The EvoDevoCI: A Concept Inventory for Gauging Students’ Understanding of Evolutionary Developmental Biology

Kathryn E. Perez,* Anna Hiatt,†,‡ Gregory K. Davis,§ Caleb Trujillo,∥ Donald P. French,† Mark Terry,¶ and Rebecca M. Price#

*Department of Biology, University of Wisconsin La Crosse, La Crosse, WI 54601; †Department of Zoology, Oklahoma State University, Stillwater, OK 74074; §Department of Biology, Bryn Mawr College, Bryn Mawr, PA 19010; ¶Department of Biological Sciences, Purdue University, West Lafayette, IN 47907; ‡Northwest School, Seattle, WA 98122; #School of Interdisciplinary Arts and Sciences, University of Washington, Bothell, Bothell, WA 98011

Submitted April 5, 2013; Revised June 1, 2013; Accepted July 1, 2013
Monitoring Editor: Diane Ebert-May

INTRODUCTION

The American Association for the Advancement of Science 2011 report Vision and Change in Undergraduate Biology Education encourages the teaching of developmental biology as an important part of teaching evolution. Recently, however, we found that biology majors often lack the developmental knowledge needed to understand evolutionary developmental biology, or “evo-devo.” To assist in efforts to improve evo-devo instruction among undergraduate biology majors, we designed a concept inventory (CI) for evolutionary developmental biology, the EvoDevoCI. The CI measures student understanding of six core evo-devo concepts using four scenarios and 11 multiple-choice items, all inspired by authentic scientific examples. Distracters were designed to represent the common conceptual difficulties students have with each evo-devo concept. The tool was validated by experts and administered at four institutions to 1191 students during preliminary (n = 652) and final (n = 539) field trials. We used student responses to evaluate the readability, difficulty, discriminability, validity, and reliability of the EvoDevoCI, which included items ranging in difficulty from 0.22–0.55 and in discriminability from 0.19–0.38. Such measures suggest the EvoDevoCI is an effective tool for assessing student understanding of evo-devo concepts and the prevalence of associated common conceptual difficulties among both novice and advanced undergraduate biology majors.

INTRODUCTION

The integrative field of evolutionary developmental biology, or “evo-devo,” enhances our understanding of evolution. Edo-devo builds on the modern synthesis by considering the developmental mechanisms of evolutionary change. For example, evo-devo biologists have identified specific genetic and developmental changes that played key roles in the loss of abdominal appendages in insects and in the modification of hind wings to balancing organs in flies (Ronshaugen et al., 2002; Hersh et al., 2007). By considering ways in which development can influence the evolutionary process, evo-devo also sheds light on evolutionary patterns that cannot be explained solely by natural selection. Examples include the extreme conservation of the number of cervical vertebrae in mammals or the fact that all centipedes have odd numbers of leg-bearing segments (Arthur and Farrow, 1999; Galis, 1999).

The insights of evo-devo present opportunities to enhance student understanding of evolution in general and are thus relevant to all biology instructors. For example, teaching evo-devo has been promoted as an important counter to the creationist claim that evolutionary novelties and other macroevolutionary changes cannot be explained by current evolutionary theory (Gilbert, 2003; Brigandt and Love, 2010). That evo-devo could potentially play this role is supported by the finding that student acceptance of evolution is more

DOI: 10.1187/cbe.13-04-0079
Address correspondence to: Kathryn E. Perez (kperez@uwlax.edu).
‡Present address: Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045
© 2013 K. E. Perez et al. CBE—Life Sciences Education © 2013 The American Society for Cell Biology. This article is distributed by The American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0).
“ASCB®” and “The American Society for Cell Biology®” are registered trademarks of The American Society for Cell Biology.
strongly correlated with instruction in macroevolution than instruction in population genetics or microevolution (Bishop and Anderson, 1990; Sinatra et al., 2003). More widespread teaching of evo-devo is likely with the increasing availability of materials for teaching evo-devo concepts, including explanatory material for teachers (Understanding Evolution, 2012a–c), online virtual labs (Howard Hughes Medical Institute, 2012), case studies (Platt, 2009), and textbooks (Schlichting and Pigliucci, 1998; Carroll et al., 2001; Wilkins, 2001; Arthur, 2011; Stern, 2011; Hall, 2012; Zimmer and Emlen, 2012).

While evo-devo holds promise for enhancing understanding of evolution, it is challenging for students to learn. This is due, in part, to students’ limited exposure to the foundational knowledge required to understand higher-order concepts in evo-devo. For example, many students lack the foundational knowledge of developmental biology, genetics, and molecular biology, with the result that both novice and advanced undergraduates face conceptual difficulties in understanding evo-devo (Hiatt et al., 2013). To address conceptual difficulties, instructors need to be able to efficiently and accurately identify their presence. Compounding these obstacles is the fact that many biology instructors were not exposed to contemporary evo-devo concepts during their own formal educations. In light of these challenges, efforts to integrate evo-devo into undergraduate life sciences curricula will require tools that effectively assess student understanding of evo-devo concepts.

Concept inventories (CIs) are powerful, research-based tools for quickly assessing student understanding of science concepts (Hestenes et al., 1992; Garvin-Doxas et al., 2007; Adams and Wieman, 2010; Smith and Tanner, 2010). Several existing CIs assess conceptual understanding of evolution (Anderson et al., 2002; Baum et al., 2005; Nadelson and Southerland, 2010; Novick and Catley, 2012) but do not assess most evo-devo concepts. Although a few tools are available to evaluate developmental biology concepts, they either include only a few items that touch on evolution (e.g., Knight and Wood, 2005) or have not been validated for use as a general diagnostic assessment tool (e.g., Darland and Carmichael, 2012). Our goal was to develop an easy-to-use CI that assesses student understanding of evo-devo concepts for undergraduate biology majors, the EvoDevoCI. We built upon previous work (Hiatt et al., 2013), in which we identified core concepts in evo-devo that experts agree ought to be taught at the undergraduate level, as well as common conceptual difficulties students encounter when trying to learn these concepts. In this paper, we describe the EvoDevoCI, its validation by experts, and extensive field-testing on undergraduate students at four institutions.

### METHODS

#### Sources of Data

We collected data at four different universities in the United States: a public master’s degree–granting comprehensive university in the Midwest (MCU), a private university in the Northeast (PU), and two large research-intensive public universities in the Midwest (RIM) and the mid-South (RIS). All test administration was deemed exempt by institutional review boards and performed with informed consent (PU, IRB #R11-033; RIM, IRB #1210012864; RIS, IRB #AS125; MCU’s IRB approved, but no IRB number was assigned).

We field-tested multiple-choice items from January 2012 to January 2013 among biology majors (“life sciences” at RIS; Tables 1 and 2). The manner in which students progressed through the curriculum at each institution ensured that each student in our study took the EvoDevoCI only once, unless he or she participated in a test/retest trial, in which case the participant took the CI twice.

#### Development of Scenarios and Items

We designed the EvoDevoCI to assess understanding of the core concepts of evo-devo (Table 3) that we previously identified by surveying evo-devo experts and educators about which concepts are central to the field and taught at the undergraduate level (Hiatt et al., 2013). We wrote several items to target each core evo-devo concept, with each item comprising a question stem and four possible responses: one correct response and three distracters (e.g., Figure 1). Initially,
Table 2. Summary of final validity and reliability testing of the EvoDevoCl

| Assessment          | Curriculum level of students | Number of responses | Response rate |
|---------------------|-----------------------------|---------------------|---------------|
| Final administration| Novice (includes EvoDevoCI) | 441                 | 20%           |
|                     | Advanced                    | 98                  | 63%           |
|                     | Total                       | 539                 | 24%           |
| Test/retest         | Novice                      | 34                  | 77%           |
|                     | Advanced                    | 48                  | 92%           |
|                     | Total                       | 82                  | 85%           |
| Redacted            | Novice (EvoDevoCI)          | 104                 | 92%           |

aShown are the number of students who took the CI, as well as response rates for novice (<5 biology courses) and advanced (≥5 biology courses) students. All tests were administered online and in a fixed order (Crayfish/Centipede/Minnow/Lizards), except for EvoDevoCI, which was administered in paper form, and EvoDevoCI, for which the scenario order was randomized. Only students identified as biology majors were included in the final analyses and listed here.

we used biological examples from plants, invertebrates, and vertebrates to develop scenarios and stems. Although the scenarios were inspired by actual biological examples, gene names were fictionalized to avoid triggering rote recall of common classroom examples, and organisms were sometimes changed to species that are familiar to students. Following Nehm and Ha (2011), we included examples from both sides of the following dichotomies: within versus between species differences, gains versus losses of traits, animals versus plants, and familiar versus unfamiliar taxa.

When writing and revising items, we weighed the plausibility, language, consistency, and breadth of each response to remove any temptation for students to guess or otherwise select a response without considering the associated scenario or stem. The distracters for each item represent conceptual difficulties we previously identified as common in student answers to questions targeting particular core concepts (Hiatt et al., 2013; Table 4). To illustrate the relationships between the core concepts and the associated conceptual difficulties used to write distracters, we constructed a diagram (Figure 2; descriptions of concepts and conceptual difficulties can be found in Tables 3 and 4) using the network diagram visualization function in Many Eyes (IBM, 2010). For the conceptual difficulty “Lack of development” (see Table 4, DV1), which includes an exclusive reliance on natural selection, we found it difficult to write distracters that were clearly implausible. Thus, for some of these distracters, we took the approach of making them implausible in some other

Table 3. Core concepts in evo-devo from Hiatt et al. (2013) included in this instrument

| Core concepts in evo-devo | Item code | Scenario code |
|---------------------------|-----------|---------------|
| CC1. A small number of mutations can make a large evolutionary difference: It is possible for novel phenotypes that evolve as the result of the fixation of a small number of mutations that cause significant changes in the regulation of developmental processes. This does not preclude the possibility that many (or even most) differences between species require a large number of small-effect mutations. | Q4 | M1 |
| CC2. Evolution can occur by changes in regulation: Given that developmental processes are often shared, novel phenotypes often evolve via changes in regulation (e.g., cooption or deployment of genetic regulatory networks to different tissues or stages of development). | Q5, Q6 | M2, M3 |
| CC3. Mutations that are less pleiotropic are more likely to contribute to evolution: Mutations that are less pleiotropic (e.g., mutations in a gene or gene product that plays only a limited role in development, in a modular cis-regulatory element, or in a modular domain of a protein) are less likely to have deleterious pleiotropic effects on fitness and thus are more likely to become fixed in populations. | Q2, Q7 | C2, M4 |
| CC4. Development can bias the direction of evolutionary change: Developmental processes can bias evolutionary outcomes by either limiting the variation available to natural selection or attaching deleterious pleiotropic effects to certain variants. | Q1, Q3 | C1, N1 |
| CC5. Developmental plasticity can evolve: The environment can select among heritable variation in a developmental response to a particular environmental change, resulting in adaptive developmental plasticity. | Q8, Q10 | L1, L3 |
| CC6. Developmental variation is part of the raw material of natural selection: Many adaptations are the result of the environment selecting among heritable variation in phenotype that is the result of heritable variation in developmental processes, which is itself the result of genetic variation. | Q9, Q11 | L2, L4 |

aItems targeting each concept are indicated by either their position in the CI (question code) or their position within each scenario: Crayfish (C1, C2), Centipedes (N1), Minnows (M1, M2, M3, M4), and Lizards (L1, L2, L3, L4).

bWe intend “developmental process” to refer to any process that is part of the development of a sexually mature adult.

While we recognize that features of development (e.g., gene expression patterns) are often considered to be part of an organism’s phenotype, for purposes of clarity we use “phenotype” here to refer only to traits (e.g., behavioral, morphological, physiological, biochemical) of the adult organism.
Centipede species vary in the number of leg-bearing segments, from as few as 5 to as many as 125, but all centipedes possess an odd number of leg-bearing segments, and thus an odd number of pairs of legs. By manipulating leg-bearing segment number, it has been determined that there is no difference in survival and reproductive success between individuals with even and odd numbers of pairs of legs.

This centipede has an odd number of leg-bearing segments and thus an odd number of leg pairs.

If there is no difference in reproductive success between individuals having an even versus an odd number of leg pairs, why don’t any centipedes have even numbers of leg pairs? Of the following, choose the best hypothesis.

A. Centipedes do not need even numbers of leg pairs if odd pairs are sufficient for survival; therefore they choose to have an odd number of leg pairs.
B. Centipedes with even numbers of leg pairs do not occur because of the way segments are added during development.
C. Centipedes do not have the gene that causes an even number of leg pairs to form during development.
D. Centipedes with an even number of leg pairs are more likely to be eaten by a predator.

Figure 1. CI item targeting CC4: “Development can bias evolutionary change.” Each of our questions followed a similar format, with a short scenario inspired by an actual biological example but for which some details (e.g., gene names, the organism, or experiments performed) may have been altered, followed by a question stem and four response options. Each distracter (incorrect response) is written to reflect one of the conceptual difficulties most often associated with that evo-devo concept. Although the biological example in this case is real, the fitness data have been imagined to suggest an explanation that does not rely on natural selection.

way. For example, the distracter might contradict additional information given in the scenario or stem (e.g., question 3, response D, as shown in Figure 1; see Supplemental Material for additional examples).

We evaluated the quality of distracters included in the final EvoDevoCI by administering an unaltered version ($n = 54$) and a version consisting only of item responses ($n = 50$), with the scenario and question stems redacted, to students at RIS. In the redacted version, students were asked to select the correct response for each item and provide an explanation for their choice. Both groups of students were in the same lecture course and should represent similar student populations. This test served two purposes: 1) to determine whether either the correct response or distracters for each item were detectable by students on the basis of “clues,” such as the length of the response or the inclusion of absolutes in the case of distracters (e.g., “never” or “all”) (Novick and Catley, 2012); and 2) to determine whether any responses were perceived by students to be inherently more or less plausible relative to one another. If item responses possess neither clues nor differences in inherent plausibility, then the percentage of students choosing each response should approximate the probability of selecting a response by chance ($0.25$). We used a Mann-Whitney $U$-test to compare response rates for the unaltered and redacted versions of the CI against a random distribution.

**Initial Assessment and Revision**

Field tests 1–3 evaluated how novice and advanced students respond to different scenarios, how often they select particular responses, and how they respond to the wording and phrasing of preliminary items (overall response rate = 31.3%). For field tests 1 and 2, we asked participants to circle or describe unfamiliar vocabulary (Patton, 2002). Using the data from each field test (1–3), we calculated the difficulty and discriminability of each item. Difficulty ($P$) was calculated as the overall proportion of students choosing the correct response for a particular item, while discriminability ($D$) was calculated by subtracting an item’s difficulty among low-performing (bottom half, $P_L$) students from an item’s difficulty among high-performing (upper half, $P_U$) students ($D = P_U - P_L$; Crocker and Algina, 1986). Ideally, all items have a difficulty greater than chance ($P > 0.25$ for items with four responses) and discriminability greater than $0.20$ ($D > 0.20$), indicating the item successfully discriminates between high- and low-performing students (Crocker and Algina, 1986; Haladyna, 2004). Furthermore, all distracters should be chosen by at least a few respondents ($> 0.05$) (Crocker and Algina, 1986; Haladyna, 2004). Easy items (e.g., $P > 0.8$) are undesirable, in that they are both less likely to discriminate between high- and low-performing students and less likely to detect conceptual difficulties associated with the target concept (Haladyna, 2004). Thus, items not meeting these measures of readability, difficulty, discriminability, and minimal selection of all distracters were eliminated or revised.

For field test 1, 34 preliminary items were divided into four subsets of 7–10 items each and administered (502 question sets taken by 421 students; some students at MCU took two question sets). As a follow-up, we asked select students to explain their answers to confirm that chosen responses truly reflected the understanding of the student (Patton, 2002). Based on these criteria, the 34 items were narrowed to 17, which together addressed each core concept two or three times. For
### Table 4. Conceptual difficulties used as the basis for distracters and the prevalence of the associated distracters in the final validation test

| Conceptual difficulties | Target concept | % Choosing this: CD novice | % Choosing this: CD advanced |
|-------------------------|----------------|-----------------------------|-----------------------------|
| **Common biological (CB)** |                |                             |                             |
| CB1 Teleology           | Attributing design and purpose to organism, environment, process, or mechanism. Responses that exhibit this difficulty include references to purpose or design. | CC4 11.8 | 6.5 |
| CB2 Vocabulary          | Misusing terms (e.g., confusing gene, allele, and genome). | CC6 44.8 | 31.8 |
| CB3 Anthropomorphism    | Attributing human qualities to nonhuman organisms, environments, processes, or mechanisms. | CC5 5.5 | 8.4 |
| **Developmental (DV), including cell and molecular biological aspects** |                |                             |                             |
| DV1 Lack of development | Failing to reference development, even when prompted. Includes invoking natural selection as a mechanism in place of more appropriate evo-devo mechanisms. | CC2 40.75* | 32.2* |
| DV2 A single gene affects a single trait | Stating explicitly or implying that each trait is determined by a single gene or that each gene determines only one trait. | CC2 17.1 | 12.1 |
| DV4 HOX genes are the only regulatory genes | Stating explicitly or implying that HOX genes are the only regulatory genes. | CC2 7.9 | 10.3 |
| **Evolutionary (EV)** |                |                             |                             |
| EV1 Characteristics that are not used are lost | Implying that characteristics that are not used by the organism are lost simply because they are not used and not because of the loss of maintenance selection. | CC4 10.4 | 6.5 |
| EV2 Inheritance of acquired traits | Implying that evolution proceeds by the inheritance of acquired characteristics. Among the latter we do not include potentially legitimate examples such as the genetic assimilation of induced phenotypes or the assimilation of learned behaviors, as in the Baldwin effect. | CC1 22.6 | 11.2 |
| EV3 Lack of selection results in stasis | Stating explicitly or implying that evolutionary stasis occurs only when selection (either stabilizing or positive) does not occur. | CC2 23.8 | 27.1 |
| EV4 Lack of understanding of population-level processes | Demonstrating a lack of understanding of population-level processes. For example, attributing evolutionary adaptation, the population-level process, to an individual. | CC3 31.9 | 26.2 |
| EV5 Exclusive gradualism | Stating explicitly or implying that all changes in the phenotype must evolve gradually. | CC1 40.4 | 40.2 |
| EV9 Selection acts on genes, not the phenotype | Stating explicitly or implying that selection acts on genes, independent of the phenotype. | CC3 26.3 | 20.6 |
| **Evo-devo (ED)** |                |                             |                             |
| ED1 Changes in gene expression result only from mutations in said gene | Stating explicitly or implying that a change in a gene’s expression must be due to a mutation in the cis-regulatory enhancers of that gene; not recognizing the potential for mutations in upstream regulators (in trans) to alter expression. | CC2 24.0 | 21.5 |
| ED2 Gene expression evolves only when genes appear or disappear | Stating explicitly or implying that gene expression evolves only because a gene appears or disappears in the genome. | CC1 13.9 | 17.8 |
| ED3 Phenotypic change can only result from a gene appearing or disappearing | Stating explicitly or implying that phenotypic change only occurs when genes appear or disappear in the genome. | CC5 49.2 | 48.6 |
| ED4 Only closely related species have conserved traits | Stating explicitly or implying that only closely related species can have conserved genes, proteins, or developmental processes. | CC3 24.4* | 22.45* |

*For each item, we used the conceptual difficulties most commonly associated with the targeted core concept. Background information on the prevalence and descriptions of the conceptual difficulties are given in Hiatt et al. (2013). Most conceptual difficulties were used in only one of the items targeting any particular concept; a few were used in both items, and in these cases, mean prevalence is shown and marked with an asterisk.
field test 2, we divided these 17 items into two subsets of 8 and 9 items for administration (n = 63). Again, items not meeting minimal criteria were eliminated or revised, resulting in 11 items in field test 3 (n = 168). A few minor changes were made to this version prior to expert assessment.

**Final Validity**

The final EvoDevoCI (see Supplemental Material) comprises four scenarios and 11 items—Crayfish (C1, C2), Centipedes (N1), Minnows (M1, M2, M3, M4), and Lizards (L1, L2, L3, L4)—which together target six core evo-devo concepts (see Table 3). Before the final 11-item CI was administered to students, it was subdivided into three surveys and subjected to scrutiny by experts. We solicited the opinions of experts who both held a PhD and were actively publishing in the field of evo-devo (defined as socially determined experts and task-oriented experts, respectively, by Ericsson et al., 2006). Experts were asked to review scenarios, items, and all responses according to their expertise: Crayfish/Centipede (C1, C2, N1; n = 3), Minnow (M1, M2, M3, M4; n = 2), and Lizards (L1, L2, L3, L4; n = 4). In particular, we asked experts whether items were plausible, accurate, and clear and whether the item addressed the target concept (Table 5). Experts also had the option to add commentary to each question and provide feedback on the targeted core concepts. The percent agreement for each criterion was calculated across individual items and combined for each scenario (Table 5). We then made final revisions designed to address problems or concerns raised by experts and administered the 11-question EvoDevoCI to students (n = 539) at MCU via Desire2Learn (Desire2Learn, Kitchener, Ontario, Canada) and at PU, RIM, and RIS via Qualtrics (Qualtrics Labs, Provo, UT). The CI and a key that identifies correct answers and describes distracters are provided in the Supplemental Material.

To determine whether the EvoDevoCI measures the intended construct among biology majors, we grouped results by the number of biology courses taken (0, 1–2, 3–4, 5–10, and more than 10). Cronbach’s alpha coefficient (Cronbach, 1951) was calculated for each group using IBM SPSS Statistics (IBM Corp., Armonk, NY). For all other analyses, the students were categorized as “novice,” having fewer than five biology courses, or “advanced,” having five or more biology courses. As none of the institutions surveyed had a fixed course sequence, we did not know precisely the exposure each student had to evo-devo concepts. Categorizing them into novice and advanced, rather than course sequence or class standing, was thus a useful, but rough categorization. The point biserial correlation coefficient was also calculated for individual test items (Anderson et al., 2002; Smith et al., 2008).

On the suggestion of reviewers, two small changes were made to the CI postvalidation: the figure caption for the Minnow scenario was edited to more accurately describe the image provided and, in question 6, one word in a distracter (response D; see Supplemental Material) was changed to ensure the response directly contradicted the stem.

**Final Reliability**

Following methods used to develop similar concept inventories (Anderson et al., 2002; Rutledge and Sadler, 2007), we conducted a test/retest trial to test the reliability of items (Table 2). In this trial, we administered the 11-question EvoDevoCI to both novice (0–4 biology courses) and advanced (≥5 biology courses) students at RIS. We administered the same test 3 wk later, giving no evo-devo instruction between test administrations. To test for any effects of question order or test format (paper vs. online), we selected a portion of students from a single course to take a paper version of the survey (n = 27), while the remaining students took an online survey with randomized question order (n = 25). These data were normally distributed and unskewed. We were thus able to use t tests to compare scores on online versus paper and random-order versus unaltered-order versions of the CI.

**RESULTS**

**Initial Field Tests and Revisions**

We subjected early versions of the EvoDevoCI to three separate field tests involving 733 undergraduate students at four institutions. After each field test, scenarios and items in the CI were eliminated or revised in order to increase readability and to ensure appropriate difficulty and discriminability of all items and optimal selection of all distracters (steps 4–8, Table 1; see Methods for more details). This process yielded
Table 5. Experts were given a subset of questions on concepts most aligned with their expertise with each expert evaluating 3–4 questions.

| Expert survey question                                                                 | Crayfish/Centipede (3 questions; n = 3) | Minnows (4 questions; n = 2) | Lizards (4 questions; n = 4) | Overall |
|----------------------------------------------------------------------------------------|----------------------------------------|------------------------------|------------------------------|---------|
| 1. Is the scenario plausible?                                                           | 89                                     | 50                           | 94                           | 86.5    |
| 2. Does this question address the target concept?                                       | 78                                     | 75                           | 81                           | 78      |
| 3. Is the question clear?                                                                | 67                                     | 75                           | 75                           | 73      |
| 4. Is the correct answer accurate given the scenario?                                    | 67                                     | 50                           | 75                           | 66      |
| 5. Do any of the other answers strike you as correct?                                     | 11                                     | 13                           | 31                           | 19      |

\( \text{n} \) indicates the number of experts that evaluated each scenario.

Table 6. Final statistical analysis of the EvoDevoCI

| Analyses                                       | n\(^a\) | Result                  | Significance |
|------------------------------------------------|----------|-------------------------|--------------|
| Item analysis                                  |          |                         |              |
| Item difficulty\(^b\)                          |          | \( P = \text{Percent correct} \) |              |
| Item discriminability\(^b\)                    |          | \( D = \text{Discriminability} \) |              |
| Redacted form                                  |          | \( P = \text{Percent correct} \) |              |
| Unaltered form                                 |          | \( P = \text{Percent correct} \) |              |
| Validity                                       |          |                         |              |
| Internal                                       |          | \( \alpha = 0.31-0.73 \) |              |
| Cronbach's alpha\(^c\)                        | 539      | \( R = 0.27-0.52 \)      |              |
| Reliability                                    |          |                         |              |
| Test-retest                                    |          | \( t = 0.960 \)          | \( p < 0.01 \) |
| Online vs. paper                               |          | \( t = 0.63 \)           | \( p < 0.05 \) |
| Random vs. fixed                               |          | \( t = 0.79 \)           | \( p < 0.05 \) |

\( \text{n} \) indicates the number of students who took the EvoDevoCI.

\( \text{P} \) indicates the number of experts that evaluated each scenario.

\( \text{P} \) is calculated for each item and reported as a range.

\( \alpha \) is reported for novice students (0 biology courses; \( \alpha = 0.31 \)) as well as highly advanced students (10+ biology courses; \( \alpha = 0.73 \)).
To ensure that student responses relied on the information provided in the scenarios and item stems rather than on any inherent plausibility of correct responses versus distractors or other clues among item responses, we field-tested a redacted version of the CI that included only item responses without scenarios or stems. The mean proportion of students who chose the correct answer across all items on the redacted version was close to chance (P = 0.26) and not significantly different from a random distribution (p > 0.05, Mann-Whitney U-test; Table 6). In contrast, the mean proportion of students who chose the correct answer on the unaltered version (P = 0.34) was significantly different from a random distribution (p < 0.001, Mann-Whitney U-test; Table 6). For example, items targeting the concept “Mutations that are less pleiotropic are more likely to contribute to evolution” (CC3) showed a 20–23% increase in the correct response rate in the unaltered as compared with the redacted version. These results indicate that the scenarios and question stems are essential for gathering responses that indicate specific conceptual difficulties with particular concepts, for eliciting evo-devo thinking, and for ensuring item responses do not provide clues or hints that would result in students choosing the correct response more often by chance.

Validity

To validate the CI, we solicited feedback from a group of evo-devo experts. To a reasonably high degree, experts agreed on the overall plausibility (86.5%), clarity (73.0%), and accuracy (66%) of the scenarios and items (Table 5). In addition, 78.1% agreed that, overall, scenarios and items did indeed target the intended concept, while 81.1% agreed that distracters could not be considered correct. In almost all cases of disagreement, experts provided explanations, and these open-ended comments were scrutinized and used for revision. Revisions included changing words and phrases that experts found problematic; rewording some distracters so as to make them clearly implausible given the scenario and stem; and targeting the concepts “A small number of mutations can make a large evolutionary difference” (CC1) and “Evolution can occur by changes in regulation” (CC2) with three items within the same scenario (Minnows) in response to feedback that these two concepts are intertwined.

Internal consistency measured by Cronbach’s alpha ranged from 0.31 to 0.73 (Table 6). Students having taken a greater number of biology courses (≥5) showed higher inter-item consistency (α = 0.73) than students having taken only one or two courses (α = 0.41). Point biserial correlation values (Rpb) for items in the CI (range 0.27–0.52) all exceeded the value of 0.20 recommended by Thorndike (1997).

Reliability

Student scores from test and retest administrations at RIS (n = 81) show consistent performance among the testing instances (R = 0.96, p < 0.01) when receiving no specific evo-devo instruction. Alternate forms of the 11-item EvoDevoCI also show no significant difference in ability to measure evo-devo understanding: the fixed-order EvoDevoCI version (C1, C2, N1, M1, M2, M3, M4, L1, L2, L3, L4; τ number correct = 3.16) given to students does not significantly vary compared with a randomized version (τ number correct = 2.88) given to students (τ(50) = 0.70, p = 0.49, two-tailed). Additionally, the fixed paper version did not significantly vary from a fixed online version (τ number correct = 2.92 and 3.16, respectively; τ(88) = 0.63, p = 0.53, two-tailed).

DISCUSSION

The results of the final field test indicate the EvoDevoCI is a valid and reliable measure of student understanding of evo-devo, with test items that reference plausible biological scenarios validated by evo-devo experts. Cronbach’s alpha is low for undergraduate students having taken few biology courses, and the tool has increased reliability among students who have taken five or more biology courses. The difficulty range for items is high (0.22–0.55; lower numbers indicate more difficult items), and the overall difficulty of the CI is not significantly different for students having taken less than five biology courses, as compared with students having taken five or more biology courses (Figure 3). This latter result is disappointing, but expected, based on our previous work indicating that both novice and advanced students lack the foundational content knowledge needed to answer evo-devo questions correctly (Hiatt et al., 2013). As evo-devo instruction gains a larger foothold in mainstream biology courses, we expect student performance on the EvoDevoCI to improve. We are currently undertaking additional research to examine learning gains after specific instruction, asking whether evo-devo understanding improves when a student possesses developmental biology or other foundational knowledge.

Student Reasoning and Item Context

Nehm and Ha (2011) have identified a number of contexts that affect how students reason about evolutionary situations: plants versus animals, familiar versus unfamiliar species, gain versus loss of traits, and evolution within versus among species. These dichotomies are notable, because, while they do not usually affect how experts interpret questions, students often view the opposing contexts as fundamentally different. In the EvoDevoCI, all of the scenarios referenced animals familiar to biology majors. With regard to gains versus
losses, most items in the CI reference character-state changes that are neither straightforward gains nor straightforward losses. With regard to within- versus between-species differences, the CI contains a mix of both. In a case in which a shared target concept allowed for comparison, we found the pattern observed by Nehm and Ha (2011), namely, that items referencing between-species differences are more difficult, particularly for novices. By either controlling for contexts that provoke differences in students’ reasoning (i.e., for animals vs. plants and gains vs. losses) or including a mix of both sides of the dichotomy (i.e., for within- versus between-species differences), we have attempted to minimize unwanted variation in student reasoning, while examining a diversity of potential contexts when possible.

**Prevalence of Evo-Devo Conceptual Difficulties**

Our previous research revealed that students often fail to master evo-devo concepts because they lack foundational concepts from developmental biology, genetics, and molecular biology (Hiatt et al., 2013). Because distractors for any particular item in the CI were based on conceptual difficulties empirically associated with the concept targeted by the item (Hiatt et al., 2013), more broadly associated conceptual difficulties have greater representation among distractors. The conceptual difficulty with the greatest representation, associated with four concepts, is “Lack of development” (DV1), followed by “Gene expression evolves only when genes appear or disappear” (ED2), which is associated with three concepts (Figure 2). In contrast, three of the conceptual difficulties associated with the concept “Mutations that are less pleiotropic (Figure 2). In contrast, three of the conceptual difficulties associated with the concept “Mutations that are less pleiotropic or disappear” (ED2), which is associated with three concepts (Figure 2). In contrast, three of the conceptual difficulties associated with the concept “Mutations that are less pleiotropic or disappear” (ED2), which is associated with three concepts (Figure 2). In contrast, three of the conceptual difficulties associated with the concept “Mutations that are less pleiotropic are more likely to contribute to evolution” (CC3) are exclusively associated with that concept and thus are represented less among distractors.

Although advanced students did not perform significantly better on the EvoDevoCI than novice students, generally speaking, advanced students did choose specific distractors/conceptual difficulties at lower frequencies than did novice students (Table 4), in some cases, much lower (8.55% lower for DV1 in an item targeting CC2; 11.4% lower for EV2 in a CC1 item; and 8.95% lower for ED2 in a CC4 item). This trend is expected if indeed students overcome the conceptual difficulties associated with evo-devo concepts as they progress from novice to advanced. Exceptions to this trend identify conceptual difficulties for which current modes of instruction have either no effect or a reverse effect. For example, the percentage of students choosing the distractor “Lack of development” (DV1) in an item targeting CC5 did not change much. In the cases of “Changes in gene expression result only from mutations in said gene” (ED1) in a CC4 item and “Inheritance of acquired traits” (EV2) in a CC5 item, advanced students chose the distractors 6.1% and 4.4% more frequently than novice students, respectively.

The fact that some conceptual difficulties in understanding evolution are encountered only or more commonly among advanced students has been reported (Andrews et al., 2012). In these cases, it could be that some conceptual difficulties actually require more knowledge and are not encountered until students have some exposure to developmental biology or evo-devo. An expert is able to apply a subset of his or her knowledge to particular problems with less effort than a novice (Bransford et al., 2000). In our study, however, while advanced students likely hold a larger repertoire of evo-devo content knowledge than novice students, they still seem to lack the ability to apply this knowledge to particular problems and instead may incorrectly associate more sophisticated concepts or supply factually correct but unlikely solutions. A caveat here is that our categories of “novice” and “advanced” are based merely on the number of biology courses taken and likely include students that have had an array of different course experiences. More precise data on prior concept exposure would be useful for any future studies of students’ conceptual difficulties with evo-devo.

**Limitations**

An instrument such as the EvoDevoCI has intrinsic limitations. For one, our goal of a short instrument that takes little class time required that our assessment be based on relatively few multiple-choice questions targeting only the most essential core concepts. This necessarily limited the breadth of the instrument, precluding the inclusion of more sophisticated evo-devo concepts that are nonetheless arguably of great evolutionary importance. These included canalization, genetic assimilation and accommodation, gene–environment interactions, epigenetic modification of DNA, gene duplication and genome evolution, serial homology, modularity, facilitated evolution, and the evolution of multicellularity. Supplementing the EvoDevoCI with 1) questions on additional topics, 2) reasoning contexts, and 3) two-tiered (Treagust and Haslam, 1986) or open-ended questions (Nehm and Schonfeld, 2008) ought to increase breadth when assessing student understanding of evo-devo.

The utility of this particular tool lies in its ability to assess understanding of a range of evo-devo concepts, all considered vital for undergraduate biology majors, rather than exhaustively assessing a single knowledge construct. While, in theory, a maximally reliable CI would examine only a single knowledge construct, the EvoDevoCI includes items examining five distinct evo-devo concepts, all of which are interdisciplinary in nature. This predictably results in a lower Cronbach’s alpha value, which is typical of similar CIs, such as the Genetics Concept Assessment (Smith et al., 2008). The construction of a CI requires balancing the reliability of the instrument to capture student understanding on the one hand with practicality and usability on the other (Adams and Wieman, 2010).

Finally, in constructing the EvoDevoCI, we have no desire to canonize any part of evolutionary developmental biology. Instead, we recognize that, as our scientific understanding of evo-devo improves, our inventory of evo-devo concepts and attendant conceptual difficulties, along with the tool we designed to assess them, must also change. Our hope is that future tools designed to assess student knowledge of evo-devo will benefit from and build upon the EvoDevoCI.

**Uses for the EvoDevoCI**

The EvoDevoCI is a diagnostic test designed to assess conceptual understanding of a set of core concepts in evo-devo among undergraduate biology majors. Given that the CI has been validated with a geographically and institutionally diverse student population, ranging from freshmen to seniors, the tool has different potential applications.
At RIS, faculty members currently use the EvoDevoCI pre- and postinstruction to assess the knowledge students gain from an evo-devo unit taught in upper-level courses in evolution and embryology and lower-level courses in animal biology. In these applications, the CI is taken online with a 2- or 3-wk interval between pre- and postadministrations. Similarly, faculty members at MCU have administered the CI during the first and last weeks of courses in organismal biology to assess newly implemented evo-devo instruction in these courses.

Our hope is that the EvoDevoCI can be used to complement the growing number of diagnostic instruments, allowing instructors to capture a more complete snapshot of student understanding of evolution. As per the recommendations of Vision and Change (Bauerle et al., 2011), the EvoDevoCI marries disciplines and focuses on assessing concepts. Because of the exclusive focus on concepts, however, we advise using the EvoDevoCI in conjunction with assessments designed to assess competencies. In this way, student knowledge of both evo-devo concepts and the practices used to arrive at them can be fully assessed.

ACKNOWLEDGMENTS

We thank the National Evolutionary Synthesis Center (NESCent), which funded the EvoCI Toolkit working group, making this work possible, and other members of the EvoCI toolkit working group. We thank the following faculty who surveyed their biology courses to gather student-response data: Anita Baines, Tim Gerber, Rick Gillis, Gretchen Gerrish, Jennifer Miskowski, David Bos, David Eichinger, Michi Tobler, Jason Belden, Arpad Nyari, Andy Dzialowski, Stephen Gardiner, Joy Little, and Mary Towner. We thank the students who participated in this study. We also thank our panel of expert reviewers, which included, among others, Alexa Bely, W. Anthony Frankino, Brian Hall, Dayalan Srinivasan, David L. Stern, and Matt Wund. Their contribution should not be seen as an endorsement of either this article or the EvoDevoCI itself. The centipede image was used with permission from www.gutenberg.org; crayfish images were modified from an image in the Wisconsin Water Resources Clip Art Collection; other images were drawn by Caleb Trujillo. G.K.D. is supported by the Elizabeth B. Jackson Fund at Bryn Mawr College and award IOS-1051643 from the National Science Foundation (NSF). NESCent is the Elizabeth B. Jackson Fund at Bryn Mawr College and award IOS-1051643 from the National Science Foundation (NSF). NESCent is the Elizabeth B. Jackson Fund at Bryn Mawr College and award IOS-1051643 from the National Science Foundation (NSF). Our hope is that the EvoDevoCI can be used to complement the growing number of diagnostic instruments, allowing instructors to capture a more complete snapshot of student understanding of evolution. As per the recommendations of Vision and Change (Bauerle et al., 2011), the EvoDevoCI marries disciplines and focuses on assessing concepts. Because of the exclusive focus on concepts, however, we advise using the EvoDevoCI in conjunction with assessments designed to assess competencies. In this way, student knowledge of both evo-devo concepts and the practices used to arrive at them can be fully assessed.

REFERENCES

Adams WK, Wieman CE (2010). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ 33, 1289–1312.

Anderson DL, Fisher KM, Norman CJ (2002). Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach 39, 952–978.

Andrews TM, Price RM, Mead LS, McElhinny TL, Thanukos A, Perez KE, Herreid CF, Terry DR, Lemons PP (2012). Undergraduate biology students’ misconceptions about genetic drift. CBE Life Sci Educ 11, 248–259.

Arthur W (2011). Evolution: A Developmental Approach, West Sussex, UK: Wiley-Blackwell.

Arthur W, Farrow M (1999). The pattern of variation in centipede segment number as an example of developmental constraint in evolution. J Theor Biol 200, 183–191.

Bauerle C et al. (2011). Vision and Change in Undergraduate Biology Education. A Call to Action, Washington, DC: American Association for the Advancement of Science.

Baum DA, Smith SD, Donovan SSS (2005). Evolution: the tree-thinking challenge. Science 310, 979–980.

Bishop BA, Anderson CW (1990). Student conceptions of natural selection and its role in evolution. J Res Sci Teach 27, 415–427.

Brunsford JD, Brown AL, Cocking RR (2000). How People Learn: Brain, Mind, Experience, and School, Washington, DC: National Academies Press.

Brigandt I, Love A (2010). Evolutionary novelty and the evo-devo synthesis: field notes. Evol Biol 37, 93–99.

Carroll SB, Grenier JK, Weatherbee SD (2001). From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, Malden, MA: Blackwell Scientific.

Crocker L, Algina J (1986). Introduction to Classical and Modern Test Theory, Orlando, FL: Holt, Rinehart and Winston.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Darland DC, Carmichael JS (2012). Long-term retention of knowledge and critical thinking skills in developmental biology. J Microbiol Biol Educ 13, 125–132.

Ericsson KA, Charness N, Feltovich PJ, Hoffman RR (2006). The Cambridge Handbook of Expertise and Expert Performance, New York: Cambridge University Press.

Galis F (1999). Do we already all mammals have seven cervical vertebrae? Developmental constraints, Hox genes, and cancer. J Exp Zool 285, 19–26.

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.

Gilbert SF (2003). Opening Darwin’s black box: teaching evolution through developmental genetics. Nat Rev Genet 4, 735–741.

Haladyna TM (2004). Developing and Validating Multiple-Choice Test Items, Mahwah, NJ: Erlbaum.

Hall B (2012). Evolutionary developmental biology (Evo-Devo): past, present, and future. Evol Educ Outreach 5, 184–193.

Hersh BM, Nelson CE, Stoll SJ, Norton JE, Albert TJ, Carroll SB (2007). The UBX-regulated network in the haltere imaginal disc of D. melanogaster. Dev Biol 302, 717–727.

Hestenes D, Wells M, Swackhamer G (1992). Force Concept Inventory. Phys Teach 30, 141–157.

Hiatt A, Davis GK, Trujillo C, Terry M, French DP, Price RM, Perez KE (2013). Getting to evo-devo: concepts and challenges for students learning evolutionary developmental biology. CBE Life Sci Educ 12, 494–508.

Howard Hughes Medical Institute (2012). The Virtual Sticklegene Evolution Lab. www.hhmi.org/biointeractive/vlabs/stickleback/index.html (accessed 1 February 2013).

IBM (2010). Many Eyes. www-958.ibm.com/software/analytics/manyeyes/page/About.html (accessed 1 February 2012).

Knight JK, Wood WB (2005). Teaching more by lecturing less. Cell Biol Educ 4, 298–310.

Nadelson LS, Southerland SA (2010). Development and preliminary evaluation of the measure of understanding of macroevolution: introducing the MUM. J Exp Educ 78, 151–190.

Nehm RH, Ha M (2011). Item feature effects in evolution assessment. J Res Sci Teach 48, 237–256.
Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach 45, 1131–1160.

Novick LR, Catley KM (2012). Assessing students’ understanding of macroevolution: concerns regarding the validity of the MUM. Int J Sci Educ 34, 2679–2703.

Patton MQ (2002). Qualitative Evaluation and Research Methods, Newbury Park, CA: Sage.

Platt JE (2009). Case of the Three-Spined Stickleback, Evolution & the Nature of Science Institutes. www.indiana.edu/~ensiweb/lessons/stickleback.html (accessed 27 March 2012).

Ronshaugen M, McGinnis N, McGinnis W (2002). Hox protein mutation and macroevolution of the insect body plan. Nature 415, 914–917.

Rutledge ML, Sadler KC (2007). Reliability of the measure of acceptance of the theory of evolution (MATE) instrument with university students. Am Biol Teach 69, 332–335.

Schlichting CD, Pigliucci M (1998). Phenotypic Evolution: A Reaction Norm Perspective, Sunderland, MA: Sinauer.

Sinatra GM, Southerland SA, McConaughy F, Demastes JW (2003). Intentions and beliefs in students’ understanding and acceptance of biological evolution. J Res Sci Teach 40, 510–528.

Smith JI, Tanner K (2010). The problem of revealing how students think: concept inventories and beyond. CBE Life Sci Educ 9, 1–5.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ 7, 422–430.

Stern D (2011). Evolution, Development, and the Predictable Genome, Greenwood Village, CO: Roberts.

Thorndike RM (1997). Measurement and evaluation in psychology and education, Upper Saddle River, NJ: Merrill/Prentice-Hall.

Treagust DF, Haslam F (1986). Evaluating secondary student’s misconceptions of photosynthesis and respiration in plants using a two-tier diagnostic instrument, San Francisco, CA. Paper presented at the 59th Annual Meeting of the National Association for Research in Science Teaching.

Understanding Evolution (2012a). Bringing Homologies into Focus. http://evolution.berkeley.edu/evolibrary/search/lessonsummary.php?thisaudience=9-12&resource_id=203 (accessed 27 March 2012).

Understanding Evolution (2012b). Eye Evolution. http://evolution.berkeley.edu/evolibrary/eye_evolution.pdf (accessed 27 March 2012).

Understanding Evolution (2012c). Why the Eye? http://evolution.berkeley.edu/evolibrary/article/1_0_0/eyes_01 (accessed 27 March 2012).

Wilkins AS (2001). The Evolution of Developmental Pathways, Sunderland, MA: Sinauer.

Zimmer C, Emlen D (2012). Evolution: Making Sense of Life, Greenwood Village, CO: Roberts.