how elements of verbal engagement and physical interaction at the exhibit might contribute to learning outcomes. Evidence suggests that both social and physical engagement play an important role in learning in museums (Crowley et al., 2001; Eberbach & Crowley, 2005; Falk & Dierking, 2000), but we know less about how they shape learning with computer-based exhibits, particularly those involving visualizations of large scientific data sets. To investigate these factors we recruited youth dyads, aged 8–15 years, at two natural history museums to participate in one of four conditions. In the first two conditions, dyads interacted with different versions of our tree of life exhibit on a tabletop display for a fixed period of 10 minutes. In the third condition, dyads watched a 10 minute video about evolution and the tree of life that addressed topics that were similar to the tabletop conditions. Because it was produced independently, the video used language and visual representations that were not directly comparable to the tabletop exhibits. We included the video as a condition in the study because it exemplifies media commonly used by museums to help visitors understand evolution (MacDonald & Wiley, 2012).

We subsequently administered a 53-item exit interview to each participant individually and compared the results to those of participants from a baseline control condition. We also collected video recordings of dyad conversation and computer logs of touch interaction. With this design we hoped to examine differences in visitor learning between the experimental and baseline conditions. We also hoped to understand how measures of dyads' verbal engagement and physical interaction in the tabletop conditions contributed to learning outcomes.

Learning Objectives

Our exhibit design and assessments were guided by learning objectives related to macro- and micro-level evolution concepts. The macro-level concepts reflect increasingly deeper levels of understanding of the tree of life, especially the concept that all living things are related and that evidence for these relationships is based on shared ancestral traits. Additionally, we hoped to improve visitors' ability to interpret phylogenetic tree diagrams. The micro-level concepts emphasize evolutionary processes that act on populations over time, resulting in the tree of life and including inheritance, variation, adaptation, and natural selection. Finally, we hoped to instill a sense of wonder at the complexity and diversity of life on earth. While our learning objectives and measures concern concepts of evolution, we believe that our findings have the potential to inform the design of other learning experiences involving the collaborative exploration of large scientific data sets.

Research Questions and Predictions

The current study was guided by the following two research questions. RQ1: What are the effects of exhibit condition (tabletop and video) and age on youth understanding of evolution? RQ2: How do elements of verbal engagement and physical interaction with the tabletop display contribute to learning outcomes? We predicted that compared to the baseline condition, both tabletop conditions would elicit a better understanding of evolutionary concepts. We expected that the video condition would also result in learning gains. However, we also hypothesized that youth would interact with one another less often while watching the video, and that this, coupled with the self-directed engagement of the tabletop conditions, might lead to differential learning outcomes in favor of the tabletop exhibits.

Method

Participants

In total, 251 youth participated in the study (Mean Age = 11.55 years; $SD = 1.69$). In all cases dyads were siblings or friends recruited from the same family group. Four youth were excluded from the analyses, three because they did not complete the exit interview, and one because of a recruiting error. The remaining 120 girls and 127 boys identified as 72% Caucasian, 11% Asian American, 5% Latin American, 4% African American, 4% Mexican American, 2% Indian, 2% Puerto Rican, and less than 1% Arabic and Native American. Dyads were randomly assigned to one of four conditions (see Table 1). For the purpose of analysis we defined two age groups by median split: 8–11 years ($M = 9.99$; $SD = 0.86$) and 12–15 years ($M = 12.87$; $SD = 0.90$). We selected these age groups because previous research has indicated that 8–11-year olds are beginning to grasp the concept of evolutionary relationships, while older children are exposed to these ideas in school (Evans, 2013). The mean age for each age group did not differ significantly by condition.

Previous research has demonstrated that background factors such as age, education, and religious beliefs are likely to influence visitor understanding of evolution and responsiveness to exhibits (Evans et al., 2010; Tare et al., 2011). We controlled for these factors by randomly assigning participants to condition and also by measuring them so that they could be statistically controlled if necessary. Parents ($N = 231$) completed a questionnaire covering demographic

Table 1
Site, age, and gender of participants by condition (DeepTree I, DeepTree II, video, and baseline)

Condition	DeepTree I	DeepTree II	Video	Baseline	Total
Total	59	62	63	63	247
Site					
Museum 1	29	32	31	32	124
Museum 2	30	30	32	31	123
Age					
8–11 years	28	28	29	28	113
12–15 years	31	34	34	35	134
Gender					
Female	33	31	32	31	127
Male	26	31	31	32	120

information including: parental educational level; parental views of religion and evolution, such as beliefs about evolutionary origins (from Spiegel et al., 2012); and characteristics of the youth participants, such as the child's knowledge of evolution. There were no significant differences by museum site ($ps > 0.05$) or by condition ($ps > 0.05$) for these measures, with one exception for parents of children in the video condition, which did not bear on our research questions (see Suppl. Table S1 for details of the measures by condition).

Materials

This study used two interactive tabletop applications called *DeepTree* and *FloTree*. We developed these applications through an iterative process of design and evaluation with a team of computer scientists, learning scientists, biologists, and museum curators (Block et al., 2012a).

DeepTree. DeepTree is an interactive visualization of the tree of life showing the phylogenetic relationships of 70,000 species. The design has three major components (Figures 1 and 2). The main display area allows visitors to zoom and pan through the entire tree of life. The tree uses a fractal layout algorithm so that branches emerge as the user zooms in. Unlike static depictions of trees that simplify information by limiting the number of species, the fractal design allows for the depiction of many thousands of species while reducing visual complexity. The second component is a scrolling image reel along the right side of the screen containing a subset of 200 species representing important evolutionary groups. When an image is held, the table highlights the specie's location in the tree and automatically flies toward it. The final component is a *relate* feature that allows visitors to compare any two species in the image reel. When activated, the system flies to the common ancestor of the two species. Visitors can then open a second screen that shows a simplified "training" tree depicting the time of divergence and major evolutionary landmarks for the two species (Figure 2). These landmark points can be activated to reveal further information about common ancestors and major shared traits.

FloTree. FloTree is an interactive visualization of a simulated population of organisms that changes over time in response to geographic separation and natural selection. When launched, visitors see colorful dots representing organisms that emerge from the bottom of the screen and repeatedly "produce" new lines of dots that steadily grow upward. Visitors can place their hands on the table to introduce virtual environmental barriers that split the population of dots into subgroups (Figure 3). If the hands remain in place long enough, the color patterns diverge into two new populations with distinctive characteristics ("species"). After each simulation run, the pattern

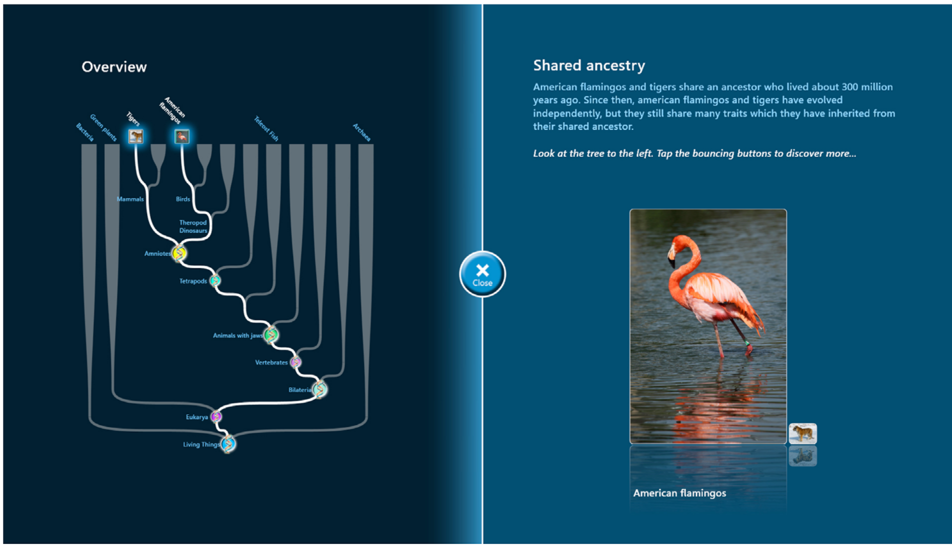

Figure 2. Screen shot of the training tree from the DeepTree exhibit. The left side of the screen shows a simplified "training tree" with important evolutionary landmarks highlighted. The right side of the screen shows information about the two species selected and their time of divergence. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

of diverging dots merges into solid branches of a tree. Expandable information bubbles explain the visualization in accessible language.

Video. We also included a third condition in which participants watched a video, *Discovering the Great Tree of Life*. This video was produced by the Peabody Museum of Natural History (Prum, 2008) and was chosen for its high production quality and the evolutionary topics it covered. Video exhibits are also common in natural history museums and are used to explain core evolutionary concepts. The video addressed all of our content-related learning objectives and featured animations, voiceovers, and interviews with prominent evolutionary biologists. The video also included a dynamic visualization of a tree of life and a segment visualizing how changes in a population of organisms (i.e., rabbits) can result in speciation. While it was not our primary objective to compare the video and the tabletop conditions directly because they differed

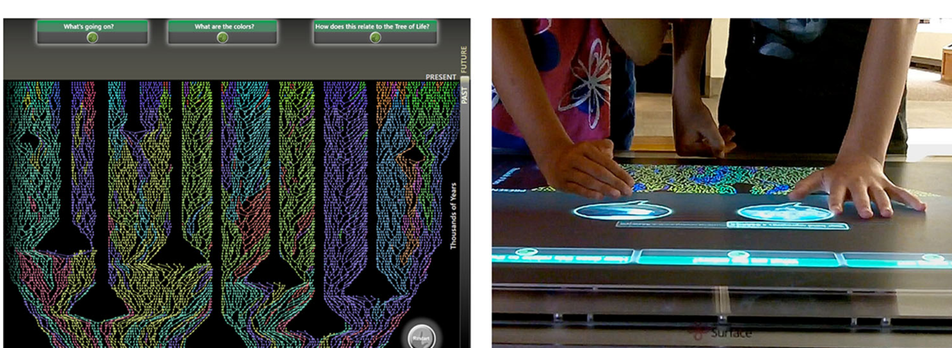

Figure 3. Screenshot from the FloTree interactive (left) and a dyad interacting with FloTree at anatural history museum (right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com].

significantly in presentation, we do report instances in which there were significant differences in visitor engagement or learning outcomes.

Exit Interview. We conducted a 15–20 minute audio-recorded interview with 53 open- and closed-ended questions to assess youths' understanding of microevolution and macroevolution concepts (see Supplementary Materials). Youth were interviewed individually without access to outside resources. For closed-ended questions, youth were trained to use five-point Likert scales (with faces representing each choice). Most of the interview items were developed specifically for this study. The natural selection questions were based on measures developed in prior research (Evans et al., 2010; Legare et al., 2013; Spiegel et al., 2012). Other measures were adapted from prior research on tree-reasoning ability (Novick, Catley, & Funk, 2010, 2011), evolutionary relatedness (Phillips, Novick, & Catley, 2011), and common ancestry (Poling & Evans, 2004).

Procedure

We recruited participants at two natural history museums, the Harvard Museum of Natural History (HMNH) and the Field Museum, Chicago. HMNH serves around 240,000 visitors annually, while the Field Museum serves over 1.2 million visitors annually. Our exhibit was installed in a hall of vertebrate paleontology at Harvard and near the entrance to the evolution hall in the Field Museum. At each site we recruited groups of visitors as they came into the vicinity of our exhibit. To be eligible to participate, visitor groups had to consist of at least one parent or guardian and at least two youth in the target age range of 8–15 years old. After obtaining informed consent, dyads were randomly assigned to one of the four conditions described below. Dyads were given a $15 gift for participating.

Conditions. In the first experimental condition (DeepTree I), youth dyads engaged in an unscripted exploration of the DeepTree for 4 minutes, followed by a forced transition to the FloTree application for 4 minutes, and concluded with an exploration of DeepTree for an additional 2.5 minutes. The exhibit software controlled the transitions between the DeepTree and FloTree. After the first 4 minutes of interaction, the software disabled the interface and prompted participants to press an Experiment button that launched the FloTree. A similar transition guided visitors back to the DeepTree for the final 2.5 minutes. In the second experimental condition (DeepTree II), youth dyads engaged in a 10.5 minute unscripted exploration of the DeepTree only. In both DeepTree conditions, if participants had not used the relate function after the first 90 seconds of interaction, they were prompted to do so by the exhibit. In the third experimental condition (Video), dyads watched the 10.5 minute video, *Discovering the Great Tree of Life*. The timing of the intervention was dictated by the video length, typical of museum settings. In the baseline condition, participants completed the exit interview before gaining access to the exhibit.

Exit Interview: Coding Open-Ended Responses and Constructing Measures

Participants' open-ended explanations provided critical information about their understanding of evolutionary concepts. Responses to 10 open-ended questions in the exit interview were evaluated with codes (see Table 2) based on systems used in prior research (Catley et al., 2013; Evans et al., 2010; Novick & Catley, 2012). Newly emergent codes were also included as needed. Two researchers achieved 96.7% agreement when coding a total of 14,586 responses ($\kappa = 0.681$). We used three main coding systems, each with a unique set of codes: *Biological terms* (11 codes) were linguistic codes assigned when the youth used the same or a closely associated term. The other coding systems were based on the concept, not necessarily the use of the correct term. *Informed reasoning* (8 codes) captured a relatively well-informed, but in most cases far from

Table 2
Open-ended codes used to construct measures of evolution understanding

Biological Terms	Example
Adaptation	"It's the one that has more chance of survival... that makes evolution and adaptation."
Ancestry	"A long time ago, they were common ancestors to us."
Branches	"I would tell them that every species splits and branches."
Genes	"... maybe because that they have the same, a couple of the same genes."
Relate	"A tree of how everyone relates."
Separate	"... everything got separated, and they all went in their different directions"
Time	"It's how creatures changed over time." And "the act of evolving over time."
Informed Reasoning	**Example**
Branching patterns	"I would say it's probably about evolution, going off to the different branches of human race and animals."
Common descent	"... humans, gorillas, orangutans, and chimpanzees evolved from these same type of ape-like creatures a long, long time ago."
Differential reproduction	"The ones that survived reproduced faster and had more population."
Differential survival	"It's the one who has more chance to survive, that in the population, that's make the evolution, and adaptation."
Environmental pressures	"Each different species adapted to its environment, and like, became different in some way, from its like ancestors ..."
Inheritance	"Those that had larger beaks were favored by the environment, so they were able to eat, breed, and then their offspring continued to do the same."
Shared traits	"They all have these characteristics of hair, amniotic sac, backbones ..."
Taxonomic relationships	"Because dolphins are mammals [...] meaning that they'll be closely related."
Intuitive Reasoning	**Example**
Connectedness	"It shows the connection between them."
Need-based reasoning	"Because each of them had their own specific need to live in a different environment, so they adapted to what they needed from the environment."

expert answer (Evans et al., 2010). For example, Taxonomic Relationships included statements that referenced valid biological groupings (e.g., "Because dolphins are mammals... they are in the same category, meaning that they will be closely related"). In contrast, *Intuitive reasoning* (9 codes) captured visitors' everyday reasoning, particularly anthropomorphic or teleological concepts. For example, Need-Based Reasoning included statements about the needs of organisms (e.g., "*Because each of them had their own specific need to live in a different environment...*"). Terms and concepts used rarely (1% or less) or considered peripheral to the main study questions were excluded from further analyses.

Individual codes were scored as 0 (absent) or 1 (used at least once). For example, if a youth mentioned Relate several times in response to a single question, they would be assigned a score of 1 for Relate for that question. For each question, youth responses were scored on a 0–1 scale for each of the 28 codes, which were then averaged across all 10 questions. We combined and averaged subsets of these codes to create measures of evolutionary reasoning. Table 2 shows only those codes used in the measures while Supplementary Table S2 shows the complete coding system. A summary of the interview protocol is also provided in supplemental materials.

Tree of life measures. To assess tree of life reasoning we constructed measures of participant use of *Tree Terms* (Terms: Ancestry, Branches, Relate) and *Tree Concepts* (Concepts: Branching

Patterns, Common Descent, Shared Traits, Taxonomic Relationships), in response to the 10 open-ended questions. In addition, to assess youths' initial responses, the first open-ended question asked "What [is] the tree of life all about?" Therefore we report participants' use of *Tree Terms* and *Tree Concepts* in response to this single question. Further, as the concepts of relationship and connectedness were hypothesized to be intuitive concepts associated with tree-of-life reasoning, we constructed a measure of the use of the *Relate Term* and its morphological variants and the use of the *Connectedness Concept* across the same 10 open-ended questions. For the latter code, we coded all responses that referred to species being "connected" without explicit reference to the degree of relatedness between species (e.g., The tree of life shows "how they are all attached to each other").

Participants were also asked to interpret a tree of life graphic with three closed-ended questions (*Tree Reading*). For each question, youth identified the species that have traits in common (e.g., "Point to the living things that have a backbone"). Accuracy on each of the three questions was averaged to produce a mean composite score of 0–1 (α: 0.76). Finally, participants were asked to indicate their agreement (1–5 scale) with five closed-ended questions related to *Common Ancestry* (adapted from Poling & Evans, 2004), each of which conveyed the idea that different kinds of organisms share ancestors (e.g., "Some kids said that bears and sunflowers had the same ancestor a long, long time ago. Do you agree or disagree with them?"). The mean score across the five questions yielded a 1–5 composite score (α: 0.81).

Evolution process (EP) measures. To assess evolution process reasoning, we coded each participant's use of *Evolution Process Terms* (Adaptation, Gene, Separate, Time) and *Evolution Process Concepts* (Differential Reproduction, Differential Survival, Environmental Pressures, Inheritance) in response to the same 10 open-ended questions. We also coded participants' use of intuitive *Need-Based Reasoning*.

For the closed-ended measures, participants indicated their agreement (1–5 scale) with five statements that evolution is an ongoing process (*Ongoing Evolution*: 1–5 scale; α: 0.70). Participants were also presented with four evolution process scenarios, each of which yielded four closed-ended statements and one of the 10 open-ended questions. The statements assessed youth informed and intuitive reasoning about evolution processes. Each EP closed-ended composite consisted of the averaged agreement score (1–5 scale) across the four statements, one for each scenario. For example, for *Natural Selection Agreement*, youth presented with the Canary Island Lizard scenario were first asked an open-ended question, "[...] How did it happen that there were so many brown-colored lizards on the sandy shores of the island?" Then they were asked how much they agreed with the statement: "[...] the seabirds ate the colorful lizards; the brown lizards lived and they had babies that looked like them." This kind of explanation was repeated for all scenarios.

These composite scores across the four scenarios yielded measures of: *Evolution Agreement* (α 0.78); *Natural Selection Agreement* (α: 0.53; M_{IIC}: 0.22); *Need-Based Agreement* (three items -α: 0.51; M_{IIC}: 0.27); *Want-Based Disagreement* (α: 0.64; M_{IIC}: 0.32); *Design-Based Disagreement* (α: 0.90). As scales with fewer than 10 items often have low alpha values, we also report the mean inter-item correlation (M_{IIC}: optimal range 0.2–0.4) for those measures with alphas below 0.70. The later two measures, *Want-Based Disagreement* and *Design-Based Disagreement* were intended to assess intentional or anthropomorphic reasoning.

Physical Interaction Measures. For our second research question, we analyzed the relationship between dyads' physical interaction with the tabletop exhibit and several learning outcomes. These measures were derived from an analysis of computer logs of participants' touch interactions with the tabletop. Seven touch logs were unavailable due to network connections problems. Because the touch sensing technology could not differentiate touches of individual participants (and because video recordings of the dyad sessions did not consistently include faces), these measures applied to the dyad as a whole. In total we had touch data for 54 dyads. We constructed one measure of dyads' overall touch interaction with the exhibit: *Total Touches*. This measure was a summation all touch-input events on the tabletop recorded by our event logging system ($M = 116.61$, $SD = 46.79$; Range = 42–221). We also recorded dyads' use of three key exhibit features. First we recorded the number of times dyads used the relate function to compare two species: *Relates Activated* ($M = 2.67$, $SD = 2.0$, Range = 0–10). Second, we recorded the number of times dyads then opened the simplified "training tree" shown in Figure 2: *Training Trees Activated* ($M = 3.00$, $SD = 2.1$, Range = 0−9). Finally, from the training tree, dyads could tap on glowing double helix icons to reveal more information about important evolutionary landmarks: *Traits Activated* ($M = 5.61$, $SD = 4.88$, Range = 0–17). When tapped, the software would display text, images, and in some cases, short video clips.

Verbal Engagement (Conversation). For the second research question, we also analyzed dyads' verbal engagement as they interacted with the tabletop exhibit. To measure verbal engagement we analyzed dyad conversation using the transcripts of discussion at the tabletop. We used a computer script to count occurrences of individual words in the transcripts grouped by the morphological stem related to a specific key concept. We then examined all words used at least ten times across all of the dyad sessions and created several categories (including the total number of words). In creating these categories we focused on key evolutionary or biological terms related to our learning outcomes as these plausibly signaled engagement with the material. We also included affect words, reasoning that these reflected deeper or more enjoyable levels of engagement and potentially better learning outcomes. Again, because we could not distinguish individual speakers from the session transcripts, these measures applied to the dyad as a group. We constructed the following five measures based on our word categories:

(1) *Total Words*. Mean: 406.92 (SD = 289.11, Range = 17–989).
(2) *Affect Words*. Mean: 7.61 (SD = 7.74, Range = 0–31); (love, cute, pretty, wow, cool, etc.).
(3) *Tree of Life Words*. Mean: 16.2 (SD = 15.32, Range = 0–62); (tree, relate, population, etc.).
(4) *Animal Words*. Mean: 9.33 (SD = 8.61, Range = 0–30); (cat, shark, banana, human, etc.).
(5) *Trait Words*. Mean: 2.76 (SD = 3.82, Range = 0–17); (eukaryotes, nuclei, DNA, cells, etc.).

Results

The results will be reported in two sections, each addressing a different research question.

RQ1: What Are the Effects of Condition and Age on Learning Outcomes?

To answer the first research question we conducted two-way ANOVAs on the learning outcomes, with condition (four: DeepTree I, DeepTree II, Video, and Baseline) and age group (two: Young, Old) as factors. Tukey post-hoc tests were used to evaluate the effects of condition[1]. Effect sizes were estimated using partial eta squared, which can be interpreted as: small 0.01–0.05, medium 0.06–0.137, or large 0.138 and higher (Cohen, 1988).

Tree of Life Results. Overall, for tree of life reasoning, results for all measures were similar. The DeepTree II condition and, to a lesser extent, the DeepTree I condition, elicited scores that were significantly higher than baseline (see Table 3). There were main effects of age and condition, but no interactions, an indication that both age groups made learning gains.

Table 3

Effects of condition and age-group (means, SD) on tree-of-life reasoning (F-, p-values)

Learning Measure	DeepTree I (DT1)	DeepTree II (DT2)	Video	Baseline	Effect of Condition (F)	DT1 versus Baseline (p)	DT2 versus Baseline (p)	Young Age-Group	Old Age-Group	Effect of Age (F)
First open-ended question: what is the tree of life all about?										
Tree terms	0.14 (0.19)	0.17 (0.20)	0.07 (0.15)	0.03 (0.10)	9.47***	0.001	0.0001	0.10 (0.17)	0.10 (0.17)	ns
Tree concepts	0.03 (0.08)	0.06 (0.12)	0.02 (0.07)	0.008 (0.04)	4.11**	ns	0.005	0.03 (0.08)	0.03 (0.09)	ns
Mean use across 10 open-ended questions										
Connect concept	0.03 (0.08)	0.04 (0.08)	0.07 (0.10)	0.05 (0.06)	2.48+	ns	ns	0.03 (0.06)	0.07 (0.10)	11.30**
Relate term	0.09 (0.11)	0.11 (0.12)	0.06 (0.09)	0.04 (0.07)	5.39**	0.049	0.001	0.06 (0.09)	0.08 (0.11)	ns
Tree terms	0.06 (0.05)	0.06 (0.05)	0.04 (0.05)	0.03 (0.04)	5.82**	0.004	0.002	0.04 (0.04)	0.05 (0.05)	6.58*
Tree concepts	0.06 (0.05)	0.07 (0.05)	0.06 (0.04)	0.04 (0.04)	2.89*	ns	0.018	0.05 (0.04)	0.06 (0.05)	8.97**
Closed-ended measures										
Common ancestry	2.93 (1.00)	2.93 (0.97)	2.80 (0.90)	2.49 (0.78)	3.49*	0.029	0.027	2.55 (0.73)	2.98 (1.02)	15.42***
Tree reading	0.63 (0.41)	0.77 (0.32)	0.66 (0.38)	0.60 (0.39)	2.87*	ns	0.034	0.52 (0.40)	0.80 (0.32)	38.48***

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

For the first open-ended question ("What is the tree of life all about?"), the use of *Tree Terms* was significantly higher for participants in both DeepTree conditions ($\eta_p^2 = 0.11$), with no effect of age and no significant interaction (see Table 3 for means, standard deviations, F tests, and p-values). Likewise, for *Tree Concepts* there was a significant effect of condition ($\eta_p^2 = 0.05$), but, in this case, DeepTree II was the only condition significantly different from baseline.

Overall, across the 10 open-ended questions there was a significant main effect of condition for the *Relate Term* ($\eta_p^2 = 0.06$), *Tree Terms* ($\eta_p^2 = 0.07$) and *Tree Concepts* ($\eta_p^2 = 0.04$), with participants in both DeepTree conditions scoring significantly higher than the baseline in all but one case (see Table 3). There was no effect of age for the *Relate Term* ($\eta_p^2 = 0.01$), but there were significant effects of age and no interactions for *Tree Terms* ($\eta_p^2 = 0.03$) and *Tree Concepts* ($\eta_p^2 = 0.04$). For *Tree Terms*, participant responses in the DeepTree II condition were also significantly different from those in the Video condition ($p = 0.005$). For the *Connectedness Concept*, a marginal main effect was found for condition: $F(3,239) = 2.48, p = 0.062, \eta_p^2 = 0.03$. In this case only, youth in the Video condition were significantly more likely to use this concept than those in the DeepTree I condition; there were no other condition effects, but there was a significant effect of age, with older children more likely to endorse connectedness (See Table 3).

A similar pattern was apparent for the closed-ended measures. For *Common Ancestry* there was a main effect of condition ($\eta_p^2 = 0.04$) and age ($\eta_p^2 = 0.06$) and no significant interaction. Both DeepTree conditions were significantly different from baseline (but not from each other), with DeepTree II also significantly different from the Video condition ($p = 0.005$). For *Tree Reading*, there were main effects for condition ($\eta_p^2 = 0.04$) and age ($\eta_p^2 = 0.18$), with older participants and those in the DeepTree II condition performing at a significantly higher level.

The overall pattern for the condition effect was a clear finding that participants in DeepTree II consistently scored at higher levels (compared to baseline) on the closed-ended measures and were more likely to use tree terms and concepts in their explanations, regardless of age. Although DeepTree I was less effective overall (compared to baseline) for the *Tree Reading* measure and for eliciting *Tree Concepts*, it was similarly effective at eliciting evolutionary terms and higher *Common Ancestry* agreement scores. The Video condition consistently elicited higher scores compared to the baseline, but the differences were not significant (see Table 3).

The main effects of age were driven by older youth who consistently performed at a higher level than younger youth on all the main measures. Two deviations from this pattern were found for *Relate Term* and responses to the initial question ("What is the tree of life all about?") where there were no significant age-related differences. This finding suggests that while younger participants were able to use the knowledge they gained from the tabletop interactions to respond to the first tree of life question, it was the older youth who were better able to extend this knowledge effectively across all 10 questions. It also suggests that relatedness, but not connectedness, may be a bridging concept enabling youth to make links between intuitive and more informed reasoning.

Evolution Process Results. In contrast to the tree of life measures, for evolution process (EP) reasoning there were main effects of age but no main effects for condition with no significant interactions (see Table 4). For the informed reasoning measures: *Evolution Process Terms* ($\eta_p^2 = 0.05$), *Evolution Process Concepts* ($\eta_p^2 = 0.05$), *Ongoing Evolution* ($\eta_p^2 = 0.07$), *Natural Selection Agreement* ($\eta_p^2 = 0.02$), and *Evolution Agreement* ($\eta_p^2 = 0.05$), older youth performed at significantly higher levels.

Similarly, for the intuitive reasoning concepts there were no effects of condition. For the open-ended questions, older youth were more likely than younger youth to use *Need-Based Reasoning* ($\eta_p^2 = 0.03$). Likewise, on the closed-ended questions across the four scenarios, older youth were more likely to endorse *Need-Based Agreement* ($\eta_p^2 = 0.02$), *Design-Based*

Table 4
Effect of age-group (mean, SD) on evolutionary process (EP) reasoning

Measure	Young age group	Old age-group	Effect of age (F)
Mean use across 10 questions			
Evolution process terms	0.05 (0.06)	0.08 (0.06)	13.64***
Evolution process concepts	0.03 (0.03)	0.04 (0.04)	14.31***
Need-based reasoning	0.12 (0.15)	0.17 (0.17)	6.04*
Closed-ended measures: informed evolutionary process reasoning			
Ongoing evolution	3.70 (0.76)	4.12 (0.71)	19.61***
Natural selection agreement	3.63 (0.72)	3.85 (0.75)	5.55*
Evolution agreement	3.55 (0.88)	3.97 (0.90)	13.40***
Closed-ended measures: intuitive evolutionary process reasoning			
Need-based agreement	3.59 (0.74)	3.82 (0.81)	5.30*
Design-based disagreement	3.98 (1.02)	4.25 (1.00)	4.57*
Want-based disagreement	3.07 (0.66)	3.30 (0.52)	9.48**

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Disagreement ($\eta_p{}^2 = 0.02$), and *Want-Based Disagreement* ($\eta_p{}^2 = 0.04$) (see Table 4). This age-related pattern of endorsements for the intuitive concepts is consistent with prior research (Legare et al., 2013; Spiegel et al., 2012).

One possible reason for the lack of a significant effect of condition for the evolution process reasoning overall may have been that the measures themselves did not have good construct validity. Theoretically, all five informed evolution process measures should be positively correlated with: one another (rs = 0.18–0.38; ps < 0.01), age (rs = 0.14–0.31; ps < 0.05), the tree of life measures, *Common Ancestry* and *Tree Reading* (rs = 0.18–0.47; ps < 0.01), *Need-Based Agreement* (rs = 0.17–0.45; ps < 0.01), *Design-Based Disagreement* (rs = 0.14–0.25; ps < 0.05), and *Want-Based Disagreement* (rs = 0.13–0.23; ps < 0.01).[2] Although there was variation in the strength of the correlations, this pattern of relationships is consistent with the argument that these variables were assessing participants' understanding of evolutionary processes.

RQ2: How do Verbal Engagement and Physical Interaction Contribute to Learning Outcomes in the Tabletop (DeepTree) Conditions?

The second research question focused on features of youth engagement that were likely to explain the learning outcomes for our tree of life measures. For this question we focused exclusively on the tabletop conditions because participants could not physically interact with the video, and because we assumed that participant speech in the Video condition would be very limited. One reason for this assumption is that the multi-touch tabletop interface often requires dyads to negotiate their exploration of the content, particularly when they have conflicting ideas about what to do. In contrast, we believed that the voiceover narrative in the Video condition would allow for less discussion. To verify this assumption, we transcribed the video recordings of the dyad discussion in the three experimental conditions. Due to background noise in the museum environment, the audio was not of sufficient quality to produce a transcript in all cases. In total, we transcribed 83 of 93 sessions (27 of 30 in DeepTree I; 29 of 31 in DeepTree II; and 27 of 32 in the Video condition). When participant voices were not clear enough, we used an *inaudible* marker in the transcripts. As described earlier, it was not possible to individuate the conversation because the video recordings did not always include the faces of the participants. As an approximation of the overall level of verbal interaction, we counted the number of words spoken by both participants. Inaudible segments were counted as one word. On average dyads in DeepTree I spoke 444.85

words per session (SD = 227.32), while dyads in the DeepTree II spoke 434.83 words per session (SD = 290.56). Three dyads across both tabletop conditions did not speak at all during their entire sessions. In contrast, dyads in the Video condition spoke an average of 6.96 words per session (SD = 14.60), only 1.6% of the words spoken by participants in the tabletop conditions. Notably, 20 dyads in the Video condition did not speak at all.

Having established that verbal exchanges were minimal in the Video condition (in comparison to DeepTree I and DeepTree II), we focused our attention on the tabletop conditions and the relationship between measures of verbal engagement, physical interaction, and learning outcomes from the exit interview. Because the two DeepTree conditions did not differ significantly from one another in RQ1, we combined their data for these analyses. We first report correlations between our verbal and physical measures and the learning outcomes. This is followed by series of analyses using Linear Mixed Models in which we examined the individual contributions of physical and verbal engagement to the different learning outcomes. Age and family background were also included in the models. Although the latter variables did not differ significantly between conditions, they were likely to contribute to learning outcomes within a condition.

Correlation Between Verbal and Physical Measures and Learning Outcomes. We first related our measures of physical interaction (described in the Procedure section) to four key learning outcome measures for which there were consistent significant effects of condition in RQ1 (see Table 3). For the two of the measures, *Common Ancestry* and *Tree Reading*, there were strong effects of age as well. From the open-ended measures, we selected *Relate Term* because there was no main effect of age, suggesting that it reflected a more basic understanding of tree reasoning accessible to both age groups; on the other hand, *Tree Concepts* elicited a strong effect for age, suggesting, in turn, that it reflected a deeper level of understanding. This analysis allowed us to explore the relationship between age, engagement, and levels of evolutionary reasoning, a necessary first step in the construction of a developmental learning trajectory.

As can be seen in Table 5 (upper half), for the measures of physical interaction youth who more often activated the relate, training tree, and trait functions in the tabletop exhibit were more likely to use the *Relate Term* in the open-ended questions and achieve higher scores on *Tree Reading* in the subsequent exit interview. Additionally, *Relates Activated* was significantly

Table 5

Correlations between learning outcomes and measures of physical engagement (top) and social engagement (bottom) across the two tabletop conditions (DeepTree I and DeepTree II)

	Relate Terms	Tree Concepts	Common Ancestry	Tree Reading
Measures of physical engagement ($n = 107$)				
Total touches	ns	0.16^{+}	0.19^{+}	ns
Relates activated	0.36^{**}	ns	0.20^{*}	0.22^{*}
Training trees activated	0.24^{*}	ns	ns	0.19^{*}
Traits activated	0.30^{**}	ns	0.16^{+}	0.26^{**}
Measures of social engagement ($n = 113$)				
Total words	ns	0.21^{*}	0.31^{**}	ns
Affect words	0.21^{*}	0.27^{**}	0.33^{**}	0.22^{*}
Tree words	ns	ns	0.29^{**}	ns
Animal words	0.30^{**}	0.16^{+}	0.24^{*}	0.16^{+}
Trait words	0.21^{*}	0.21^{*}	0.30^{**}	ns

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

correlated with *Common Ancestry*. There was no significant correlation between *Total Touches* and any learning outcome. Further, none of the physical interaction measures were correlated with the use of *Tree Concepts* in the exit interview.

The correlations between these measures of verbal engagement (described in the Procedure section) and the same four outcome measures (Table 5, lower half) demonstrate a consistent pattern. There were positive correlations between use of particular content-related words in dyad conversation and most of the subsequent learning outcomes. Most notably, the more *Affect Words* used during the exhibit interaction, the more likely youth were to score at higher levels on all four learning outcomes.

Relationship Between Verbal and Physical Measures. As we expected, there were significant correlations between our measures of physical interaction and verbal engagement. This finding indicates that conversation and physical activation of the content went hand-in-hand. Specifically, *Relates Activated* was positively correlated with *Affect Words* ($r = 0.25$, $p = 0.018$) and *Animal Words* ($r = 0.34$, $p = 0.001$); *Traits Activated* was positively correlated with *Animal Words* ($r = 0.30$, $p = 0.001$) and *Trait Words* ($r = 0.23$, $p = 0.028$); and *Total Touches* was positively correlated with *Affect Words* ($r = 0.21, p = 0.049$).

Although there were positive relationships between the physical and verbal measures, the pattern of correlations suggests that they contributed to learning outcomes in different ways. For example *Common Ancestry* agreement was significantly correlated with all of the measures of verbal engagement, but only one measure of physical interaction, *Relates Activated*. *Tree Reading*, on the other hand, was positively correlated with the three key measures of physical interaction, but with only one verbal engagement measure, *Affect Words*.

Linear Mixed Models. To understand the contribution of our engagement measures to learning outcomes, in our final analysis we used linear mixed models (LMMs). Our models focused on the effects of verbal engagement and physical interaction on the four overall outcome measures as well as on two outcome measures from the first open-ended question (see Table 3). In each of these analyses we used a LMM with a random effect per dyad to take into account the correlation among measures for individuals within the dyad. All other variables were entered as fixed effects. It should be noted that for LMM analyses there is no commonly accepted assessment of the overall variance explained by the model (Nezlek, 2008); however, as it is the individual contributions of each predictor that is of interest in this study, those statistics will be reported.

The key question addressed with these analyses is whether the physical and verbal measures elicited different learning outcomes. We were further interested in whether the significant correlations between measures of dyadic engagement and learning outcomes (reported above) were a reflection of age, prior knowledge, or family background of participants. In other words, were more knowledgeable youth more likely to find the exhibit engaging and thus more likely to do well on the learning measures? Or, did higher levels of youth engagement elicit better learning outcomes independent of family background?

To assess these questions, we used parents' endorsement of evolutionary origins (see Table S1) as an indicator of family background. This measure resembled the content of the learning outcomes and was positively correlated with other relevant parent variables including parental rating of the importance of evolution for scientists/self ($r = 0.60, p = < 0.001$) and parent education level ($r = 0.36, p = < 0.001$). Parental rating of youth evolution knowledge was used as a proxy measure of youth prior knowledge. This measure was positively correlated with all four learning outcomes (rs 0.14–0.20, $ps < 0.05$). Thus, for each of the following analyses we included youth age, parent endorsement of evolutionary origins (*Parent Belief*), and youth evolution knowledge (*Youth Knowledge*) as predictors, along with the most highly correlated measures of physical and/or verbal engagement for each of the main learning outcomes (see Table 5). Here we present the LMM analysis for each learning outcome:

Relate Terms Used. Youth age, *Parent Belief*, *Youth Knowledge*, *Animal Words*, and *Relates Activated* were included in the model for *Relate Terms*. Significant effects independent of the other variables were found for *Animal Words* (Est. 0.003, SE 0.001, df 43.7, $t = 2.23, p = 0.031$), *Relates Activated* (Est. 0.02, SE 0.006, df 53.2, $t = 2.65, p = 0.010$), and *Parent Belief* (Est. 0.2, SE 0.01, df 46.13, $t = 2.08, p = 0.043$).

Common Ancestry. Youth age, *Parent Belief*, *Youth Knowledge*, *Affect Words*, *Trait Words*, and *Relates Activated* were included in the model for *Common Ancestry Agreement*. Significant effects independent of the other variables were found for *Affect Words* (Est. 0.03, SE 0.01, df 45.2, $t = 2.39, p = 0.021$) and *Trait Words* (Est. 0.07, SE 0.03, df 44.5, $t = 2.43, p = 0.019$). A marginal effect was found for youth age as well (Est. 0.11, SE 0.06, df 82.27, $t = 1.94, p = 0.056$).

Tree Concepts. Youth age, *Parent Belief*, *Youth Knowledge*, *Affect Words*, and *Traits Activated* were included in the model for *Tree Concepts*. Significant effects independent of the other variables were found for youth age only (Est. 0.08, SE 0.003, df 78.7, $t = 2.32, p = 0.023$).

Tree Reading. Youth age, *Parent Belief*, *Youth Knowledge*, *Affect Words*, and *Traits Activated* were included in the model for *Tree Reading* accuracy. Significant effects independent of the other variables were found for age only (Est. 0.08, SE 0.02, df 84, $t = 3.26, p = 0.002$).

What Is the Tree of Life All About?. LMMs for the two measures used to assess youth responses to the first open-ended question about the tree of life were included because they offered insights into the immediate effects of the exhibit. The *Tree Terms* and *Tree Concepts* found in youth explanations for the opening question were significantly correlated with *Relates Activated* (*r*s 0.29, $ps = 0.003$) in the DeepTree exhibit and the *Trait Words* in dyads' conversation (*r*s 0.19–0.24, $ps = < 0.05$). Youth age, *Parent Belief*, *Youth Knowledge*, *Trait Words*, and *Relates Activated* were included in the models for the two outcomes, in turn: (i) For *Tree Terms*, significant effects independent of the other variables were found for *Relates Activated* (Est. 0.26, SE 0.01, df 47.8, $t = 2.56$, $p = 0.014$), *Trait Words* (Est. 0.12, SE 0.005, df 41.57, $t = 2.12$, $p = 0.040$), and *Parent Belief* (Est. 0.4, SE 0.02, df 42.95, $t = 2.20$, $p = 0.033$); (ii) For *Tree Concepts*, significant effects independent of the other variables were found for *Relates Activated* (Est. 0.01, SE 0.005, df 50.2, $t = 2.21, p = 0.032$) and a marginal effect for *Trait Words* (Est. 0.05, SE 0.003, df 43.8, $t = 1.87, p = 0.068$).

Overall, the LMMs demonstrate that measures of verbal engagement and physical interaction explain variance in the learning outcomes independent of one another and independent of prior knowledge and parent acceptance of evolution. It should be noted, however, that age was the main independent predictor of two outcomes: *Tree Reading* accuracy and *Tree Concepts* found in youth responses to the 10 open-ended questions. For these two outcomes, older youth were more likely to benefit from the exhibit interaction, regardless of family background. However, measures of engagement did predict other learning outcomes, regardless of age and family background. Specifically, activation of the relate function on the tabletop and the use of animal words in dyad conversation predicted the frequency of *Relate Terms* in the overall explanations. Similarly, activation of the relate function and use of trait words in dyad conversation predicted the frequency of *Tree Terms* and *Tree Concepts* in response to the first open-ended question. Moreover, the frequency of trait and affect words in the dyad conversation predicted the likelihood that youth would endorse the rather abstract concept of *Common Ancestry.*

Discussion

The popularity of interactive surfaces in museums has created unique opportunities for visitors to “touch” and explore large scientific datasets. Beyond reflecting the increasingly computational nature of science, such experiences may create new opportunities for learning. While we know that large evolution exhibitions with multiple interactive components can provide effective learning experiences (Evans et al., 2015; Spiegel et al., 2012; Tare et al., 2011), the current study addressed whether learning occurs in a brief interaction with a dynamic visualization of the tree of life including over 70,000 species. We were also interested in understanding how different features of physical interaction and verbal engagement contributed to visitor learning with the multi-touch tabletop.

Our first research question focused on the effects of exhibit condition and age on youth understanding of evolution concepts. The DeepTree conditions engaged youth dyads in the exploration of a large interactive phylogenetic tree. The DeepTree I condition also included an embedded activity on evolutionary processes called FloTree. The Video condition, meanwhile, consisted of a video of the same length on similar evolution concepts. Outcomes were compared to those of youth in a baseline condition with no intervention. The overall pattern of our results comparing conditions was very clear. Youth in the DeepTree conditions (and DeepTree II, in particular) consistently scored at higher levels than youth in the baseline condition on both open and closed-ended measures of shared ancestry, common descent, and the tree of life. Specifically, youth in the DeepTree conditions were significantly more likely to invoke tree of life concepts and terminology in their open-ended responses. These subjects were also significantly more likely to correctly interpret a phylogenetic tree diagram and endorse ideas of common ancestry in closed-ended items. Surprisingly, a brief, open-ended museum experience yielded consistent learning outcomes about phylogeny, a complex and difficult science concept. Furthermore, there were significant main effects of age for many of our measures. Older youth demonstrated a more consistent and informed understanding of evolution than younger youth, with the exception of basic concepts of relatedness, which were the same for both groups.

Our study design also included a Video condition as a way to represent a typical learning experience that visitors might encounter at a natural history museum. Our results show that while there were positive trends across many of our measures for the Video condition, almost none of the learning gains were significant with respect to the baseline. Notably, apart from connectedness, the expert language used by the narrators in the video did not seem to elicit significant comparable language in the youth explanations. Participation in the tabletop conditions, in contrast, was associated with an increase in evolutionary language and concepts.

Although the current study focused heavily on youth understanding of macroevolutionary concepts, the FloTree component of the DeepTree I condition addressed microevolutionary processes as well. Counter to our predictions, the FloTree application did not facilitate youth understanding of processes such as differential survival and differential reproduction. One possible explanation is that the forced transition to FloTree may have distracted participants while shortening the overall exposure to the individual components. However, the animated portrayal of natural selection in the Video condition was also unsuccessful in this regard.

Understanding Contributions of Verbal Engagement and Physical Interaction

In our second research question we investigated the effects of verbal engagement and physical interaction on youth learning in the tabletop conditions (DeepTree I and DeepTree II). Using video transcripts and computer logs, we constructed several measures of physical interaction with the tabletop and verbal interaction between participants. We then examined correlations between these measures of engagement and the key learning outcomes. These analyses revealed significant relationships. Even though our physical and verbal measures were inter-correlated, the pattern of relationships suggested that they contributed to learning outcomes in different ways. Optimal learning outcomes occurred when youth dyads both activated relevant exhibit functions and conversed about the specific experience. This pattern was confirmed through the use of linear mixed models. These models indicated that several measures of engagement specifically predicted higher learning outcomes for our tree of life measures. In particular, youth who activated the relate function more frequently were more likely to use the relate term in their responses to open-ended questions and to use tree terms and tree concepts in their response to the first open-ended question on the tree of life. Moreover, dyads whose conversation included higher numbers of affect and trait words were more likely to endorse the idea that diverse species have an ancestor in common. Notably these relationships held even when controlling for family background, youth age, and prior knowledge. These results also highlight the fact that the overall level of verbal engagement (total number of words spoken) and the overall level of physical interaction (total number of touches) were not the best predictors of learning. Rather, learning depended on the specifics of *what* youth were saying and *how* they used the table. Moreover, affect words (such as *wow*, *cool*, and *hah*) were significantly correlated with *all* of the learning measures we considered. Our measures of engagement do not address more nuanced elements of dyadic interaction and shared meaning making. However, we have conducted a detailed qualitative analysis of interaction and learning based on video recordings of ten dyads from this study, which is the focus of another paper (Davis et al., 2015).

Towards a Developmental Learning Trajectory

Our age-related findings also offer insight into the concept of a developmental learning trajectory for understanding common descent. Youth in both age groups benefited from interacting with DeepTree, indicating that the exhibit was successful for different levels of prior knowledge. Moreover, the age-related patterns suggested a learning trajectory for the acquisition of tree-of-life concepts, from relatedness, to shared ancestry, to more complex tree concepts.

Activation of the relate function in the exhibit and use of "animal terms" in the conversation were associated with an increased understanding of evolutionary relationships in the exit interview, for both age groups. Moreover, there were no significant age-related differences in youth use of the relate term. In this case, youth appeared to be relying both on intuitive notions of family relatedness (e.g., the tree of life is about "how you are related to someone's family" 10-year-old #571b) as well as more expert explanations of evolutionary relatedness (e.g., the tree of life is about "how things relate... like billions of years ago... it shows how, like, bananas and squids... how they were like kind of the same, once" 12-year-old #556a). These data suggest that reasoning about family relationships may facilitate rather than impede youth's interpretation and understanding of common descent. Older youth, though, were better at decoding these relationships in the tree of life graphic and employing more complex tree concepts, such as branching patterns and shared traits, in their explanations in the exit interview.

This pattern for the relatedness concept is consistent with prior research suggesting that intuitive reasoning patterns are not necessarily abandoned or "overcome" as students acquire evolutionary constructs. Rather, they may provide a foundation for a more scientifically accurate understanding (Evans et al., 2012). For example, in this study, in contrast to their younger siblings, older youth were more likely to incorporate need-based reasoning (e.g., "because the different kinds [of anoles] need to adapt to their different environments" 14-year-old #122b) in their responses, while rejecting the anthropomorphic explanations (e.g., The lizards changed over time because "they do not like to get eaten" 11-year-old #559b). Moreover, in contrast to anthropomorphic reasoning, need-based reasoning was positively associated with the evolutionary process learning outcomes. These findings bolster the argument that need-based reasoning can potentially provide a foundation for a more sophisticated understanding of microevolution, if it is disassociated from anthropomorphic concepts (Legare et al., 2013; Spiegel et al., 2012).

This kind of logic could also be applied to essentialist reasoning. Perhaps essentialism is not necessarily the barrier to macroevolutionary reasoning that prior research has claimed (Gelman & Rhodes, 2012; Shtulman & Schulz, 2008). By activating the relate function and conversing about species and their shared traits, youth were repeatedly exposed to the idea that diverse species are related. Such youth were more likely to endorse the idea of common ancestry. We propose that these youth generalized their concept of "essence," from its original application to a single species or kind, to all living things. DNA now represents the "essence" of our shared evolutionary heritage, the family of all living things on Earth.

Limitations

There were limitations of this study that should be taken into account when interpreting these results. Foremost, we assessed only short-term learning outcomes immediately following the intervention. While we acknowledge this as a limitation, we point out that establishing short-term learning gains is a crucial first step. Furthermore, the prevalence of affect words in participant speech gives us some hope that long-term gains in youth understanding are feasible. Research on the neurobiology of memory, for example, indicates that emotionally arousing stimuli are more likely to be consolidated and preserved over the long term (McGaugh, 2006). We also note that there were limits to the ecological validity of our design. In particular, youth were recruited, video recorded, and asked to participate with a sibling or friend for a fixed period of time, all of which are known to affect participant behavior (Block et al., 2015). However, some degree of control was necessary for us to collect in-depth data on engagement and to establish statistically significant differences based on youth age and condition. Based on naturalistic observations conducted as part of a summative evaluation of the exhibit, we found that active and prolonged periods of engagement were not uncommon among dyads or visitor family groups, suggesting that our experimental setup had some correspondence to the types of engagement we might expect to see with more informal use of the DeepTree exhibit (Block et al., 2015). Finally, our sample reflects audiences that typically attend natural history museums in that most participant families were well educated and not necessarily representative of the broader population (Korn, 1995).

Implications

Taken together, these findings suggest important implications for the design of exhibits featuring visualizations of large scientific datasets. The most obvious implication is to provide adequate support for social interaction. Large interactive surfaces such as multi-touch tabletops can be effective for encouraging simultaneous use by multiple visitors, but this does not imply that visitors will interact or work together in productive ways. In fact, conflict, interference, and confusion are more likely outcomes in the absence of careful design and testing. Given the level of verbal engagement that we observed and its positive contribution to learning outcomes, we believe that promoting effective social interaction warrants special attention in the design process.

A second implication relates to self-directed engagement. Through our iterative design work we found that it was important to provide visitors with the opportunity for open-ended exploration with the support of "gentle guidance" (Humphrey & Gutwill, 2005) built into the interaction. Along these lines, including small amounts of video or expository text seemed valuable provided that they did not interfere with visitors' sense of control. The forced transition to the FloTree in the first tabletop condition seemed, in retrospect, counterproductive to learning. In this case, participants were presented with a highly interactive experience, yet in the absence of sufficient guidance were unable to interpret the microevolutionary processes displayed. Video of participants in this condition showed that the forced transitions were often confusing, interrupting otherwise productive sessions.

A third implication derives from the significant effect of the built-in relate function on the learning outcomes. DeepTree gives visitors the ability to repeatedly compare species across the span of all domains of life on Earth. This provided novices with an intuitive stepping stone from which to transition from an everyday understanding of "relationships" toward the *scientific* concept of evolutionary relatedness. Our results indicate that such scaffolding, in the form of repeated use of the relate function, contributed to the successful learning experiences. Similar intuitive conceptual mappings will likely apply to other scientific disciplines.

The final implication is that interactive visualizations of large scientific datasets hold promise for promoting learning about complex science concepts in museums. These exhibits can be useful as a way for natural history museums to reflect the changing nature of scientific inquiry, an endeavor that increasingly relies on large data sets and computational tools and methods. But they can also be used to create new types of learning experiences for visitors. In sum, while our learning objectives and measures concerned concepts of evolution and biodiversity, we believe that our findings make a compelling case that such experiences are worthy of further study across a broader array of science concepts.

We are grateful to the Harvard Museum of Natural History and the Field Museum for allowing us to conduct this research in their galleries. We thank our science advisers, Gonzalo Giribets, James Hanken, Hopi E. Hoekstra, Jonathan Losos, David Mindell, Sebastian Velez, and Mark Westneat, and the researchers who assisted with data collection and analysis, Elizabeth Bancroft, Pryce Davis, Ashley Hazel, Linying Ji, Christina Krist, Novall Khan, Kay Ramey, Laurel Schrementi, Amy Spiegel, Azalea Vo, and Nan Xin. Finally, we thank the National Science Foundation for their support of this project through grant, DRL-1010889. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Notes

[1]Two alternative analyses were also conducted to investigate the age effects: (1) Using ANCOVAs, with age as a continuous covariate, we checked whether the age effect was underestimated in the ANOVAs (2) Using linear mixed models to account for possible non-independent age data for dyads in the tabletop conditions. As the results were essentially the same as those for the ANOVAs, we used the latter analysis as it was easier to present the age-group results (in RQ2, age effects in the table top conditions were evaluated using linear mixed models).

[2]Exceptions to this pattern were the non-significant correlations between *Natural Selection Agreement* and (1) *Want*- and (2) *Design-Based Disagreement*; these occurred because of interactions with age-group, assessments of which are beyond the scope of this paper.

References

Antle, A. N., Tanenbaum, J., Seaborn, K., Bevans, A., & Wang, S. (2011). Futura: Design for collaborative learning and game play on a multi-touch digital tabletop. In Proceedings of Tangible, Embedded and Embodied Interaction (pp. 93–100). ACM Press.

Banet, E., & Ayuso, G. E. (2003). Teaching of biological inheritance and evolution of living beings in secondary school. International Journal of Science Education, 25, 373–407.

Baum, D. A., Smith, S., & Donovan, S. (2005). Tree-thinking challenge. Science, 310, 979–980.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. Journal of Research in Science Teaching, 27, 415–427.

Block, F., Hammerman, J., Horn, M. S., Phillips, B. C., Evans, E. M., Diamond, J., & Shen, C. (2015). Fluid grouping: Quantifying group engagement around interactive tabletop exhibits in the wild. In Proceedings ACM Conference on Human Factors in Computing Systems (pp. 867–876). ACM Press.

Block, F., Horn, M. S., Phillips, B. C., Diamond, J., Evans, E. M., & Shen, C. (2012a). DeepTree exhibit: Visualizing the tree of life to facilitate informal learning. IEEE Transaction on Visualization & Computer Graphics, 18(12), 2789–2798.

Block, F., Wigdor, D., Phillips, B. C., Horn, M. S., & Shen, C. (2012b). FlowBlocks: A multi-touch UI for crowd interaction. In Proceedings ACM Symposium on User Interface Software and Technology (pp. 497–508). ACM Press.

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, 2327–2348.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cracraft, J., & Donoghue, M. J. (2004). Assembling the tree of life. New York, NY: Oxford University Press.

Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., & Shrager, J. (2001). Shared scientific thinking in everyday parent-child activity. Science Education, 85, 712–732.

Diamond, J., & Evans, E. M. (2007). Museums teach evolution. Evolution, 61, 1500–1506.

Diamond, J., & Scotchmoor, J. (2006). Exhibiting evolution. Museums & Social Issues, 1, 21–48.

Davis, P., Horn, M. S., Block, F., Phillips, B., Evans, E. M., Diamond, J., & Shen, C. (2015). "Whoa! We're going deep in the trees!": Patterns of collaboration around an interactive information visualization exhibit. International Journal of Computer-Supported Collaborative Learning, 10, 53–76.

Eberbach, C., & Crowley, K. (2005). From living to virtual: Learning from museum objects. Curator: The Museum Journal, 48, 317–338.

Evans, E. M. (2000). The emergence of beliefs about the origins of species in school-age children. Merrill-Palmer Quarterly, 46, 221–254.

Evans, E. M. (2001). Cognitive and contextual factors in the emergence of diverse belief systems: Creation versus evolution. Cognitive Psychology, 42, 217–266.

Evans, E. M. (2013). Conceptual change and evolutionary biology: Taking a developmental perspective. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 220–239). New York, NY: Routledge [Second Edition].

Evans, E. M., Rosengren, K., Lane, J. D., & Price, K. S. (2012). Encountering counterintuitive ideas: Constructing a developmental learning progression for biological evolution. In K. R. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), Evolution challenges: Integrating research and practice in teaching and learning about evolution (pp. 174–199). New York, NY: OUP.

Evans, E. M., Spiegel, A. N., Gram, W., Frazier, B. N., Tare, M., Thompson, S., & Diamond, J. (2010). A conceptual guide to natural history museum visitors' understanding of evolution. Journal of Research in Science Teaching, 47, 326–353.

Evans, E. M., Weiss, M., Lane, J., & Palmquist, S. (2015). The spiral model: Integrating research and exhibit development to foster conceptual change. In D. M. Sobel, & J. Jipson (Eds.), Cognitive development in museum settings: Relating research and practice (pp. 36–64). New York, NY: Psychology Press.

Falk, J. H., & Dierking, L. D. (2000). Learning from museums: Visitor experiences and the making of meaning. Walnut Creek, CA: Altamira Press.

Foster, I. (2006). 2020 computing: A two-way street to science's future. Nature, 440(7083), 419–419.

Gelman, S. A., & Rhodes, M. (2012). Two-thousand years of stasis. In K. R. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), Evolution challenges: Integrating research and practice in teaching and learning about evolution (pp. 200–207). New York, NY: Oxford.

Hinrichs, U., & Carpendale, S. (2011). Gestures in the wild: Studying multi-touch gesture sequences on interactive tabletop exhibits. In Proceedings ACM Conference on Human Factors in Computing Systems (pp. 3023–3032). ACM Press.

Hinrichs, U., Schmidt, H., & Carpendale, S. (2008). EMDialog: Bringing information visualization into the museum. IEEE Transactions on Visualization and Computer Graphics, 14, 1181–1188.

Horn, M. S., Leong, Z. A., Block, F., Diamond, J., Evans, E. M., Phillips, B., & Shen, C. (2012). Of BATs and APEs: An interactive tabletop game for natural history museums. In Proceedings ACM Conference on Human Factors in Computing Systems (pp. 2059–2068). ACM Press.

Hornecker, E. (2008). "I don't understand it either, but it is cool"-visitor interactions with a multi-touch table in a museum. Proceedings IEEE Tabletop, 2008, (pp. 113–120).

Humphrey, T., & Gutwill, J., Exploratorium APE Team. (2005). Fostering active prolonged engagement. San Francisco, CA: The Exploratorium.

Korn, R. (1995). An analysis of differences between visitors at natural history museums and science centers. Curator: The Museum Journal, 38(3), 150–160.

Legare, C. H., Lane, J. D., & Evans, E. M. (2013). Anthropomorphizing science: How does it affect the development of evolutionary concepts? Merrill-Palmer Quarterly, 59, 168–197.

Louw, M., & Crowley, K. (2013). New ways of looking and learning in natural history museums: The use of gigapixel imaging to bring science and publics together. Curator, 52, 87–104.

Ma, J., Liao, I., Ma, K. L., & Frazier, J. (2012). Living liquid: Design and evaluation of an exploratory visualization tool for museum visitors. IEEE Transactions on Visualization and Computer Graphics, 18, 2799–2808.

MacDonald, T., & Wiley, E. O. (2012). Communicating phylogeny: Evolutionary tree diagrams in museums. Evolution: Education and Outreach, 5, 14–28.

MacFadden, B. J., Dunckel, B. A., Ellis, S., Dierking, L. D., Abraham-Silver, L., Kisiel, J., & Koke, J. (2007). Natural history museum visitors' understanding of evolution. BioScience, 57, 875–882.

McGaugh, J. L. (2006). Make mild moments memorable: Add a little arousal [Review]. Trends Cognitive Science, 10, 345–347.

Meir, E., Perry, J., Herron, J. C., & Kingsolver, J. (2007). College students' misconceptions about evolutionary trees. The Amer. Biology Teacher, 69, e71–e76.

Miller, J. D., Scott, E. C., & Okamoto, S. (2006). Public acceptance of evolution. Science, 313, 765–766.

Newport, F. (2014). In US, 42% believe creationist view of human origins: Americans' views related to religiousness, age, education. Gallup Poll, May 2014 [data set]. Retrieved from: http://www.gallup.com/poll/170822/believe-creationist-viewhuman-origins.aspx

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. Social and Personality Psychology Compass, 2, 842–860.

Novick, L. R., & Catley, K. M. (2012). Assessing students' understanding of macroevolution: Concerns regarding the validity of the MUM. Science Education, 34, 2679–2703.

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, 138–177.

Novick, L. R., Catley, K. M., & Funk, D. J. (2010). Characters are key: The effect of synapomorphies on cladogram comprehension. Evolution: Education and Outreach, 3, 539–547.

Novick, L. R., Catley, K. M., & Funk, D. J. (2011). Inference is bliss: Using evolutionary relationship to guide categorical inferences. Cognitive Science, 35, 712–743.

Phillips, B. C., Novick, L. R., & Catley, K. M. (2011). Form, function, & habitat: Similarity of responses among college students and children to questions about evolutionary relatedness. Paper presented at the 42nd Annual Meeting of the Jean Piaget Society. Berkeley, CA.

Phillips, B. C., Novick, L. R., Catley, K. M., & Funk, D. J. (2012). Teaching tree thinking to college students: It's not as easy as you think. Evolution: Education & Outreach, 5(4), 595–602.

Poling, D. A., & Evans, E. M. (2004). Religious belief, scientific expertise, and folk ecology. Cognition and Culture: Studies in the Cognitive Anthropology of Science, 4, 485–524.

Prum, A. J. (Producer). (2008). Discovering the great tree of life. Available from the Peabody Museum of Natural History, Yale University, P.O. Box 208118, New Haven, CT. http://archive.peabody.yale.edu/exhibits/treeoflife/film_discovering.html

Roberts, J., Lyons, L., Cafaro, F., & Eydt, R. (2014). Interpreting data from within: Supporting human-data interaction in museum exhibits through perspective taking. In Proceedings Interaction Design and Children (pp. 7–16). ACM Press.

Seoh, K. H. R., Subramaniam, R., & Hoh, Y. K. (2015). How humans evolved according to Grade 12 students in Singapore. Journal of Research in Science Teaching, 53, 291–323. doi: 10.1002/tea.21256.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. Cognitive Science, 21(2), 147–177. doi: 10.1207/s15516709cog2102_2.

Shtulman, A., & Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. Cognitive Science, 32, 1049–1062.

Snibbe, S. S., & Raffle, H. S. (2009). Social immersive media: Pursuing best practices for multi-user interactive camera/projector exhibits. In Proceedings ACM Conference on Human Factors in Computing Systems (pp. 1447–1456). ACM Press.

Spiegel, A. N., Evans, E. M., Frazier, B., Hazel, A., Tare, M., Gram, W., & Diamond, J. (2012). Changing museum visitors' conceptions of evolution. Evolution: Education & Outreach, 5, 43–61.

Tare, M., French, J., Frazier, B., Diamond, J., & Evans, E. M. (2011). Explanatory parent-child conversation predominates at an evolution exhibit. Science Education, 95, 720–744.

Wilensky, U., & Reisman, K. (2006). Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—An embodied modeling approach. Cognition and Instruction, 24, 171–209.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.