Biological Variation as a Threshold Concept: Can We Measure Threshold Crossing?

Elise Walck-Shannon,[†] Janet Batzli,[‡*] Josh Pultorak,[§] and Hailey Boehmer[‡]

[†]Center for Integrative Research on Cognition, Learning, and Education (CIRCLE) and Department of Biology, Washington University in St. Louis, St. Louis, MO 63130; [‡]Biology Core Curriculum (Biocore) and [§]Wisconsin Institute for Discovery and Department of Integrative Biology, University of Wisconsin–Madison, Madison, WI 53706

ABSTRACT

Threshold concepts are fundamental to a discipline and, once understood, transform students' understanding and perception of the subject. Despite the value of threshold concepts as a learning "portal" for heuristic purposes, there is limited empirical evidence of threshold crossing or achieving mastery. As a threshold concept, biological variation within species is fundamental to understanding evolution and provides a target for analyzing threshold crossing. We aimed to 1) examine student understanding of variation using four dimensions of a threshold concept (discursive, troublesome, liminal, and integrative), 2) measure "threshold crossing," and 3) investigate the utility of the threshold concept framework to curriculum design. We conducted semistructured interviews of 29 students affiliated with a "variation-enriched" curriculum in a cross-sectional design with precurriculum, current, and postcurriculum groups (Pre, Current, and Post) and an outgroup of three postbaccalaureate advanced learners (Outgroup). Interview transcripts revealed that Current students expand their "variation discourse," while the Post group and Outgroup displayed conformity in word choice about variation. The Post and Current groups displayed less troublesome and more integrative responses. Pre, Post, and Outgroup explanations' revealed liminality, with discomfort and uncertainty regardless of accuracy. When we combined all four threshold concept dimensions for each respondent, patterns indicative of threshold crossing emerged along with new insight regarding curricular design.

INTRODUCTION

Understanding the origins, structure, and processes that produce and limit biological variation within species is fundamental to understanding biology and the diversity of life. Variation originates through changes in genetic information, presents itself as an array of structures and functional phenotypes, and is a prerequisite for evolution, thereby bridging three of the four core concepts outlined for biology undergraduates in *Vision and Change* (American Association for the Advancement of Science [AAAS], 2011). Despite being fundamental, variation is not typically or explicitly taught; instead, many instructors assume that students already understand and can apply this concept (Smith, 2010a,b). Within biology, "biological variation within species" aligns well with characteristics of a threshold concept (Ross *et al.*, 2010).

Land *et al.* (2010) describe threshold concepts as fundamental concepts of a discipline that, once understood, transform a student's perception of a whole subject and permit access to a previously inaccessible way of thinking, understanding, or interpreting something. First described by Meyer and Land (2003), threshold concepts have been established across many disciplines, including biology (Taylor, 2006; Ross *et al.*, 2010). Of the eight benchmark features recognized for threshold concepts (most notably assembled by Flanigan, 2018), we focused on the following four as measurable within our context:

Jennifer Loertscher, *Monitoring Editor*

Submitted Dec 14, 2018; Revised Apr 15, 2019; Accepted Apr 18, 2019

CBE Life Sci Educ September 1, 2019 18:ar36

DOI:10.1187/cbe.18-12-0241

*Address correspondence to: Janet Batzli (jcbatzli@wisc.edu).

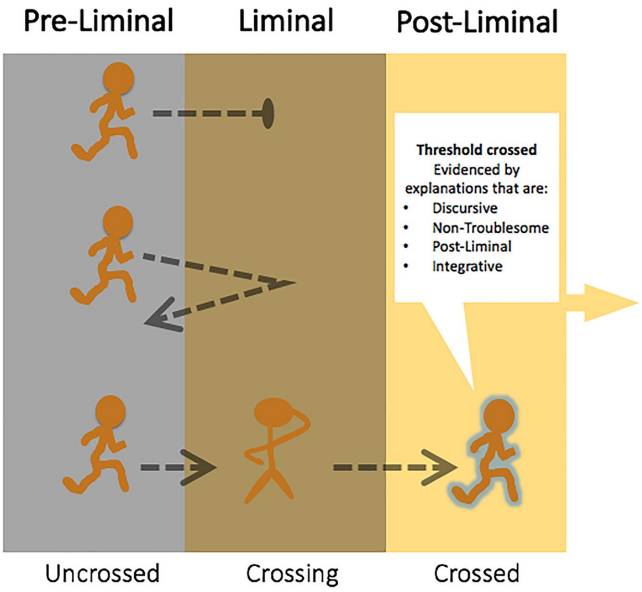

FIGURE 1. Hypothetical model of threshold crossing inspired by Land *et al.* (2014). Learners can take multiple paths, oscillating in and out of a liminal space as they approach, learn, and master a threshold concept. While we recognize thresholds may exist sequentially or in an overlapping or web-like matrix, this simplified model suggests a single threshold. The process of crossing a threshold of learning is accompanied with disciplinary language acquisition that is bounded and specific to the threshold concept (discursive); the precision and accuracy of understanding (nontroublesome); no longer confused or uncertain in understanding (postliminal); and connected with other related concepts (integrative).

- *Discursive* or important for influencing disciplinary language during discourse;
- *Troublesome* or difficult for students to gain accurate understanding;
- *Integrative* or requiring integration of many ideas within a discipline; and
- *Liminal* or requiring time in a liminal state of uncertainty in which the learner self-evaluates his or her limited understanding or misunderstanding of a concept. Notably, we view liminality as distinct from being "tentative" to accept predictions without experimentation (as observed in Halmo *et al.*, 2018). While tentativeness is a defining feature of scientists' critical view of truth, we view liminality as one's uncertainty of one's own personal knowledge base and conceptual understanding.

Figure 1 illustrates how learners could approach their understanding of a threshold concept from a variety of states of understanding and may even move back and forth between preliminal and liminal states in a nonlinear or oscillating way before ultimately crossing the threshold of understanding into a postliminal state (Meyer and Land, 2005, 2006). Liminality is eventually accompanied by a shift in the learner's understanding, and sometimes his or her identity, which can be transformative (Meyer and Land, 2003; Meyer *et al.*, 2006). As suggested by Land *et al.* (2014), we recognize that thresholds may be consecutive and overlapping, with learners seamlessly moving between one threshold concept and another. For example, a student may confront "randomness" and "variation within species" as two overlapping and related threshold concepts (Ross *et al.*, 2010) when tackling a deep understanding of evolution.

Despite the utility of the threshold concept model as a heuristic for learning about biological variation within species, there is no empirical evidence of "threshold crossing" (Batzli *et al.*, 2016). Using a simplified model (Figure 1), we attempted to measure threshold crossing within the context of a biological variation–enriched curriculum. In this model, we defined a threshold as being "crossed" when an individual can express the concept using discipline-specific language (discursive), free from uncertainty (postliminal), in an accurate or nontroublesome and integrated way. The purpose of this research was 1) to examine students' capacity to observe, explain, and represent the basis of variation within species using a threshold concept framework; 2) to generate an approach to detect threshold crossing; and 3) to test the utility of the threshold concept model at different stages within a variation-enhanced curriculum.

Theoretical Framework

As suggested previously, we view the threshold concepts model as a helpful heuristic, rather than an empirically grounded theoretical framework within the context of an undergraduate biology curriculum (Batzli *et al.*, 2014). Instead, we conceive of threshold concepts as compatible with the learning progressions framework (Batzli *et al.*, 2016). Learning progressions describe testable hypotheses of phases of understanding that learners progress through in mastering a particular skill, which often takes place over a period of several years (Corcoran *et al.*, 2009; Duncan and Rivet, 2013). Examples of certain sticking points in a learning progression at which progression temporarily stalls have been noted by Mohan *et al.* (2009) and White and Maskiewicz (2014). We suspect that some of these sticking points actually require understanding threshold concepts (Batzli *et al.*, 2016). Therefore, we view a threshold concepts model as being a helpful tool in designing curricula for a learner to surmount some of those challenging phases.

Experimental Questions and Predictions

While there have been many publications outlining the constructs behind threshold concepts (for a comprehensive bibliography, see Flanigan, 2018), there is little empirical evidence of threshold crossing itself (Nicola-Richmond *et al.*, 2018). Here, we attempt to find evidence for threshold crossing for the proposed threshold concept of biological variation within species.

We focused on biological variation within the context of a three-semester laboratory curriculum, parts of which have been designed with consideration of variation as a threshold concept (Batzli *et al.*, 2014). Students who complete this variation-enriched curriculum observe, model, manipulate, measure, analyze, and explain variation firsthand through experimentation in an iterative manner over three semesters. Over three semesters, students also participate in eight to 10 nongraded, "feedback presentations," similar to research lab meetings (Batzli *et al.*, 2018), in which student research teams are given time and permission to propose research, ask questions, share understanding, discuss misunderstandings and uncertainties, visualize and analyze data, tease apart conclusions, and otherwise

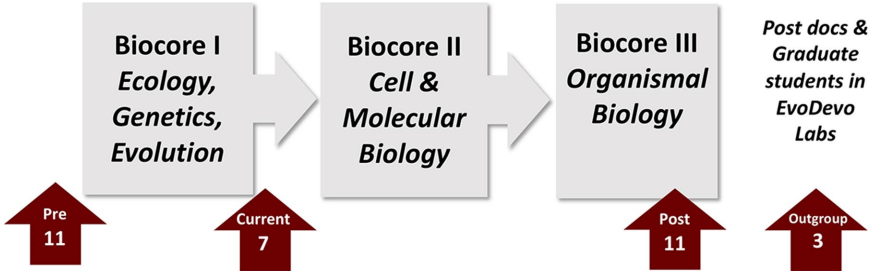

FIGURE 2. Cross-sectional groups of respondents in the semistructured think-aloud interviews. Each box represents one semester of Biocore laboratory course work, which is inquiry-based. The number in the red arrow denotes the number of respondents interviewed from each group. Further descriptions of the groups can be found in the *Context and Study Sample* section.

situate their learning in a liminal state. It is within this context that we ask the following questions:

1. Can we detect dimensions of threshold concepts (discursive, troublesome, liminal, integrative) within students' explanations of variation?
2. Can we combine these dimensions to detect evidence of threshold crossing in students' explanations of variation?
3. Can we detect a difference in students' explanations of variation at different points in a variation-enriched curriculum using threshold concept dimensions?

To answer these questions, we conducted 32 semistructured interviews during which respondents observed firsthand a novel case of biological variation (10 preserved specimens of the same bird species). We used a cross-sectional experimental design, with 29 respondents being interviewed before, during, or after taking part in the curriculum. We also interviewed three advanced learners (postdocs or doctoral graduate students) as an outgroup, because we assumed they had experiences similar to our variation-enriched curriculum (observe, model, analyze variation, etc.) but outside and independent of our curricular context. We qualitatively examined all interview responses for four threshold concept dimensions discernible given the study design: 1) discursiveness, or use of discipline-specific language; 2) troublesomeness, or lack of accuracy; 3) liminality, or evidence of self-described uncertainty; and 4) integrative thinking, or capacity to explain variation across biological scales and processes. We then combined these four dimensions to examine holistic evidence of threshold crossing.

METHODS

Context and Study Sample

During the 2015–2016 academic year, 32 respondents were interviewed at the University of Wisconsin–Madison, a large, public, land-grant institution. Twenty-nine respondents were affiliated with the Biology Core Curriculum (Biocore) program, a 2-year honors biology program that emphasizes inquiry-based learning, group work, and process of science skills. (For more information about this program, see Batzli, 2005.) There were three cross-sectional groups of students affiliated with the Biocore program: 1) pre-Biocore students who had been accepted into the program based on chemistry and math prerequisites but had not yet begun Biocore course work (referred to as "Pre," n = 11), 2) current Biocore students who had completed the first semester of Biocore I (referred to as "Current," n = 7), and 3) post-Biocore students who had completed the third inquiry-based laboratory course, Biocore III (referred to as "Post," n = 11). All students enrolled in Biocore were invited to participate. Additionally, we recruited advanced learners (one senior-level doctoral student and two postdoctoral fellows) from research laboratories focusing on genetics, developmental, and/or evolutionary biology (referred to as "Outgroup," n = 3; Figure 2). We included the Outgroup as a comparison to the Pre, Current, and Post groups in our formalized curriculum, and perceived the Outgroup as advanced learners who had similar but even more extensive variation-enriched experiences through their own independent research. We expected these advanced learners to demonstrate mastery and to have crossed the variation threshold within the context of this study. All respondents signed an informed consent form before participating in the study (IRB 2015 00005399), which included obtaining student records and demographic data via an institutional database and respondents' approval to participate in an audiotaped interview.

To assess the representativeness of students participating in interviews compared with Biocore students as a whole, we compared demographic data from student records between interviewed (n = 29) and noninterviewed (but still consenting) Biocore students (n = 46). There were no statistically significant differences between Biocore students who were interviewed and their noninterviewed peers, respectively, for gender (65.5 vs. 65.2% female, chi-square = 0.001, df = 1, p = 0.98), ethnicity (3.9 vs. 8.9% minority groups, chi-square = 0.64, df = 1, p = 0.42), or cumulative GPA (3.58 vs. 3.56%, independent-samples t test, t = 0.12, df = 73, p = 0.91), indicating our respondent sample was representative of Biocore students as a whole for these demographic factors.

Semistructured Interviews

Interview Script Development. Overall, the purpose of the interview was to expose respondents to a genuine example of biological variation, while asking them to observe, depict, explain, interpret, and predict the consequences of that variation for phenotypes of future generations. To do so, respondents observed 10 preserved *Sturnus vulgaris* (common starling) specimens. The phenotype of these specimens varied for a multitude of traits, including feather mottling, beak size, beak color, and tarsus length. Respondents were able to choose a trait of interest to examine for the duration of the interview. In general, the questions were phrased in vernacular language so that respondents could make sense of the questions without presuming any baseline vocabulary. Additionally, to help respondents make sense of the purpose of the interview, they were first presented with familiar examples of variation within species (i.e., color morphs of various animals) before being asked more specific questions about the *S. vulgaris* specimens.

We employed a backward design approach to generate an interview script that elicited respondents' general capacity to observe, depict, explain, interpret, and predict biological variation within species. As two specific examples of the observation

and prediction skills that we aimed to assess, we asked students to identify and describe a trait that differed among the group of *S. vulgaris* specimens (observe variation) and then asked them how they expected the contents of the birds' cells to compare (predict variation). We derived questions by first focusing on the purpose or skill that we aimed to observe, then formulating a question that we expected to elicit potential responses from respondents based on the goal of the question. Then, we performed pilot interviews for two consenting Biocore students to verify that our questions were being interpreted as intended for purposes of internal validity. Only slight modifications to wording were required after piloting, so those interviews are also analyzed here. A full interview script, including the backward design approach, is available in Supplemental Material 1.

Interview Implementation and Transcription. During Fall 2015 and Spring 2016, a single researcher (E.W.S.) completed 32 one-on-one semistructured interviews. Each interview lasted between 35 and 60 minutes. Each respondent was asked the same set of questions, with follow-up questions flexibly applied based on the respondent's initial response. During the interview, we instructed students to "think aloud" about their reasoning and to make it clear when they felt uncertain or confused (Ericsson and Simon, 1980; Dancy and Beichner, 2002). We began each interview with a brief training session to allow feedback and conditioning of respondents to think aloud as they explained a graph unrelated to the biological focus of the interview (see Appendix A, Supplemental Material 1). Following this training, respondents were shown images with examples of variation within species and asked to describe examples of variation they had observed in their own lives (question 1; See Appendix B, Supplemental Material 1). The interviewer was careful not to introduce any discipline-specific language (including the words "species" and "variation") until it had been uttered by the respondent. Then, respondents were shown 10 preserved specimens of *S. vulgaris* and were asked to describe their observations about the differences between specimens (question 2). Next, they choose a trait of interest that varied among the specimens and were instructed to order the specimens from one extreme of the trait to the other and to generate an illustration of that variation diagrammatically (question 3). Then, the interviewer asked respondents to describe and summarize the cellular and molecular origin of the variation they had just diagrammed (question 4). Finally, 10 new specimens were presented as hypothetical offspring of a subset of the initial "parent" specimens after one generation. At this point, respondents were asked to interpret any changes in the variation across generations and formulate a plausible scenario for this observation (question 5). Scripted follow-up questions were asked for each main question (see full script in Supplemental Material 1). In addition, spontaneous, unscripted follow-up questions were asked to clarify responses for each respondent. For the purposes of the analysis of threshold crossing, we focus on questions 1, 4, and 5, which provided respondents with opportunities to articulate threshold concept dimensions. Each interview was audio-recorded, then the entire interview was transcribed verbatim for further analysis. Transcripts were deidentified from the respondents' names and given randomly generated four-alphanumeric character identifiers so that coding could be done while researchers were blind to group affiliation.

Qualitative Data Analysis

Word Counts and Discipline-Specific Word Usage. To begin the coding of the discursive dimension, we first generated a library of discipline-specific words used by our respondents. We focused on question 1 (respondents described images of biological variation within species from their own observations) early in the interview and on question 4 (respondents predicted differences within the cells of birds whose phenotype varied) later in the interview. Two researchers (E.W.S. and H.B.) read through transcripts and flagged potential discipline-specific words for further discussion. Then, with reference to textbooks and biology dictionaries, three researchers (E.W.S., H.B. and J.B.) formed a consensus about words that were disciplinary and those that were vernacular in this context. (As an example, a vernacular term would be "color," while a disciplinary word could be "pigment" or "melanin" to describe the same trait.) This word-sorting process resulted in a library of 286 discipline-specific words that were cross-referenced to each respondent's transcript and compiled using word identification and counting formulas in Microsoft Excel. There were no discipline-specific words in the stem of question 1 and only the word "cell" in the stem of question 4, which, if uttered in the response, was also counted. Disciplinary words were counted independent of the accuracy of their use.

Coding for Threshold Concept Dimensions. Given our overall aim to analyze respondents' descriptions of variation for evidence of threshold crossing, we were specifically interested in coding the transcripts along four dimensions: 1) use of discursive language, 2) troublesome explanations, 3) liminal comments, and 4) integration among multiple biological scales. The development of these nonoverlapping coding schemes was a highly iterative process that largely occurred in two stages (Table 1).

During the first stage, independent rubrics were developed for each threshold concept dimension. Within each dimension, one researcher (E.W.S.) explored a subset (5–10 out of 32) of transcripts to develop a rubric in discussion with another researcher (J.B.). These rubrics were implemented to score each respondent's descriptions on several levels. Two researchers (E.W.S. and J.B.) then independently coded each respondent's description for each dimension. When there was disagreement, the nuances of each response were discussed until we a consensus was reached (as described by Stanton *et al.*, 2015; Dye and Stanton, 2017).

Important to coding for the troublesome dimension in the first stage of coding, we used Perkins (2006) to further define troublesome knowledge as knowledge that is: memorized without deep understanding (ritualized), retrievable for an exam but not readily transferred to new scenarios (inert), incompatible with personal beliefs (foreign), used without awareness (tacit), or incompatible with previous experiences (conceptually difficult). Within biology and specifically evolution, Coley and Tanner (2012, 2015) have further looked at students' explanations of conceptually difficult knowledge and found other nuanced categories of troublesome reasoning, including the overapplication of intuitive reasoning (counterintuitive), such as believing that every biological structure or form was derived for a biological function (teleological reasoning), every organism in a biological taxa has the same form and function (essentialist reasoning), or all organisms can be explained in human terms (anthropocentric).

TABLE 1. First- and second-stage coding schemes for each threshold concept dimension[a]

Dimension	First-stage coding	Second-stage coding
Discursive	Respondents earned 1 point for each type of variation that was described using discipline-specific words. The following types of variation were described: • Allelic • Chromosomal • Gene expression regulation • Environmental • Gene products/biochemicals • Development and aging • Cell signaling Respondents' scores ranged from 0 to 4.	• No types of variation were described using discipline-specific words (scored 0 in first-stage coding). • At least one type of variation was described using discipline-specific words (scored 1–4 in first-stage coding).
Troublesome[b]	Respondents' descriptions were examined for the following troublesome categories. All occurrences were summed with equal weight: • Essentialism (OAI) • Teleological (OAI) • Anthropocentric (OAI) • Answer given without reasoning (R) • Overapplication of Mendelian thinking to describe multigenic traits (R) • Inaccurate use of gene and allele (I) • Genetic equivalence among individuals (I) • Inaccurate understanding of gene expression (I) Respondents scores ranged from 0 to 3. Land *et al.*, 2005, 2010; Perkins, 2006; Shtulman and Schulz, 2008; Coley and Tanner, 2012, 2015; Speth *et al.*, 2014; Emmons and Kelemen, 2015; Richard *et al.*, 2017	• Explanation contained one or more troublesome categories (scored 1–3 in first-stage coding). • Explanation was free from all troublesome categories (scored 0 in first-stage coding).
Liminality	Respondents' descriptions were examined for the following evidence of liminality. All occurrences were summed with equal weight: • Oscillating between more than one answer • Self-reported mimicry • Self-reported discomfort or uncertainty Note that a tentativeness in accepting a new assertion without further data (Halmo *et al.*, 2018) is not coded as liminal. Respondents' scores ranged from 0 to 2. McCartney *et al.*, 2009; Land *et al.*, 2014	• Explanation contained one or more liminal categories (scored 1 or 2 in first-stage coding). • Explanation was free from all liminal categories (scored 0 in first-stage coding).
Integrative	Respondents' descriptions of variation were examined for the integration of the following biological scales. All occurrences were summed with equal weight: • Genes or alleles • Gene products/biochemicals • Seasonal, environmental, or developmental • Population • Population over time Respondents scores ranged from 0 to 5. Batzli *et al.*, 2016	• Explanation contained one or no additional biological scales (scored 0 or 1 at first-stage coding). • Explanation contained two or more additional biological scales (scored 2–5 at first-stage coding).

[a]The categories in first-stage coding were generated based on observed respondent descriptions. References for the rubrics for each dimension are shown following the respective entries.
[b]For "troublesome" categories, the following acronyms are defined: OAI, overapplication of intuitive reasoning; R, ritualized; or I, general inaccuracy.

In the second stage of coding, for the purpose of combining the threshold concept dimensions for overall evidence of threshold crossing, we converted the first-stage rubric score for each dimension to a binary score of either 0 or 1 based on adequate evidence of mastery. The rubric scores from the first stage of coding, which ranged from 0 to 5, were binned into a binary code (i.e., 0 = not sufficient evidence toward mastery or 1 = sufficient evidence toward mastery). To determine bin size for the binary code, we examined the responses for natural breaks in the distribution to select the cutoff. First-stage and second-stage coding schemes for each dimension are summarized in Table 1 and described in detail in Supplemental Material 2. Only second-stage coding scores are presented in the "Results" section. All rubric development and coding were done while researchers were blind to the group affiliation for each respondent.

Combining Dimensions for Determining Threshold Crossing. Given the overall aim to detect threshold crossing, we next sought to combine the four threshold concept dimensions into

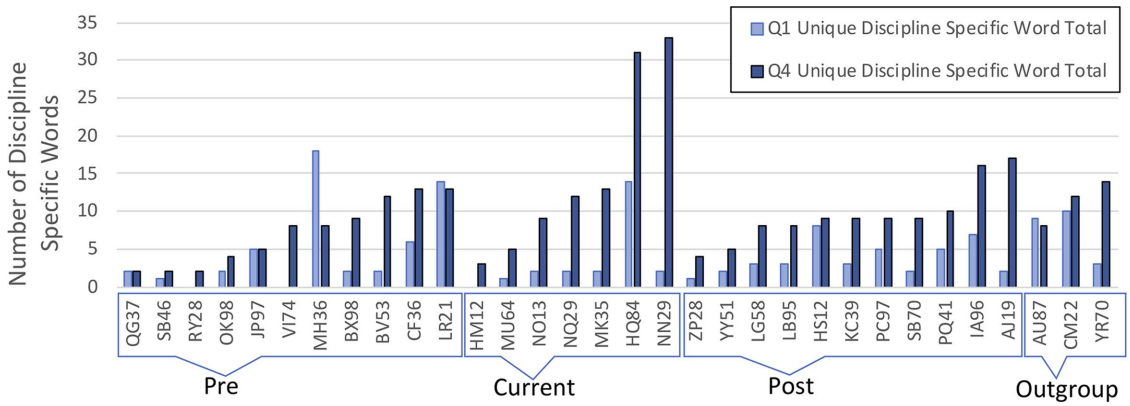

FIGURE 3. Discipline-specific word usage early (question 1) and later (question 4) in the interview for each respondent. For reference, the question 1 prompt was, "Have you ever seen this [the same kind of animals that all look really different] [images provided] in your own life? Can you provide a few examples?"; and the question 4 prompt was, "If you think about these two individuals [pick up birds] that differ for 'X' [trait selected by respondent], how would you expect the contents of their cells to compare?"

a single score for each respondent. For the integrative, troublesome, and discursive codes, this was done as a simple addition of the second-stage component scores (as seen in Table 1). However, the liminality code was complicated, due to the fact that both preliminal and postliminal states would be expected to lack evidence of liminality (Meyer and Land, 2005; Figure 1). Therefore, we created a rule to qualify a respondent's liminality score based on his or her troublesome score, as we would expect a postliminal learner to also be free of troublesome explanations, while a preliminal learner would still exhibit troublesome explanations. Based on this prediction, if a respondent's description was considered nontroublesome, the absence of liminality counted toward threshold crossing (+1). Thus, the combined scores of the four threshold concept dimensions included here ranged between 0 and 4.

Statistical Analysis of Qualitative Codes. We used nonparametric statistics for comparisons or associations between cross-sectional groups due to our small sample sizes. Specifically, we used Fisher's exact test for group comparisons of binary code outcomes (e.g., proportion of respondents exhibiting evidence for each threshold concept dimension) and co-occurrence of binary outcomes, Kruskal-Wallis tests for group comparisons of ordinal or continuous outcomes (e.g., word counts), and Spearman's rank correlation for associations of ordinal outcomes (e.g., additive dimension scores along the experience-of-subject matter axis indicated by group). For the discursive dimension, we additionally used Levene's test to assess homogeneity of variance in word count across groups, and logistic regression to test whether word count was a predictor of second-stage discursive binary code score. All statistical tests were run using SPSS Statistics v. 23.

RESULTS

Word Counts and Analysis

We report word count comparisons of each respondent by group in Figure 3. The counts represent discipline-specific words from a library of a total of 286 words that we classified as "scientific" or related to the discipline of biological science. Each count represents a unique word in a respondent's explanation either early (question 1) or later (question 4) in the interview, with no word counted more than once within the response to each question.

In composite, each respondent's utterances to question 1, in which they were asked to provide examples from their own observations of "the same kind of animals that all look really different" when given images for reference, elicited between 0 and 18 unique discipline-specific words. Word counts for question 4, in which respondents are asked how the "contents of birds' cells compare," increased from 2 to 33 words as the interview and discourse about biological variation progressed. Comparing word counts between these two questions indicates a shift in the respondent's word choice, as modified by conversation or discourse with the interviewer and interaction with the materials.

For question 1, the respondents used words that were sometimes very specific to their personal experience. For instance, word counts included references to "dark-eyed juncos"; respondents' family pets and specific dog breeds; flower varieties in the family garden; variant characteristics in their friends, family, or classmates (human eye, hair, skin color, height); examples from lab experiences working with *Brassica rapa* FastPlants or another model species (*Lumbriculus variegatus*); or mimicked examples from a recent lecture on evolution of "rock-pocket mice." In other instances, the respondents referred to general taxa (e.g., birds, fish) with self-proclaimed difficulty in identifying what constituted variation within a species, stating "within species variation … it's hard to think of specific examples. I don't really look at nature too much." When respondents referred to variant phenotypes as described with the word "color," these were not counted as discipline-specific words (e.g., brown- vs. blue-eyed humans) as compared with differences in *pigmentation*, which was considered a discipline-specific word (e.g., anthocyanin pigmentation in Brassica rapa stems has 3 discipline-specific words). Regardless of the context, there was no significant difference in mean word count among groups for question 1 (Kruskal-Wallis test, $H = 4.94$, $p = 0.18$)

For question 4, respondents' word counts increased, were less personal, and included more discipline-specific terminology that bioscientists would recognize. The most frequent discipline-specific words uttered were "genes," "DNA," "allele(s)," "cell(s)," "dominant," and "recessive," in order of frequency.

Although not statistically significant across the four groups (Kruskal-Wallis test, $p = 0.26$), the Current group showed the highest average word count and SD (~15 ± 12 words). All other groups (Pre, Post, and Outgroup) revealed average word counts of 7 ± 4, 9 ± 4, and 11 ± 3, respectively. Despite the lack of difference in the mean word count, we noted the large SD in word count in the Current group was mostly due to respondents HQ84 and NN29 uttering 31 and 33 different discipline-specific words, respectively. Levene's test showed that the differences in variance across groups was statistically significant ($p = 0.002$). This result suggests that students of the Current group had the greatest diversity of word counts, having freshly learned or being in the midst of learning new vocabulary, with at least two students eager to share and demonstrate their nuanced vocabulary. Conversely, Post and Outgroup respondents consistently used a more parsimonious, discipline-specific shorthand requiring fewer words. A few students, mostly in the Pre group, were either ill prepared or could not summon much of a response (see Figure 3, QG37, SB46, OK98, HM12, ZP28), resulting in responses containing fewer than 5 discipline-specific words.

Discursive Dimension

Analysis of discipline-specific word count and quality was a productive initial step for assessment of threshold crossing, because word choice and language use is the first step in evaluating one's ability to communicate a concept.

As described earlier, disciplinary language was brought out by discourse during the interview, with question 1 asking students about differences in organisms they had observed in their own lives, whereas question 4 asked about differences between specimens at the cellular level (see interview script in Supplemental Material 1). We further analyzed the discursive dimension pertaining to respondents' word choice when specifically describing biological variation. Sufficient mastery was defined as an explanation describing at least one form of variation within species using discipline-specific words. If respondents described one or more forms of biological variation within species with discipline-specific terminology, they received a score of 1 for the discursive dimension. Furthermore, we used binary logistic regression to determine whether respondent word counts predicted binary scores. For question 1, 28 of the 32 respondents achieved a binary score of 1, and word count for question 1 was not predictive (Nagelkerke $R^2 = 0.27$, $p = 0.13$). For question 4, however, 22 of the 32 respondents achieved a binary score of 1, and word count for question 4 was positively predictive (Nagelkerke $R^2 = 0.87$, $p = 0.001$), with higher word counts resulting in better articulation of variation. This suggests that the discipline-specific word choice in question 4 responses was related to respondents' descriptions of variation within species, while word choice in question 1 responses was not. While respondents were able to retrieve and use more discipline-specific words later in the interview in question 4, question 1 responses helped frame the "variation discourse" between the interviewer and respondent. But for the purpose of analyzing variation as a threshold concept, we focused our analysis on question 4 responses, because those could reveal the most about respondents' understanding of variation within species.

Binary coding for the discursive dimension revealed 10 respondents who did not demonstrate evidence of sufficient mastery (were nondiscursive). Among respondents who achieved the discursive score of 1, all used a minimum number of at least nine words (Figure 3), mostly drawn from a specific set (genes, DNA, allele(s), cell(s), dominant, recessive, phenotypic, genotype, gene variants), to articulate (regardless of accuracy) biological variation at the cellular level. When we compared the proportion in each curricular group categorized as nondiscursive versus discursive (Figure 4A), we found no differences (Fisher's exact test, $p = 0.15$) and determined that all groups had the capacity to use discipline-specific words to describe at least one form of variation at the cellular level.

Troublesome Dimension

Most respondents exhibited some inaccuracy in their explanation of variation, and the proportion differed among the groups. We found that 22 out of 32 respondents across groups who expressed inaccurate or ritualized (Perkins, 2006), overly intuitive, essentialist, or teleological (Coley and Tanner, 2012, 2015) aspects within their explanations of variation and that the proportion of troublesome responses differed among groups (Figure 4B; Fisher's exact test, $p = 0.03$).

An example of a troublesome response was from MK35 in the Current group: "I would expect the DNA [in all bird specimens] to be exactly the same. Maybe the rate at which it is transcribed and translated is different" (Table 2). Another example of a troublesome response is from Pre respondent MH36: "DNA is the same [among bird specimens]. DNA is universal, but like maybe different effects like epigenetics, different restrictors on segments of DNA that would allow for the certain expression." Both of these examples reveal a lack of congruence or nuanced understanding for how genotypic or allelic variation influences phenotypic variation at the whole-organism level for the 10 starling specimens, but the respondents recognize that the rate, timing, and effectors of gene expression all influence phenotypic variation.

We found that all Pre respondents had some inaccuracy or overapplication of intuitive reasoning in their explanations. For the Current and Post groups, we found 57 and 64% of the responses to be troublesome, respectively. The use of precise language and accurate explanation is challenging to master. Even the Outgroup consisting of postdocs and doctoral students studying evolution and/or developmental biology revealed troublesome explanations. Consider the following from Outgroup respondent YR70: "Differences will be expressed in the DNA, but the DNA is identical between these two birds…. I will pretend I am talking to my mom…You live in Texas, you need more white tipped feathers, you get a longer hormonal cue because of an environmental driven signal. Therefore, the gene stays on longer." This explanation reveals teleological reasoning, conflation of the term "gene" with "allele," and anthropomorphism of the bird as a resident of Texas. In addition, it is unclear whether the respondent is attempting to use a shorthand to simplify the explanation or whether his or her understanding is truly troublesome.

Liminality Dimension

Respondents varied in the types and levels of liminality they revealed. For instance, Pre respondent LR21 offered a simple answer: "I don't really know the answer. I'm not too sure." However, Post respondent IA96 struggled with recalling vocabulary: "I think it's epi, no I don't think its epistasis, maybe it is…"

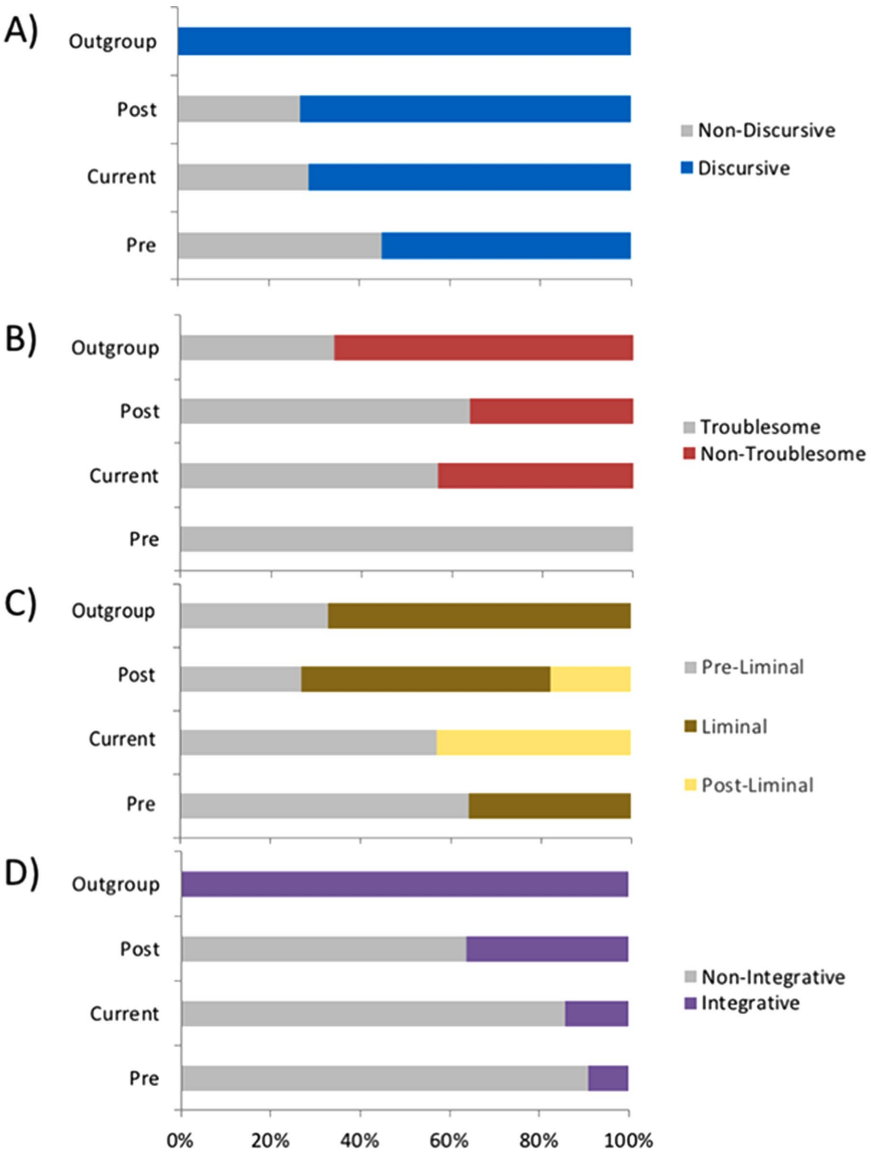

FIGURE 4. Proportion of respondents (Pre *n* = 11, Current *n* = 7, Post *n* = 11, Outgroup *n* = 3) displaying evidence for each threshold concept dimension among cross-sectional groups. Significant differences in responses across groups were found for the troublesome dimension (B, Fisher's exact test, *p* = 0.03), liminal dimension (liminal vs. nonliminal; C, Fisher's exact test, *p* = 0.05), and integrative dimension (D, Fisher's exact test, *p* = 0.02), but no significant differences were found for the discursive dimension (A, Fisher's exact test, *p* = 0.15).

(Table 2). Even Outgroup respondent AU87 revealed liminality: "I think that I don't know the answer to that question," indicating that liminality is a continuum and associated with all levels of understanding. We also acknowledge that liminality has multiple meanings associated with self-confidence, how much one values or is interested in the topic, competence, identity, and comfort level. With that, we found liminality to be an intriguing dimension unique to coding for threshold concepts, given how it should be considered in combination with the troublesomeness dimension for threshold crossing (Land *et al.*, 2014).

Many respondents (20/32) self-identified their understanding as uncertain, confused, or lacking completeness (i.e., liminal), but the proportion of liminal versus nonliminal (whether preliminal or postliminal) responses differed among groups (Figure 4C; Fisher's exact test, $p = 0.05$). About a third of the Pre respondents and more than half of the Post respondents revealed some indication of liminal explanations. Interestingly, there were no respondents from the Current group who exhibited liminal responses. This could be associated with the recency, confidence in recall, and more proximal nature of learning the subject material in their cognition.

Integrative Dimension

When respondents were asked, "What is one plausible scenario that would result in the situation that you see here in a natural environment?" (referring to the phenotypic variation in the offspring specimens as predicted by phenotypic variation of parents; question 5) few of the Pre, Current, and Post respondents could offer integrative explanations of variation from genes and gene products to the biochemical and physiological level up to the organismal level and evolutionary processes. For instance, most confirmed the notion that phenotypic variation is present in parent and offspring populations and that the genetic variation that exists in a parent population is transmitted to offspring, but few offered an evolutionary explanation (e.g., selection) or one that included a molecular/genetic explanation of the phenotype and an environmental scenario for variation in expression. In comparison, all of the Outgroup respondents provided integrative explanations of the variation they observed, including variation in genotype, gene product (biomolecule), organismal phenotype, population, and the environment. We provide example quotes in Table 2 with underlined words and phrases that are indicative of biological scales. The integrative dimension revealed a potential developmental trend and group differences (Figure 4D; Fisher's exact test, $p = 0.02$), wherein more integration was seen as the experience-level of the respondents increased.

Co-occurrence across Dimensions

Following our analysis of each dimension independently, we analyzed for co-occurrence to examine the relationships between dimensions. We found co-occurrence to be statistically significant between discursive and troublesome dimensions (Fisher's exact test, $p = 0.03$), discursive and integrative dimensions (Fisher's exact test, $p = 0.03$), and troublesome and integrative dimensions (Fisher's exact test, $p = 0.006$; see Supplemental Material 3 for data tables). While these findings are

TABLE 2. Quotes from respondents' explanations that illustrate the troublesome, liminal, and integrative threshold concept dimensions

Explanation	Level	Example quotations from student interviews[a]
Troublesome	Pre	"If you have the Punnett square, this [bird] has maybe all the recessive alleles. But molecularly, I'm pretty sure they're [the birds], like, are all the same."
	Current	"I would expect the DNA to be exactly the same. Maybe the rate at which it is transcribed and translated to be different."
	Post	"Each bird has slightly different alleles and it needs to express different proteins in order to adapt to a specific environment as needed."
	Outgroup	"I will pretend I am talking to my mom.… You live in Texas, you need more white tipped feathers, you get a longer hormonal cue because of an environmental driven signal. Therefore, the gene stays on longer."
Liminality	Pre	"I don't really know the answer. I'm not too sure. Yep."
	Current	N/A
	Post	"I think it's epi, no I don't think its epistasis, maybe it is…"
	Outgroup	"I think that I don't know the answer to that question."
Integrative	Pre	"If they are the same species, they should at least have the same cells … Well most cells. The only different part would be maybe some of the proteins within the beak cells that I guess are coding. But like DNA is universal, but like maybe different … different restrictors on segments of DNA would allow for certain expression [in beaks]. I would argue that they have the same genes. But I think the alleles are different."
	Current	"The organelles and stuff, that would be the same. Most of the difference you'd see would likely be um, in the DNA … a lot of really small variations and so like, alleles … so whether it's just a single base pair or like a few base pairs that like compose one allele, you can have spots. Or, it might be you have alleles that control whether other alleles can be expressed, and control how many spots you have, what color the spots are, how big the spots are, where the spots are, all of that stuff. … The white spots don't have the black pigment, so you know in those cells the enzyme that produces the pigment isn't going to be functioning."
	Post	"I would expect more proteins involved in the synthesis of the pigments, and more mRNA and everything that is used to produce those. I would expect the genes to be variations within like the nucleotide sequence of the DNA. Um, and in one case it could, um, encode a protein, or a protein that's better capable of synthesizing those pigments within the cell. You know like a transcription factor could cause the expression of that gene and lead to more mRNA, leading to proteins that are expressed in the cell to give those pigments.… I wouldn't know the exact mechanism by which they respond but the environment would have to act on something that would then lead to the transcription of these multiple genes involved."
	Outgroup	"I would imagine there was different amounts of growth signaling. And so whatever gene is signaling how much beak grow, there is probably more in the larger bird than the smaller bird. With development, you have a transcription factor that is turning on a gene that expresses white pigment. So differences in transcription factor abundance between cells. They generally have the same genome, and the same genes, but different alleles. However, there are copy number variations and other um insertions and deletions that can lead to inner individual variation in gene number. Also, I think that environment at different times in different seasons, in different ages, in different sexes, can lead to huge phenotypic effects."

[a]Quotes are taken from a variety of respondents and are meant to provide specific examples of our interpretation of these benchmark dimensions. These quotes came from the points in the interview when respondents were asked, 1) "Recently, one of my friends show me these examples of the same kind of organisms that all look really different. Have you ever seen this in your own life? Can you provide a few examples?"; and 2) "Based on 'X' (trait observed in 10 variant *S. vulgaris* specimens) how would you expect the content of the birds' cells to compare?"

perhaps intuitive, the liminality dimension was less so. There was no statistically significant co-occurrence between liminality and discursive dimensions (Fisher's exact test, $p = 0.15$) or between liminality and integrative dimensions (Fisher's exact test, $p = 0.18$). There was, however, a strong relationship between liminality and troublesome dimensions (Fisher's exact test, $p < 0.0001$), such that nonliminal responses were *never* troublesome; but this was unsurprising, because we defined nonliminal responses, in part, as containing no troublesome attributes.

Combined Analysis for Threshold Crossing

We combined binary responses for each threshold concept dimension, which resulted in an additive threshold-crossing score for each respondent. Figure 5 reveals the score for each respondent, as well as the differences and developmental patterns within and between all groups. Treating the groups as ordinal categories related to knowledge of subject matter (e.g., Pre = 1, Current = 2, Post = 3, Outgroup = 4), we found a significant positive association, such that additive dimension score increased with experience (Spearman's rank correlation, $r_s = 0.42$, $p = 0.02$, $n = 32$; Figure 5). Two respondents, NN29 (Current) and AJ19 (Post), exhibited the maximum score possible of 4, while all other respondents (including those in the Outgroup) scored between 0 and 3. The predominance of zero values for the Pre group indicates that the majority of respondents in this group exhibited troublesome and preliminal understanding. In other words, these individuals were confident and certain in their inaccurate explanations or knew that they did not understand at all. In comparison, most respondents in the Post group had the disciplinary language skills and were accurate in the explanations offered but recognized they still had liminal explanations with limited capacity to integrate their understanding. Despite the complexities of the task and oversimplification associated with this model (Figure 1), we believe the additive function illustrated in Figure 5 provides a

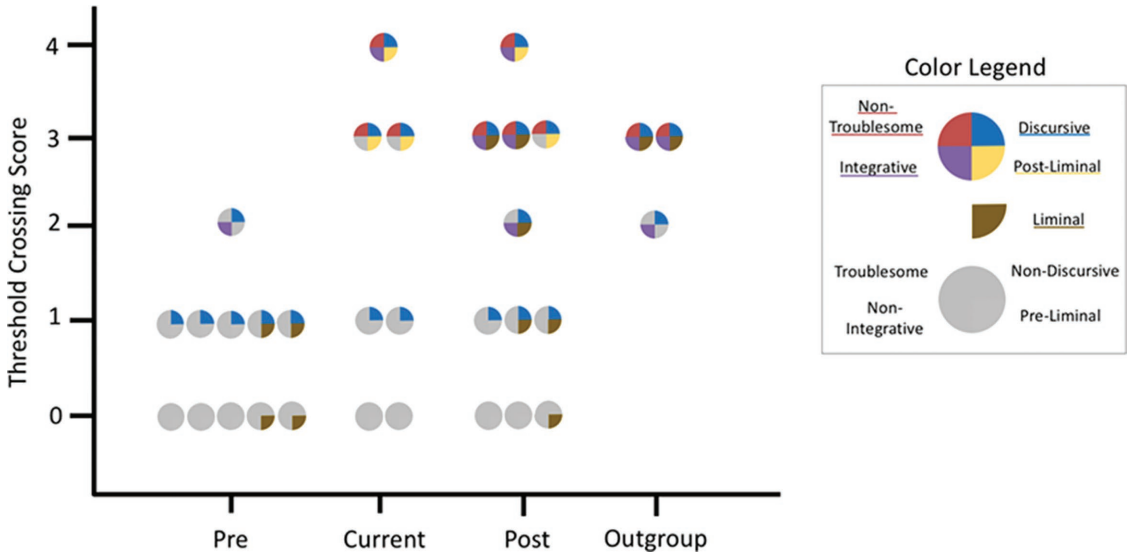

FIGURE 5. Individual respondents shown by group (Pre, Current, Post, and Outgroup) and additive threshold dimension score. Responses for each coded dimension are indicated by color. Additive dimension scores were compiled from responses using the following rule: discursive (+1), nontroublesome (+1), postliminal (+1), and integrative (+1). See the text for a complete explanation. Spearman's rank correlation, $r_s = 0.042$, $p = 0.02$.

quantitative and empirically derived baseline for analysis of threshold crossing.

DISCUSSION

Measuring Threshold Crossing

One purpose of this study was to examine whether we can detect threshold crossing for honors biological science students at precurriculum, current, and postcurriculum stages. Although these are honors students, we believe that our findings are applicable to non–honors majors and nonmajors alike, as others have found that both groups similarly endorse intuitive reasoning (Coley and Tanner, 2015). We conceive that students have crossed a threshold in their conceptual learning when they demonstrate discipline-specific dialogue through word choice (discursive) to provide accurate explanations (lacking troublesomeness), integrative explanations, and evidence of conceptual understanding that is articulated with a certain degree of comfort and confidence (postliminal).

Did we detect threshold crossing? Yes, in a simplistic way. On the basis of the model illustrated in Figure 1 and how we define transformation as evidenced by explanations that are discursive, nontroublesome, postliminal, and integrative, we believe those scoring between 1 and 3 and still liminal are in the midst of crossing, while those scoring 4 have crossed. However, using a more nuanced view, we found that our differentiation between "crossing" and "having crossed" depends on 1) students' language use and our interpretation (Green *et al.*, 2017), 2) how we define "threshold" and "sufficient mastery" of each dimension, and 3) how liminality is incorporated into the additive aspect of the cumulative threshold concept score. In this study, we chose to use a two-staged coding method wherein respondents' explanations were assessed qualitatively as to whether they met criteria that made them discursive, troublesome, liminal, or integrative; and then we assigned a binary code (0 or 1) for each dimension indicating sufficient level of mastery within each dimension. Land *et al.* (2014) describes "troublesomeness in a liminal space," with a postliminal state being transformative and a conceptual explanation being neither troublesome nor liminal (refer to model illustration Figure 1). With this in mind, liminality became a discriminator in our investigation, with nontroublesome responses that were also nonliminal deemed "postliminal."

Our method of binary coding for each dimension and adding binary scores together for each respondent may be overly simplistic, but it allowed for a straightforward approach to combine dimensions and compare respondents. This approach would be suitably flexible to qualify liminal responses in relation to troublesome explanations, to be expanded to include additional threshold concept dimensions (e.g., reconstitutive, irreversibility, transformational), and to be used to make comparisons over time (longitudinal analysis). In addition, we believe the approach can be standardized and then deployed for specific threshold concepts of interest across disciplines. To our knowledge, this approach is unique to the threshold concepts literature.

In a recent literature synthesis and meta-analysis by Nicola-Richmond *et al.* (2018), the researchers reviewed 19 studies that claimed to measure threshold crossing. Despite the arguments by Rowbottom (2007) and O'Donnell (2010) that measuring threshold crossing is impossible, the Nicola-Richmond research team support the possibility using the promising work by Cope and Staehr (2008), Shanahan *et al.* (2006), and others in a broad array of disciplines that have experimented with qualitative, quantitative, and mixed-methods designs to detect threshold crossing. Common limitations of these studies include small sample size, lack of details in reporting (interpreted as lack of rigor in coding and analysis), analysis by only one researcher, or use of a single method for data collection. Given our study of 29 students at various stages of experience (Pre,

Current, and Post) and three advanced learners (Outgroup), the use of a semistructured interview, and a standardized coding rubric used by three independent researchers, we believe our study provides a valuable approach and foundation to build upon.

Our two-staged coding method allowed us to compare multiple dimensions of threshold crossing between respondents and evaluate differences among the respondents in a relatively robust way. A larger sample size would strengthen our ability to detect similarities and differences among groups, as well as developmental patterns from Pre to Current to Post groups for troublesome, integrative, and liminal dimensions (Figure 4). In addition, the set of questions we used for the interview could be streamlined and more focused, allowing for more succinct responses that could be paired with a concept assessment such as the ACORNS instrument used to diagnose conceptual competency about natural selection (Nehm *et al.*, 2012) or the RaProEvo instrument used to measure competency of the application of randomness and probability to the context of evolution (Fiedler *et al.*, 2017), but coded using the four threshold concept dimensions. Additionally, our study treated the four dimensions of troublesome, integrative, liminality, and discursiveness weighted equally in a binary and additive way. Alternatively, one could weight dimensions differently depending on the context and the learning outcomes of the curriculum. For instance, a course intended to grow students' vocabulary in a particular subject might weight the discursive dimension more prominently and have more stringent criteria for achieving mastery, while a capstone course might place more emphasis on the integrative dimension.

In our study, the Outgroup of three advanced learners served as an illuminating comparator as well as an important model group for testing our approach for measuring threshold crossing. We expected that all three advanced learners would provide strong indicators for threshold crossing (i.e., score 4 on our additive threshold concept dimension scale). Instead, the composite scores for the Outgroup were 2 (YR70), 3 (CM22), and 3 (AU87). On the basis of their interview responses, we believe the three Outgroup members existed in a liminal space different from yet proximal to the other 29 undergraduate study respondents in the Pre, Current, and Post groups. In other words, the Outgroup members already had sufficient mastery of the threshold we were measuring and were skipping steps in logic or using an advanced shorthand to jump to a research domain or area of familiarity that they were more comfortable discussing (e.g., evolutionary biology of fruit flies and mice). For instance, one Outgroup respondent said,

> It's generally the case that people have to do a lot of work to figure out how genes underlie a trait. It's been discovered that things like size tend to be attributed to a lot of different genes, whereas, things like differences in color can be achieved by just turning on a pigment or off, so it's often monogenic. I know that because I've read a lot of papers.

There was also hesitancy to guess or generate scenarios for explaining variation given the perceived consequences of talking outside their range of research knowledge. For example, one Outgroup respondent said,

> I am not an ornithologist. I actually have no idea how feathers are made, or what kind of cells make up a feather … I'm not confident … it depends on what you define to be a gene.

Or sometimes Outgroup respondents felt they needed to support their identities as researchers and provide examples to convince the interviewer that they knew what they were talking about. For instance, YR70 said,

> There can be subtle variation in the way you turn on a gene. You can imagine a binding site is different between these two [birds]. I think that's everyone's favorite example.…in my lab we study classic examples where binding sites differ.

Later in the interview, YR70 revealed,

> I'm trying to be very careful here, because I don't want to say the DNA differs between these same species. I just want to leave it as kind of like, the signal that comes in is perceived differently … I'm trying to think within my own work, at a particularly important trait, I've found differences in the genes that lead to that trait. So it's the same species, one type has some genes, the other type doesn't have those genes, so I don't want to say that there can be no variation within species at the gene level.

In this case, YR70 recognizes the importance of speaking carefully within the context of his or her own work and disciplinary focus, yet uses the shorthand word "gene" instead of the more precise term "allele" or perhaps is considering a genetically modified organism with a novel gene. Regardless of whether this respondent's imprecise word use was due to cognitive shorthand or conceptual misunderstanding, we would code this response as troublesome in the context of this study.

Despite these issues, we felt the Outgroup served as a valuable comparator to the Pre, Current, and Post groups in the nature of their explanations. Yet we recognize that the context and proximity of the groups for comparison needs to be well defined and bounded. Ideally, this approach for measuring threshold crossing could be used in a longitudinal way, interviewing the same individuals repeatedly over time, before and after they have experienced a particular curriculum pertaining to the threshold concept of interest.

Using the Threshold Concept Model to Inform Curriculum

Another aim of our study was to explore the utility of the threshold concept model to inform curricular change. Disciplines are typically bounded by a curriculum (e.g., biology curriculum as opposed to English rhetoric and language curriculum), but even within a discipline, a curriculum is often context specific, and the context influences the concepts, skills, and affective inputs and outcomes that can be expected of the curriculum. In this study, we examined 29 students who had experienced different amounts or "doses" of the same variation-enriched biology curriculum.

Beginning with discursive analysis, we found that novice students first expand their variation discourse while engaged in variation-enhanced course work, and then display conformity in language and parsimony, with explanations being similar for the Post group compared with the Outgroup. A vocabulary

ranging from 9 to 33 discipline-specific words (Figure 3) was adequate to achieve the level of mastery defined in this study using the binary coding scheme (Figure 4A) for evaluating at least one form of biological variation at the genetic, cellular, or organismal level. Even though some respondents far exceeded the nine-word benchmark (e.g., HQ34 and NN29 with 31 and 33 discipline-specific words for question 4, respectively), the average number of words uttered across all groups, including the Outgroup, was 13 words and included "DNA," "gene," "allele," "genotype," "code," "RNA," "expression," "protein," "phenotype," "selection," "environment," "variant," and "mutation." This finding is interesting and may helpfully circumscribe a vocabulary list that is more realistically learnable by students than the full dictionary of biology terminology that is typically presented within a course. While both NN29 (Current) and AJ19 (Post) were identified as having crossed the threshold in this study, NN29 used 33 discipline-specific words and AJ19 used 17 discipline-specific words to articulate their explanations. Although an expansive vocabulary within biology may help a student generate his or her identity as a biologist and may be perceived as an entry point into the biologists' community of practice, our study indicates that a strong handle on a few choice, fundamental vocabulary words and the use of that vocabulary in a precise, efficient, and effective way are sufficient for communicating meaningfully about variation within species. Furthermore, our findings complement a study on systems thinking by Dauer *et al.* (2013), who examined the change in students' gene-to-evolution models over time through a curriculum. These researchers found that, when students' models gained accuracy and complexity, they also were more parsimonious, requiring efficiencies in language and concepts as the systems they were attempting to represent became more complex.

We recognized discomfort and conflicting or disoriented reasoning when respondents attempted to explain how genetic and cellular variation leads to phenotypic variation, with acknowledgment by students when their thinking became confused. For some respondents, their self-doubt seemed disturbing, particularly for those in the Outgroup. Although the state of not knowing should be a common feeling for learners at all stages in the learning process, recognition of a liminal understanding summons self-dissatisfaction. As educators, we often critique the cognitive aspects of our students' troublesome knowledge (e.g., if they only understood this important concept, they could understand this other important concept) without recognizing or empathizing with the affective aspect of the liminal journey (Rattray, 2016). There are often very specific values or alternative commitments that have nothing to do with the pedagogy or the curriculum but nevertheless influence a learners' capacity to move forward and cross a threshold. Our task as educators is to make space and time for both the cognitive and affective shift and to create opportunities for points of multiple entry, as well as off-ramps and on-ramps, as students move through liminal space of a threshold concept–based curriculum. Creating time and having patience for moving through liminal space are not only important for threshold concepts, they are essential for promoting metacognition (Land *et al.*, 2014; Tanner, 2017). Traditional methods of assessment focus, by and large, on vocabulary and accuracy of understanding without recognizing important aspects such as liminality and integration, which are fundamental to learning threshold concepts. It would be more sensible to include alternative modes on our assessments that could examine students' tolerance to uncertainty, their comfort and confidence for embracing complexity in ill-structured problems with more than one solution, and their application and analysis of randomness. These ideas (e.g., randomness and uncertainty), which have been identified as candidate threshold concepts (Ross *et al.*, 2010), are at the core of conceptual understanding in biology (Garvin-Doxas and Klymkowsky, 2008; Fiedler *et al.*, 2017) and are embedded in the philosophy for the core competencies articulated in *Vision and Change* (AAAS, 2011).

At a time when risk-taking and uncertainty are discouraged in the classroom, when confirmatory investigations are the norm, and when test scores and fixed mind-sets guide educational decisions and outcomes, making time for the liminal space is most needed. Liminal space needs to be incorporated as an intentional aspect of the curriculum, with students being encouraged to sort knowns from unknowns, take risks and dive into their uncertainties, learn to think creatively and pose questions, generate novel connections, and be comfortable with not knowing. At the same time, being comfortable in a liminal space (as an instructor or a student) is challenging and requires patience and a different set of standards/expectations for describing success and mastery. Instead of factual content knowledge being most prized, in a threshold concepts–focused curriculum, the capacity to tolerate uncertainty, deal with messiness and complexity, think critically, pose questions, and problem solve given ill-structured problems would be emphasized.

Using the Threshold Concept Model to Understand Students' Explanations of Variation

Our remaining aim was to use the threshold concept model to examine students' capacity to observe, explain, and represent the basis of variation within species. Our results indicate that two threshold concept dimensions were particularly hard for students to achieve: 1) nontroublesome explanations and 2) integration of multiple biological scales in their explanations. Given that variation within species is so fundamental for understanding genetics and evolution, there have already been studies that examine students' explanations of variation in other contexts (Shtulman and Schulz, 2008; Nehm and Ridgway, 2011; Coley and Tanner, 2015). Here, we will relate our findings to these published results.

We find the majority of respondents' explanations of variation contain some inaccuracy or overapplication of intuitive reasoning. In respondents' explanations of the cellular differences among different birds, many resorted to overly intuitive reasoning. Specifically, among our sample, essentialist reasoning was the most commonly overused, followed by teleological reasoning. This fits with previous work, where both majors and nonmajors most frequently endorsed teleological misconception statements but most frequently used essentialist reasoning in their written justifications (Coley and Tanner, 2015). Previous literature also provides evidence that within-species variation is particularly counterintuitive. In one study, just under half of adults held essentialist beliefs that all members of a biological species are the same (Shtulman and Schulz, 2008). In another study, researchers described differences in evolutionary experts and novices in card-sorting tasks, problem-solving tasks,

and interviews (Nehm and Ridgway, 2011). They found that evolutionary novices were more likely to hold cognitive biases, such as teleological reasoning—which resembled that of children—while these biases were absent in experts. This is slightly different from our results, as we found that our Post and Outgroup respondents often demonstrated overly intuitive reasoning in their explanations. However, it is unclear whether this was due to fundamental misunderstanding or a cognitive shorthand to explain variation.

Apart from the overapplication of intuitive reasoning, students' explanations of how the contents of cells vary for different birds were also commonly troublesome due to inaccuracies. We found that students often conflated the meaning of "gene" and "allele" in their explanations, as we have observed previously (Batzli *et al.*, 2014). This is despite our emphasis on conveying the importance of precise language within the courses. Additionally, we found that respondents seemed to ritualize Mendelian thinking to the point where they tried to apply monohybrid crosses and Punnett square analysis to polygenic traits, even though our curriculum goes to some length to delineate that discrete traits are influenced predominantly by a single gene and continuous traits are influenced by a large number of genes. We were particularly struck by respondents' willingness (even within the Outgroup, as discussed earlier) to view all phenotypes related to pigmentations as completely discrete and monogenic, likely due to the considerable emphasis on pea and petunia flower color examples used in common genetics curricula. Consistent with previous research, these results suggest that variation within species is particularly challenging to accurately explain.

Using a threshold concepts model, we also observed that a majority of respondents failed to integrate biological scales in their explanations of how variation changed over one generation. Similar to our results, Speth *et al.* (2014) also found that introductory biology students struggled to integrate multiple biological scales in their structure–behavior–function models of the origin of variation. Notably, all of the members of our Outgroup demonstrated sufficient mastery with explanations that integrated phenotypic variation at the organismal level with two or more other scales of variation (genetic, population, environmental, etc.).

Together, our results suggest that, while students are quick to acquire discipline-specific terminology of variation, it takes them considerably longer to develop conceptual genetics knowledge and to integrate the concept of variation among biological scales in order to provide well-reasoned, accurate explanations. The threshold concept model was useful in helping us bring together information, much of which is supported by previous studies, and combine these dimensions toward a deep understanding of variation within species. Understanding of evolution more generally is even more fraught as a threshold concept, because it involves multiple thresholds converging (e.g., variation and randomness): an appreciation of variation within species is required to understand what evolutionary mechanisms act upon, while an appreciation of randomness is required to understand how certain evolutionary mechanisms function (such as genetic drift). While our analysis has focused on the former, recent work suggests that randomness is a measurable threshold concept that preservice teachers in particular struggle to achieve competence in (Fiedler *et al.*, 2017).

Limitations

In addition to the potential oversimplicity of coding already described, a major remaining limitation to the current study was its cross-sectional design. While efficient, this design was limited in its ability to measure transformation in a single student over time, which would have required a long-term longitudinal study. Further, given the time-intensive nature of qualitative data analysis, we were only able to examine students' explanations in one context (associated with *S. vulgaris* specimens) and were unable to include a large sample for each cohort, and thus our statistical power is limited by sample size. Future experimentation may aim to use these data to design more directed questions in several contexts that can be given to many students over time with greater efficiency. Finally, we recognize our basic approach of creating cutoff values to define "sufficient mastery" and threshold crossing may be creating inappropriate stringency or could be problematic if used for formal assessment. Therefore, we caution against an overly quantitative approach to parsing responses as "crossed" or "uncrossed," because a threshold does not seem absolute, but rather is "fuzzy" and influenced by context, identity, and values of the learner.

Application to Research, Pedagogy, and Practice

The main implication for instruction that we have taken away from studying the threshold concepts framework is the importance of deliberately making time in the curriculum for students to be in a liminal space and for us as instructors to empathize with the challenge of being uncertain. Particularly, our data suggest that it is particularly challenging for learners to explain variation accurately (i.e., nontroublesome) in a way that integrates biological scales (i.e., integrative). So, we find it especially important to implement curricular tasks that focus on the accuracy and integration of variation in a highly formative way. Below, we suggest two specific examples from the literature that may satisfy these parameters:

- In a lecture-based setting, students can generate models that follow molecular origins to evolutionary outcomes (so called gene-to-evolution models), which they iteratively revise throughout the semester with the use of feedback from peers and instructors (Dauer *et al.,* 2013; Speth *et al.,* 2014). This provides students with the task of integrating biological scales related to variation but also reveals inaccuracies in their thinking. And, doing so in an iterative, formative way, rich with feedback, allows students the opportunity to accept their uncertainty and learn from it.
- In a lab-based setting, students can give ungraded feedback presentations on research plans related to variation in which they receive input from their instructors and peers on their plans to observe, explain, predict, and measure variation, similar to a research lab meeting (Batzli *et al.,* 2014, 2018). Through this context, students are provided with feedback about the accuracy (i.e., troublesomeness) and missing links (i.e., integrative nature) of their explanations. Further, because all students give feedback presentations in the same period, they share the experience of a liminal space together.

No matter the specific implementation plan, we find it especially important that we as instructors create an environment

where uncertainty is accepted and treated as a useful step toward deep understanding.

SUMMARY AND CONCLUSION

In summary, through our analysis, we found evidence for threshold concept dimensions (i.e., troublesome, integrative, discursive, and liminal) in respondents' explanations of variation within species (aim 1). On the basis of our rubrics, we found that only the integrative, liminal, and troublesome dimensions discriminated among different cross-sectional groups (aim 2). The discursive dimension was achieved by the majority of respondents in all cross-sectional groups, while adequate accuracy (i.e., lack of troublesomeness) and adequate integration of biological scales were achieved by the majority of Outgroup respondents, but not the curricular groups. While trying to detect threshold crossing (aim 3), we realized that, by definition, the liminal space and the threshold are hard to define, which also makes the exact point of threshold crossing hard to define. While we believed that we detected threshold crossing in our study, we realize that our approach and interpretation may be overly simplistic. Instead, we conclude that pinpointing the exact moment of threshold crossing is not as important as identifying the lessons learned from the threshold concept framework as applied to curricular design. We find resolve to focus on how to destigmatize the liminal space so that learners do not reside there but, with instructors' guidance, can work through it into a postliminal space.

ACKNOWLEDGMENTS

We thank Jeff Hardin and the Department of Integrative Biology at the University of Wisconsin–Madison for funding through the Michael Guyer Post-Doctoral Fellowship. We also thank students in the Biocore program and the UW Zoological Museum for lending *S. vulgaris* specimens. We are grateful for interview transcription by Maddie Batzli and many conversations with colleagues, including Michelle Harris, Emily Jobe, and Robin Forbes-Lorman, whose insights and questions have been invaluable.

REFERENCES

American Association for the Advancement of Science (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Batzli, J. M. (2005). Points of view: A survey of survey courses: Are they effective? A unique approach? Four semesters of Biology Core Curriculum. *Cell Biology Education*, *4*, 125–128.

Batzli, J. M., Harris, M. A., & McGee, S. A. (2018). It takes time: Learning process of science through an integrative, multi-semester lab curriculum. Tested studies for laboratory teaching. In McMahon, K. (Ed.), *Proceedings of the Association for Biology Laboratory Education*, *39*(21), www.ableweb.org/volumes/abstract/?ResourceId=%201350

Batzli, J. M., Knight, J. K., Hartley, L. M., Maskiewicz, A. C., & Desy, E. A. (2016). Crossing the threshold: Bringing biological variation to the foreground. *CBE—Life Sciences Education*, *15*, es9.

Batzli, J. M., Smith, A. R., Williams, P. H., McGee, S. A., Dósa, K., & Pfammatter, J. (2014). Beyond Punnett squares: Student word association and explanations of phenotypic variation through an integrative quantitative genetics unit investigating anthocyanin inheritance and expression in *Brassica rapa* Fast Plants. *CBE—Life Sciences Education*, *13*, 410–424.

Coley, J. D., & Tanner, K. D. (2012). Common origins of diverse misconceptions: Cognitive principles and the development of biology thinking. *CBE—Life Sciences Education*, *11*, 209–215.

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE—Life Sciences Education*, *14*, ar8.

Cope, C., & Staehr, L. (2008). Improving student learning about a threshold concept in the IS discipline. *Informing Science, 11*, 349–364.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (*CPRE Research Report # RR-63*). New York: Consortium for Policy Research in Education, Center on Continuous Instructional Improvement Teachers College, Columbia University.

Dancy, M. H., & Beichner, R. J. (2002). But are they learning? Getting started in classroom evaluation. *CBE—Life Sciences Education*, *1*, 87–94.

Dauer, J. T., Momsen, J. L., Bray Speth, E., Makohon-Moore, S. C., & Long, T. M. (2013). Analyzing change in students' gene-to-evolution models in college-level introductory biology. *Journal of Research in Science Teaching*, *50*, 639–659.

Duncan, R. G., & Rivet, A. E. (2013). Science learning progressions. *Science*, *339*, 396–397.

Dye, K. M., & Stanton, J. D. (2017). Metacognition in upper-division biology students: Awareness does not always lead to control. *CBE—Life Sciences Education*, *16*(2), ar31.

Emmons, N. A., & Kelemen, D. A. (2015). Young children's acceptance of within species variation: Implications for essentialism and teaching evolution. *Journal of Experimental Child Psychology*, *139*, 148–160.

Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. doi: 10.1037/0033-295X.87.3.215

Fiedler, D., Tröbst, S., & Harms, U. (2017). University students' conceptual knowledge of randomness and probability in the contexts of evolution and mathematics. *CBE—Life Sciences Education*, *16*(2). https://doi.org/10.1187/cbe.16-07-0230

Flanigan, M. (2018). *Threshold Concepts: Undergraduate Teaching, Postgraduate Training and Professional Development: A Short Introduction and Bibliography*. Retrieved June 20, 2018, from www.ee.ucl.ac.uk/~mflanaga/thresholds.html

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE—Life Sciences Education*, *7*, 227–233.

Green, D. A., Loertscher, J., Minderhout, V., & Lewis, J. E. (2017). For want of a better word: Unlocking threshold concepts in natural sciences with a key from the humanities? *Higher Education Research & Development*, *36*(7), 1401–1417.

Halmo, S. M., Sensibaugh, C. A., Bhatia, K. S., Howell, A., Ferryanto, E. P., Choe, B., … Lemons, P. P. (2018). Student difficulties during structure-function problem solving. *Biochemistry and Molecular Biology Education*, *46*(5), 453–463.

Land, R., Cousin, G., Meyer, J. H. F., & Davies, P. (2005). Threshold concepts and troublesome knowledge: Implications for course design. In Rust, C. (Ed.), *Improving student learning diversity and inclusivity* (pp. 53–64). Oxford, UK: Oxford Centre for Staff and Learning Development.

Land, R., Meyer, J. H. F., & Baillie, C. (2010). Editors' preface: Threshold concepts and transformational learning (pp. ix–xlii). In Land, R., Meyer, J. H. F., & Baillie, C. (Eds.), *Threshold concepts and transformational learning*. Rotterdam: Sense Publishers.

Land, R., Rattray, J., & Vivian, P. (2014). Learning in the liminal space: A semiotic approach to threshold concepts. *Higher Education*, *67*, 199–217

McCartney, R., Boustedt, J., Eckerdal, A., Moström, J. E., Sanders, K., Thomas, L., & Zander, C. (2009). Liminal spaces and learning computing. *European Journal of Engineering Education*, *34*, 383–391.

Meyer, J. H. F., & Land, R. (2003). *Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practicing within the disciplines* (Occasional Report 4). Edinburgh, UK: Enhancing Teaching-Learning Environments in Undergraduate Courses Project, Higher and Community Education, School of Education, University of Edinburgh.

Meyer, J. H. F., & Land, R. (2005). Threshold concepts and troublesome knowledge (2): Epistemological considerations and a conceptual framework for teaching and learning. *Higher Education*, *49*, 373–388.

Meyer, J. H. F., & Land, R. (2006). Threshold concepts and troublesome knowledge: Issues of liminality. In Meyer, J. H. F., & Land, R. (Eds.), *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge* (pp. 19–32). London: Routledge.

Meyer, J. H. F., Land, R., & Davies, P. (2006). Implications of threshold concepts for course design and evaluation. In Meyer, J. H. F., & Land, R. (Eds.),

Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge (pp. 195–206). London: Routledge.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, *46*, 675–698.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *American Biology Teacher*, *74*(2), 92–98.

Nehm, R. H., & Ridgway, J. (2011). What do experts and novices "see" in evolutionary problems? *Evolution: Education and Outreach*, *4*, 666–679.

Nicola-Richmond, K., Pépin, G., Larkin, H., & Taylor, C. (2018). Threshold concepts in higher education: A synthesis of the literature relating to measurement of threshold crossing. *Higher Education Research*, *37*, 101–114.

O'Donnell, R. (2010). *A critique of the threshold concepts hypothesis and an application in economics* (Working Paper No. 164). Sydney, Australia: Finance Discipline Group, UTS Business School, University of Technology.

Perkins, D. (2006). Constructivism and troublesome knowledge. In Meyer, J. H. F., & Land, R. (Eds.), *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge* (pp. 33–47). London: Routledge.

Rattray, J. (2016). Affective dimension of liminality. In Meyer, J. H. F., Land, R., & Flanagan, M. T. (Eds.), *Threshold concepts in practice* (pp. 67–76). Rotterdam: Sense Publishers.

Richard, M., Coley, J. D., & Tanner, K. D. (2017). Investigating undergraduate students' use of intuitive reasoning and evolutionary knowledge in explanations of antibiotic resistance. *CBE—Life Sciences Education*, *16*, ar55.

Ross, P. M., Taylor, C. E., Hughes, C., Kofod, M., Whitaker, N., Lutze-Mann, L., & Tzioumis, V. (2010). Threshold concepts: Challenging the way we think, teach and learn in biology. In Land, R., Meyer, J. H. F., & Baillie, C. (Eds.), *Threshold concepts and transformational learning* (pp. 165–177). Rotterdam, Netherlands: Sense Publishers.

Rowbottom, D. P. (2007). Demystifying threshold concepts. *Journal of Philosophy of Education*, *41*(2), 263–270.

Shanahan, M., Foster, G., & Meyer, J. (2006). Operationalising a threshold concept in economics: A pilot study using multiple choice questions on opportunity cost. *International Review of Economics Education*, *5*(2), 29–57.

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

Smith, M. U. (2010a). Current status of research in teaching and learning evolution. I. Philosophical/epistemological issues. *Science Education*, *19*, 523–538.

Smith, M. U. (2010b). Current status of research in teaching and learning evolution. II. Pedagogical issues. *Science Education*, *19*, 539–571.

Speth, E. B., Shaw, N., Momsen, J., Reinagel, A., Le, P., Taqieddin, R., & Long, T. (2014). Introductory biology students' conceptual models and explanations of the origin of variation. *CBE—Life Sciences Education*, *13*, 529–539.

Stanton, J. D., Neider, X. N., Gallegos, I. J., & Clark, N. C. (2015). Differences in metacognitive regulation in introductory biology students: When prompts are not enough. *CBE—Life Sciences Education*, *14*(2), ar15.

Tanner, K. (2017). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120. https://doi.org/10.1187/cbe.12-03-0033

Taylor, C. (2006). Threshold concepts in biology: Do they fit the definition? In Meyer, J. H. F., & Land, R. (Eds.), *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge* (pp. 87–99). London: Routledge.

White, J. S., & Maskiewicz, A. C. (2014). Understanding cellular respiration in terms of matter and energy within ecosystems. *American Biology Teacher*, *76*, 408–414.