Is Active Learning Enough? The Contributions of Misconception-Focused Instruction and Active-Learning Dosage on Student Learning of Evolution

ROSS H. NEHM, STEPHEN J. FINCH, AND GENA C. SBEGLIA

Prior studies of active learning (AL) efficacy have typically lacked dosage designs (e.g., varying intensities rather than simple presence or absence) or specification of whether misconceptions were part of the instructional treatments. In this study, we examine the extent to which different doses of AL (approximately 10%, 15%, 20%, 36% of unit time), doses of misconception-focused instruction (MFI; approximately 0%, 8%, 11%, 13%), and their intersections affect evolution learning. A quantitative, quasiexperimental study (N > 1500 undergraduates) was conducted using a pretest, posttest, delayed posttest design with multiple validated measures of evolution understanding. The student background variables (e.g., binary sex, race or ethnicity), evolution acceptance, and prior coursework were controlled. The results of hierarchical linear and logistic models indicated that higher doses of AL and MFI were associated with significantly larger knowledge and abstract reasoning gains and misconception declines. MFI produced significant learning above and beyond AL. Explicit misconception treatments, coupled with AL, should be explored in more areas of life science education.

Keywords: active learning, misconceptions, learning, evolution, instruction

Evolutionary theory is considered to be one of the most valuable scientific theories within the life sciences and across many other disciplines (Cosmides 1989, Dosi and Nelson 1994, Rutledge and Warden 2000). Because of this, the theory of evolution has been appropriately highlighted as a central component of science literacy more broadly, and many policy documents have urged educators to help students throughout the educational hierarchy understand this core biological concept (Brewer and Smith 2011, NRC 2012). Evolution serves an additional function: It unites the diverse array of biological concepts, systems, and subdisciplines into a coherent conceptual structure (NRC 1958, Brewer and Smith 2011). Despite such importance, the goal of instilling a robust understanding of evolution has been hard to achieve; decades of research on students from primary (Brown et al. 2020) to graduate school (Gregory 2007) indicate that students misunderstand basic features and mechanisms of evolutionary change and harbor a diverse array of nonnormative beliefs colloquially referred to as *misconceptions* (for reviews, see Gregory 2007, Nehm and Reilly 2007).

Misconceptions about the natural world are common, and many tend to persist for long periods of time. Constructivist learning theories emphasize that individuals begin to build their understanding of the natural world prior to and outside of formal schooling at a young age (NRC 2001). Understanding generated by these less formal learning experiences tends to differ from understanding generated by sustained and principled scientific effort (Vosniadou and Brewer 1992). For example, many students falsely believe that our planet is closer to the sun in summer than in winter because, in everyday situations, temperature is related to the distance from a heat source (Caravita and Halldén 1994, Hammer 1996). To students, their own ways of thinking are more plausible and less effortful than scientific practices. Students (and many teachers) remain satisfied with their personally constructed models of how the natural world works (NRC 2001).

BioScience 72: 1105–1117.
https://doi.org/10.1093/biosci/biac073 Advance Access publication 7 September 2022

Table 1. Limitations of prior studies of active learning of evolution and corresponding approaches for addressing these limitations.

Limitations of prior studies	How limitation can be addressed
Instructor self-reports of time devoted to active learning and misconceptions.	Use instruments (e.g., COPUS) to independently measure instructional behaviors via direct observation along with reliability tests. Independent verification of time devoted to misconceptions.
Measures of learning outcomes only used parts of a validated[a] instrument or assessment items lacking robust validation.	Use multiple complete, validated assessment instruments adopting different perspectives to measure learning of the same constructs.
No control of evolution acceptance in measures of evolution learning.	Use validated measures of acceptance (e.g., I-SEA) as a control variable in analyses.
Atypical instructors (i.e., BER faculty) used to implement instructional innovations, or different instructors used to compare conditions.	Study impact with biologists with no formal education training or DBER expertise and typical (average) teaching evaluations. Study the same instructor across course iterations.
Delayed posttesting of learning outcomes unclear or absent.	Use a pretest, posttest, and delayed posttest design.
Lack of control of a wide array of relevant demographic and background variables.	Incorporate background variables (e.g., binary sex, race/ethnicity, EL status, prior biology courses) as control variables.
Sample sizes moderate to small with unreported participation rates.	Collect large student samples and report participation rates.
Lack of replication of comparison differences	Examine replicates of findings (e.g., across instruments, semesters, instructors)
Binary (e.g., presence/absence) versus dosage designs	Incorporate dosage (or intensity) of classroom activities/interventions

Abbreviations: BER, biology education research; DBER, discipline-based education research. [a]Although instrument validation is a continuous process, we use the phrase *validated instruments* as opposed to *instruments with validity evidence* throughout this table to conserve space.

Misconceptions have been discussed in the scientific literature for at least 100 years (e.g., Osbern 1922) and are formally defined as "understandings or explanations that differ from what is known to be scientifically correct" (NRC 2012, p. 58). In education, *misconception* is a term that has changed in meaning and has encompassed many different forms of ideation, including universally false beliefs (e.g., vitalist forces exist and cause evolutionary change), normative ideas that become false beliefs as a result of overextension (e.g., mutations are always caused by the abiotic environment; needs cause goal-oriented change), and unproductive combinations of normative ideas and false beliefs. Misconceptions have also been differentiated by their cognitive stability (e.g., theory-like permanence versus spontaneous mental assembly), timing (e.g., whether they occur before or after formal instruction on the topic), and value for learning (e.g., as necessary stepping stones to normative understanding versus problematic cognitive baggage). Although misconceptions are a category of ideation composed of diverse entities and labeled in different ways, they are united by the role they often play: as barriers to normative scientific understanding (e.g., evolution; Gregory 2007, Kampourakis 2020).

Constructivist theories suggest that explicitly engaging with students' models of the world is essential for meaningful learning to occur (NRC 2001). Empirical research indicates that student misconceptions are often impervious to traditional (nonconstructivist) teaching (Ausubel 1968, Guzzetti et al. 1993). Generating cognitive dissonance in learners through active engagement during experiences that conflict with or challenge their thinking appears to be an essential feature of successful interventions (see Guzzetti et al. 1993 for a meta-analysis). Many different active learning (AL) approaches may be used to generate such dissonance (e.g., peer-led discussions, labs, refutational worksheets; see Guzzetti et al. 1993 for numerous examples).

In a large amount of literature, student misconceptions about evolution have been catalogued (e.g., Gregory 2007, Nehm 2018; see supplemental table S1), and many assessment tools have incorporated misconceptions into items intended to measure understanding (Anderson et al. 2002, Nehm et al. 2012, Kalinowski et al. 2016). Documenting student misconceptions and incorporating them into measurement instruments have laid the groundwork for developing and testing interventions for improving student learning of evolution and other areas of biology (e.g., Nehm and Reilly 2007, Andrews et al. 2011, Beardsley et al. 2012).

Intervention studies remain comparatively rare in evolution education, however, and are often small scale (e.g., one or two classes), lack robust research designs (e.g., no comparison groups, univariate designs), do not control for background variables (e.g., sex, race or ethnicity, evolution acceptance), use education-trained instructors (versus biologists; see Andrews et al. 2011), and measure change using single instruments (see table 1; NRC 2014). Some studies have adopted only a few of these quality control criteria (e.g., Nehm and Reilly 2007), whereas others have adopted more (e.g., Andrews et al. 2011). Additional studies that tackle these methodological limitations (table 1) are urgently needed to inform practice; the current body of intervention studies has produced ambiguous findings about evolution learning (e.g., Nehm and Reilly 2007, Andrews et al. 2011).

There is remarkably little large-scale work empirically testing whether explicit attention to evolution misconceptions is beneficial, neutral, or harmful (e.g., by perpetuating problematic ideas). Indeed, in undergraduate education, although, many biology textbooks and curricula appear to focus on teaching only normative scientific ideas even

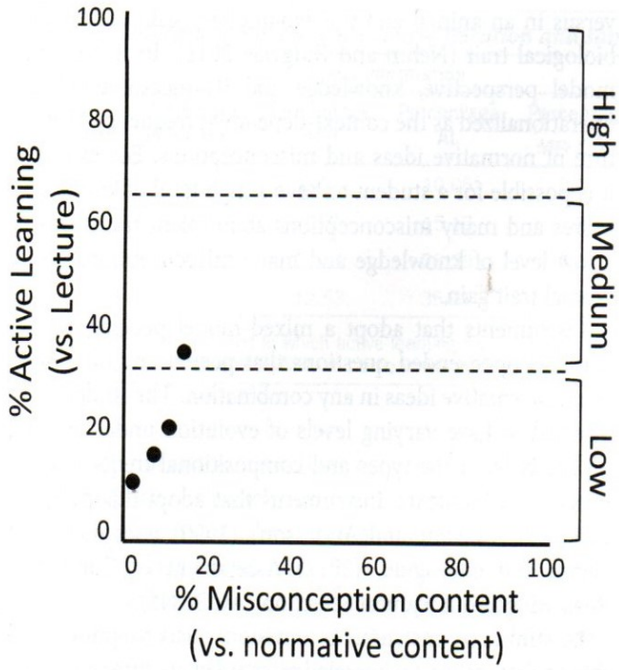

Figure 1. Active learning and attention to misconceptions situated within a gradient or dosage framework (as a percentage). The low, medium, and high designations on the right hand side reflect the active learning intensity categories outlined by Theobald and colleagues (2020). The dots indicate the doses used in the present study. Importantly, dosages of active learning in this study were calculated relative to the overall time in the evolution unit (i.e., time spent on evolution content both in and out of class), unlike Theobald and colleagues' (2020) estimates of active learning, which seem to consider only in class time.

though this approach is at odds with constructivist models of meaningful learning (NRC 2001). Despite a century of discussion of evolution misconceptions in undergraduate education (e.g., Osbern 1922, Gregory 2007), it remains unknown whether and to what extent undergraduate biology courses are addressing these misconceptions.

In contrast to misconception-focused pedagogies, the biology education community has focused much more attention on AL and its impact on student outcomes (Freeman et al. 2014). Initial studies of this approach to learning have mostly been focused on the presence or absence of AL. Freeman and colleagues (2014) proposed that varying intensities of AL should be an avenue of future work. Recently, Theobald and colleagues (2020) used this approach and reported that high-intensity AL narrowed exam score achievement gaps between minoritized and nonminoritized students (44% for high intensity versus 22% for low intensity) and reduced passing rate disparities (76% for high intensity versus 16% for low intensity). These findings support the efficacy of AL, as well as the need to move away from binary (e.g., presence or absence) perspectives on classroom activities and interventions.

In contrast to other biology domains, the impacts of AL on evolution outcomes have shown that AL may (e.g., Nehm and Reilly 2007) or may not (Andrews et al. 2011) lead to improved learning outcomes. Given such ambiguity, as well as the lack of robust large-scale studies investigating the impact of misconception-focused instruction (MFI) and numerous methodological weaknesses of prior work (table 1), we aim in this study to examine the outcomes of evolution instruction that varies along two primary axes: the extent to which misconceptions are addressed and the extent to which AL is implemented (see figure 1). By incorporating varying doses of AL and MFI, we build on the insights and advances of Freeman and colleagues (2014) and Theobald and colleagues (2020) described above.

The present study took place in a large introductory biology class taught by a typical faculty member (a biologist with no formal education preparation) using a quantitative, quasiexperimental, pretest, posttest, delayed posttest design with large student samples, multiple control variables, varying treatment doses, and measures that conceptualize evolution learning in multiple ways. The inclusion of multiple measures is an important feature of this study because Freeman and colleagues' (2014) meta-analysis results showed that different types of assessments (e.g., concept inventories versus instructor-designed tests) led to different inferences about the impact of AL on student outcomes.

Three research questions guided the study: Is the amount of evolution learning conditional on the amount (or dose) of AL? Is the amount of evolution learning conditional on the amount (or dose) of MFI above and beyond the contributions of AL alone? And do measurement instruments adopting different perspectives on learning support the same inferences about the roles of AL and misconceptions in evolution learning?

Approach to answering our research questions

Study setting. The introductory biology course in which this study was conducted was characterized by many common constraints at research universities: large size (typically 500 or greater), lecture-hall-style rooms, and pressure to cover as many concepts as possible. The prerequisites were high school biology and freshman-level math. The course was taught twice a week for 80 minutes, and the content was aligned with five core concepts of biological literacy (Brewer and Smith 2011). Four iterations of the course (A–D) were studied (figure 2, table 2). In each iteration, an evolution unit was taught in the first half of the semester by the same biologist instructor, with fewer than 20 years of teaching experience but no formal preparation or training in biology education or AL. The instructor did not design the study or the materials, nor were they aware of the study's goals or hypotheses. The total time devoted to this unit in each course iteration is shown in table 2 (see also section 1 of the supplemental material). In iteration D, the course was split into two sections (approximately 250 students each), one of

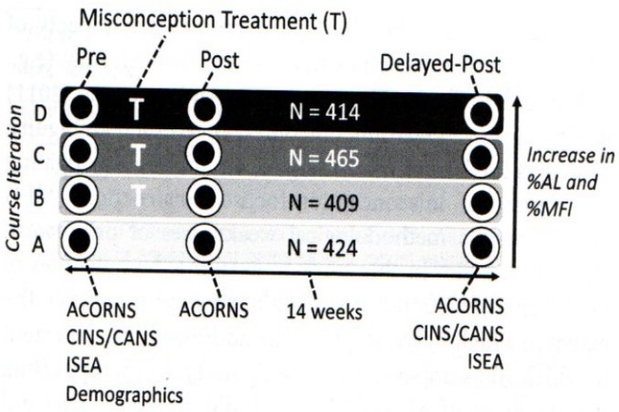

Figure 2. Summary of the study design, timing, and measures. *Posttest* ***refers to the assessment that occurred following evolution instruction, and*** *delayed posttest* ***refers to the assessment that occurred at the end of the semester (7–11 weeks after evolution instruction). See the text for measurement instrument details (e.g., ACORNS). Abbreviations: AL, active learning; MFI, misconception-focused instruction.***

which was taught by a new instructor with 1 year of teaching experience and minimal preparation in biology education and AL. The data from this instructor were included only to test the replicability of the findings across instructors (see section 4.1 of the supplemental material for more information about these procedures).

Measurement instruments. Different instruments reflect somewhat different perspectives on the relationship between misconceptions and understanding, and as a result, they have implications for how to measure evolution learning. To simplify, one perspective operationalizes evolution knowledge as the converse of misconceptions (what may be termed an *either-or* model). In this perspective, students are asked questions that include one normative answer option and a series of misconception options. The students are required to choose only one of the options, and if they choose a misconception, then they are inferred to lack understanding of the concept that the question targets (e.g., differential survival). The total number of normative statements selected is calculated to establish a knowledge measure. Instruments that could be considered to score student understanding in this way include the Conceptual Inventory of Natural Selection (CINS; Anderson et al. 2002) and the Conceptual Assessment of Natural Selection (CANS; Kalinowski et al. 2016).

A slightly different perspective on the relationship between misconceptions and knowledge assumes that these different ideas coexist in students' minds and that using one idea does not preclude the use of another. Different types and combinations of ideas (even those appearing contradictory from an expert perspective) may be activated or repressed on the basis of the reasoning situation. Examples of different reasoning situations include evolution in a plant versus in an animal and the evolutionary gain or loss of a biological trait (Nehm and Ridgway 2011). In this mixed-model perspective, knowledge and its measurement are operationalized as the context-dependent frequency or mixture of normative ideas and misconceptions. For example, it is possible for a student to have a high level of knowledge scores and many misconceptions about plant trait loss but a low level of knowledge and many misconceptions about animal trait gain.

Instruments that adopt a mixed-model perspective ask students open-ended questions that permit any normative or nonnormative ideas in any combination. The students are inferred to have varying levels of evolution understanding on the basis of the types and compositional frequencies of ideas across contexts. Instruments that adopt this perspective include Bishop and Anderson's (1990) assessment and Nehm and colleagues' (2012) Assessment of Contextual Reasoning about Natural Selection (ACORNS).

In summary, instruments envision misconceptions and their relationship to knowledge somewhat differently, but they all consider misconceptions to be an important aspect of assessment.

Three previously published instruments with substantial validity evidence were used to measure knowledge of evolution and were also found to generate valid inferences in the local sample. The first two, CINS (Anderson et al. 2002) and the CANS (Kalinowski et al. 2016), may be considered either-or knowledge measures. The CINS measures 10 concepts using 20 multiple-choice items; higher scores are intended to indicate more knowledge. Each CINS item has one correct response and several misconception distractors. Although the CINS has psychometric problems at a fine-grained level, Nehm and Schonfeld (2008) reported that the instrument generates valid inferences about overall evolutionary knowledge using total scores. After our study began, the CANS (Kalinowski et al. 2016) was developed. To establish whether the CANS would corroborate patterns produced using the CINS, both instruments were used in two semesters. Like the CINS, the CANS adopts an either-or measurement model, with one normative idea and multiple misconception distractors for each item. The CANS has 24 multiple-choice items, and higher scores are intended to indicate more evolution knowledge.

The third instrument used to study evolution knowledge was the ACORNS (Nehm et al. 2012). The ACORNS is a constructed response instrument that measures three aspects of evolution understanding: evolution knowledge (i.e., normative ideas), evolution misconceptions, and the coherence of evolution knowledge (i.e., consistency across evolution problem types). Validity and reliability evidence has been gathered for this measurement in comparable undergraduate settings (e.g., Nehm et al. 2012, Opfer et al. 2012, Beggrow et al. 2014). The ACORNS items were designed to vary in features known to affect novices but not experts (e.g., plant versus animal; Nehm and Ridgway 2011). Unlike the CINS and the CANS, the ACORNS permits the

Table 2. Information about each course iteration and sample information for each course iteration.

	Class information				Sample information				
Course iteration	Unit time (in minutes)	Percentage *MFI*	Percentage AL	Percentage *MFI* ∩ *AL*	Sample size	Percentage Participation	Percentage female	Percentage URM	Percentage no prior bio
A	756	0.00	10.05	0.00	424	79	61	21	25
B	714	7.70	15.13	1.96	409	78	60	21	26
C	907	11.47	20.29	6.84	465	90	58	25	31
D	734	12.53	35.97	8.92	414	83	54	17	25

Abbreviations: AL, unit time in which active learning occurred; AL ∩ MFI, unit time in which active learning and misconception-focused instruction overlapped (i.e., intersected); MFI, unit time in which misconception-focused instruction occurred.

use of both normative ideas (core concepts) and naive ideas (misconceptions) in each response. Two ACORNS items were used: one about animal trait gain and one about plant trait loss (supplemental table S2). Each item was scored separately and given a score of either 0 or 1 for each of three core concepts (variation, heritability, and differential survival; max = 3 points per item). The response was also scored for whether or not misconceptions (i.e., adapt or acclimation, needs or goals, use or disuse inheritance) were present (1, present; 0, absent). On the basis of the pattern of core concepts and misconceptions found in each ACORNS response, students were assigned a scientific model consistency (MODC) score, which indicates whether normative ideas alone were used across responses at a single testing time point (MODC = 1) or not (MODC = 0). Responses lacking core concepts or misconceptions (e.g., repeating the question, mentioning irrelevant information) were coded as *no model* (MODC = 0). ACORNS responses across the four course iterations were scored using the machine-learning-based tool EvoGrader (see Moharreri et al. 2014 for validity details and human–computer agreements), eliminating common problems with human scoring drift and inconsistency. For further information on the ACORNS, see Nehm and colleagues (2012) and Nehm (2018).

Given that knowledge and acceptance of evolution have been found to be related to some extent, controlling for acceptance in studies of learning is warranted but frequently lacking (e.g., Nehm and Reilly 2007, Andrews et al. 2011). Evolution acceptance was measured using the I-SEA (the Inventory of Student Acceptance of Evolution; Nadelson and Southerland 2012). The I-SEA is a 24-item Likert-scale instrument with five answer options. The instrument has three subscales: microevolution, macroevolution, and human evolution. Validity and reliability evidence has been gathered for measurement of evolution acceptance in undergraduate settings (e.g., Nadelson and Southerland 2012, Sbeglia and Nehm 2018, 2019).

Active learning was operationalized following Stains and colleagues (2018) and measured using the Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith et al. 2013). The COPUS is a published instrument with substantial validity evidence. It is designed to collect observational data about the behaviors of students and instructors in undergraduate classrooms (Smith et al. 2013). Trained COPUS raters score 25 specifically defined behaviors as either present or absent at 2-minute intervals throughout the classes studied. Several authors have used COPUS behaviors characteristic of student-centered teaching as measures of AL (e.g., Stains et al. 2018, Sbeglia et al. 2021). COPUS data were generated by three observers certified to conduct COPUS observations by an expert evaluator (Michelle K. Smith, one of the COPUS developers) and achieved Cohen's kappa interrater scores above .80. Further description about the measurement of AL using the COPUS is available in section 2 of the supplemental material.

Sample. The students were given the opportunity to complete the ACORNS, the CINS, and the CANS (if relevant) during the first week of the course and at the end of the semester. These surveys were voluntary, and the participants received extra credit for complete responses. The students also completed the ACORNS during the first midterm, which counted toward the midterm grade (figure 2). Prior work in a comparable sample has shown that there were no meaningful differences in ACORNS scores between required and voluntary assessment conditions (Sbeglia and Nehm 2022). At all assessment time points, the students were administered an ACORNS plant loss item and an animal gain item (see above).

The students reported prior biology coursework (i.e., no prior biology, AP biology, one introductory biology course, two or more introductory biology courses), binary sex, and race or ethnicity (i.e., White, Asian, underrepresented minority [URM]; table 2). URM students were classified as those who identified as Black or African American, American Indian or Alaska Native, Hispanic of any race, or Native Hawaiian or other Pacific Islander (see table 2). Data were also collected on English language learner status, but this variable was insignificant and was removed to simplify the models.

Of the 1955 students enrolled in the course across the four semesters studied, 1629 consented to participate (see table 2 for participation rates). Overall, 7.5% of the consenting students had at least one missing data point (the maximum missing data for any one variable was 2%; see supplemental table S3), leaving 1507 students who completed all relevant

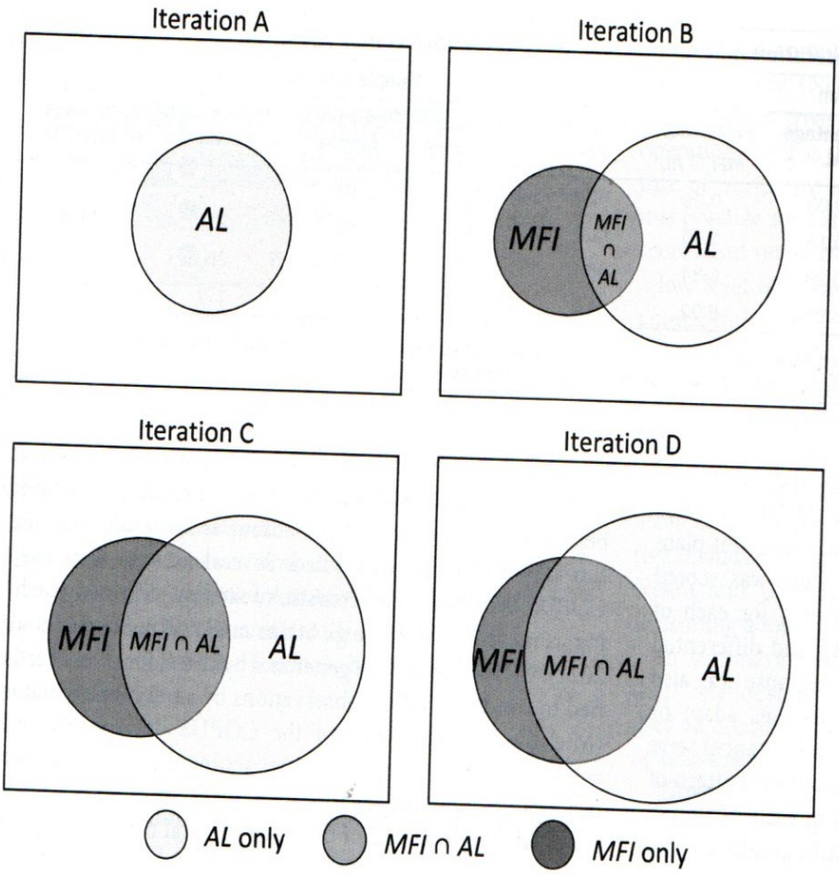

Figure 3. Relative amounts of active learning (AL) and misconception-focused instruction (MFI), and their intersections in each course iteration (iteration A–D). Iteration A had the lowest amount of AL and no MFI. The relative amount of AL increased from iteration A to D (also see table 1). Course iterations B–D had both AL and MFI with different amounts of intersection. The term intersection and its associated symbol (∩) are from set theory. In the intersection of two sets, such as hypothetical sets A and B, every element of A ∩ B belongs to both A and B.

assessments and were included in the analyses ($N = 1507$). In a recent study in different semesters of the same course (Sbeglia and Nehm 2022), the percentages of participating minoritized and female students were comparable to their distribution in the class. This finding suggests that participation bias is probably not significant. The study was approved by the university's institutional review board (protocol no. 504,271) and was classified as *not human subjects research*. The procedures outlined in the present article are in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Declaration of Helsinki 1975. Raw data and model R code are available from the corresponding author on emailed request.

Study design. Four consecutive 14-week spring semester iterations (i.e., iterations A–D; figure 2) were studied. Only spring semesters were chosen to reduce unintended variability in the data. In particular, the fall semester at this institution tends to enroll students with lower high school GPAs (91.8 versus 93.0 in spring), fewer arts and sciences majors (76.2% versus 84.5% in spring), more transfer students (8.7% versus 4.7% in spring), more commuters (15.3% versus 11.5% in spring), and more students in their first term at the university (18.1% versus 4.7% in spring; Bertolini et al. 2021). The precourse demographic variables and instrument scores were used to control for between-group differences. The dosages of AL and MFI increased across course iterations (table 2, figure 3).

MFI can be defined as student engagement with content that explicitly attends to "understandings or explanations that differ from what is known to be scientifically correct" (NRC 2012, p. 58). It therefore involves some form of pedagogy but specific content. The materials used as part of MFI in this study differed somewhat among the four course iterations: absent, present with varying dosage, and different types of misconception exposures (e.g., homework, in-class lecture, in-class activity, individual work, group work; supplemental table S4). Course iteration A lacked MFI altogether, whereas in the remaining three iterations, instructional materials designed to explicitly address student misconceptions (supplemental table S1) were implemented between the pretest and the posttest (figure 2).

The percentage of MFI in each course iteration was calculated by documenting the start and end times of MFI to the nearest minute using archived Echo recordings of each semester. The total time spent in MFI was then divided by the total time (both in and out of class) in the evolution unit overall. The percentage of time devoted to AL in each course iteration was calculated by dividing the total number of 2-minute COPUS intervals within which an AL-aligned behavior occurred by the total number of 2-minute intervals in the unit.

The percentage of time that the students experienced MFI and AL simultaneously (see figure 3 for visualization) was calculated by summing the number of minutes scored as both MFI and AL and dividing by the total time in the evolution unit.

Model description. Hierarchical linear and logistic models were used to determine the extent to which AL and MFI (both at varying doses) contributed to evolution learning (if at all). The statistical package lme4 in R (Bates et al. 2022) was used for all of the models. For outcome variables with three timepoints (see figure 2), the hierarchical linear and logistic model was built as a piecewise slope

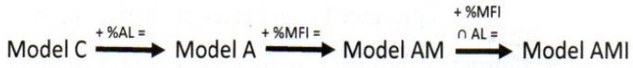

Figure 4. Visualization of the model comparison design used in this study.

model in which one slope was modeled from the pretest to the posttest (time 1) and a second was modeled from the posttest to the delayed posttest (time 2). The coefficient of each time period represents the linear rate of change (i.e., the slope) of the outcome variable across that time period (Hoffman 2015). In line with Theobald and Freeman (2014), we controlled for four student-level variables: prior biology courses, binary sex, race or ethnicity, and pretest I-SEA score. Each observation was also characterized by three continuous, time-invariant treatment variables: the percentage of AL, the percentage of MFI, and the percentage of AL ∩ MFI (table 2, see figure 3 for a visual representation). In addition, the treatment variables were modeled as having an interaction with the piecewise slope for time 1 and (for all outcomes but the CINS) time 2, allowing testing of whether the slope of evolution learning through time was conditional on the treatment dosage. Student identity was modeled as a random intercept. For each evolution learning outcome variable, four models were fit: a control model (model C), an AL-only model (model A), an AL+MFI model (model AM) and an AL+MFI+AL ∩ MFI model (model AMI; figure 4). See section 3 of the supplemental material and supplemental table S5 for additional details about the models.

Analyses. Likelihood ratio tests (LRT) and Bayesian information criteria were used to determine the extent to which the percentage of AL contributed to evolution learning outcomes (model C versus model A; see figure 4). To determine how varying AL doses contributed to evolution knowledge growth through time, the model coefficients for the interactions between the percentage of AL and each time period in model A were examined. The coefficient of the interaction represents the change in slope of evolution understanding at varying doses of AL.

To determine whether MFI significantly contributed to evolution learning outcomes above and beyond AL alone, the percentage of MFI was added as a predictor (model AM) and compared with model A using LRT (figure 4). To determine how varying MFI doses contributed to learning, model coefficients were examined for the interaction between the percentage of MFI and each time period in model AM.

Finally, the percentage of AL ∩ MFI was added as a predictor (model AMI) and compared with model AM (figure 4) to determine whether it significantly and uniquely contributed to evolution learning. To determine how varying the percentage of AL ∩ MFI doses contributed to evolution knowledge growth, the coefficients for the interaction between the percentage of AL ∩ MFI and each time period in model AMI were examined.

To test for replicability across instructors, the patterns of evolution learning in iteration D were compared between two sections of the course, one of which was taught by the experienced instructor who taught the prior three iterations of the course and one of which was taught by a novice instructor. See section 4 of the supplemental material for additional details.

For all analyses, standardized coefficients (β), odds ratios (OR), and partial omega squared (ω^2_p; Lakens 2013) were used to measure effect sizes where appropriate.

Key findings about the contribution of active learning and misconception-focused instruction to evolution learning

In all of the course iterations we studied, knowledge of and abstract reasoning about evolution significantly increased, and misconceptions significantly decreased (figure 5). Specifically, the CINS ($B_{\text{pretest-posttest}} = 14.8$, $\beta_{\text{pretest-posttest}} = 0.33$, $p < .001$, $\omega^2_p = 0.37$), ACORNS core concepts ($B_{\text{pretest-posttest}} = 1.03$, $\beta_{\text{pretest-posttest}} = 0.46$, $p < .001$, $\omega^2_p = 0.54$, $B_{\text{posttest-delayed}} = 0.18$, $\beta_{\text{posttest-delayed}} = 0.08$, $p < .001$, $\omega^2_p = 0.05$), and ACORNS MODC ($\text{OR}_{\text{pretest-posttest}} = 9.82$, $\beta_{\text{pretest-posttest}} = 2.22$, $p < .001$, $\text{OR}_{\text{pretest-delayed}} = 1.51$, $\beta_{\text{pretest-delayed}} = 0.40$, $p < .001$) significantly increased from both the pretest to posttest and the posttest to delayed posttest. Likewise, ACORNS misconceptions decreased significantly both from the pretest to the posttest and from the posttest to the delayed posttest ($\text{OR}_{\text{pretest-posttest}} = 0.26$, $\beta_{\text{pretest-posttest}} = -1.6$, $p < .001$, $\text{OR}_{\text{posttest-delayed}} = 0.63$, $\beta_{\text{posttest-delayed}} = -0.55$, $p < .001$). The percentage of students with evolution misconceptions declined from approximately 33% at the pretest to approximately 19% at the delayed posttest in the baseline iteration (iteration A, no MFI) and from approximately 30%–40% at the pretest to approximately 12% (iteration B), 6% (iteration C), and 5% (iteration D) at the delayed posttest in the intervention iterations (figure 5). The finding that the posttest to delayed posttest time period had a significant slope for all outcome measures indicated that the growth in knowledge and abstract reasoning, as well as the reduction in misconceptions, continued even after the posttest (i.e., during units on phylogenetics, diversity, matter and energy; figure 5). Most of the learning, however, occurred from the pretest to the posttest (i.e., larger ω^2_p value for CINS and core concepts; OR values further from 1 for MODC and misconceptions). The semester-specific effect sizes of instruction increased as AL and MFI dosage increased (table 3). The results from the final iteration were replicated with the novice instructor (supplemental figure S1 and section 4.2 of the supplemental material).

AL was found to be a significant contributor to evolution learning. Specifically, including the percentage of AL into the model (model A) significantly improved the model above and beyond the control model (model C) for all evolution knowledge outcome variables (table 4).

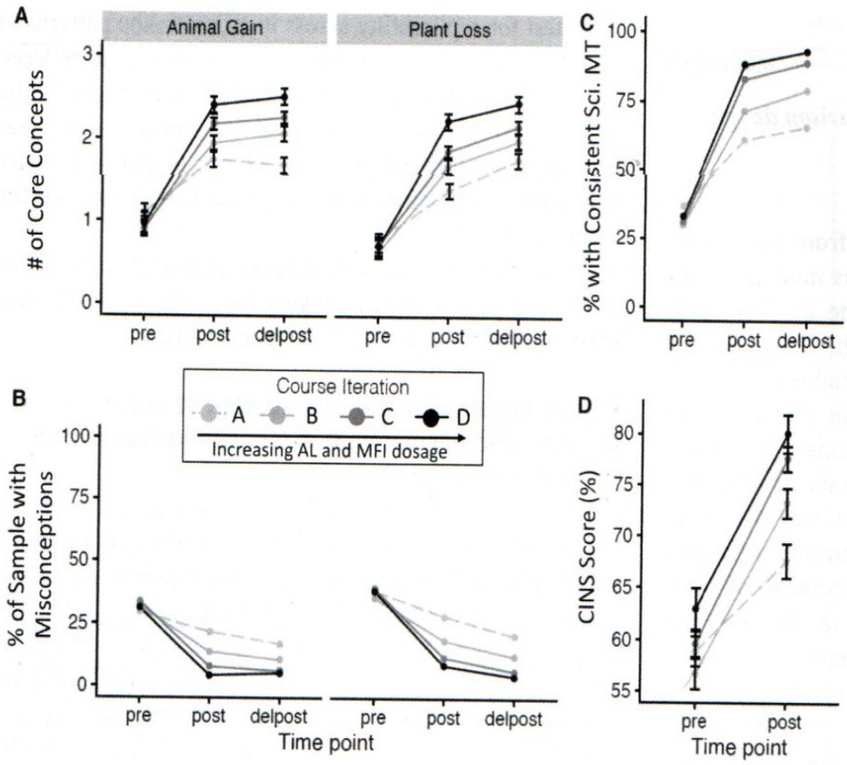

Figure 5. Raw scores for ACORNS core concepts (a), the percentage of sample with misconceptions (b), the percentage of sample with consistent scientific model type (c), and CINS (d) across all four course iterations. The error bars represent two standard errors. The patterns for the CINS in iterations C and D are comparable to the patterns found for the CANS in the same course iterations (see supplemental figure S2).

Furthermore, the growth in evolution knowledge from the pretest to the posttest was conditional on the amount of AL: the higher the percentage of AL, the larger the growth (i.e., change in slope) from the pretest to the posttest for all outcome measures and with a medium effect size (CINS, $B = 0.30$, $\beta = 0.15$, $\omega^2_p = 0.02$, $p < .001$; core concepts, $B = 0.03$, $\beta = 0.34$, $\omega^2_p = 0.08$, $p < .001$; misconceptions, OR = 0.49, $\beta = -1.46$, $p < .001$; MODC, OR = 2.11, $\beta = 1.23$, $p < .001$). This finding is visualized by pretest–posttest slope steepness with an increasing percentage of AL (figure 6a–d). However, although evolution knowledge continued to grow significantly from the posttest to the delayed posttest (see above), the magnitude of growth was not conditional on the amount of AL. In other words, all course iterations showed gains in evolution knowledge at a similar rate during this time period, regardless of the amount of AL in the prior period. This finding is visualized by the parallel posttest and delayed posttest slopes in figure 6e–g (see model A in supplemental table S6a–d for more details). Although we have evidence that participation incentive (e.g., a required posttest, a voluntary delayed posttest; see Sbeglia and Nehm 2022) did not affect assessment performance, even if it did, evolution knowledge continued to increase from the posttest (required) to the delayed posttest (voluntary) in all course iterations, and the general finding that higher amounts of AL and MFI produced larger rates of change through time still holds.

MFI was found to be a significant and unique contributor to evolution learning, and this contribution was above and beyond that of AL alone. Specifically, adding the percentage of MFI to the model (model AM) significantly improved the model above and beyond the AL-only model (model A) for all outcome measures (table 4). Furthermore, the growth in evolution knowledge was conditional on the amount of MFI from the pretest to the posttest; the higher the percentage of MFI, the larger the growth (i.e., change in slope) during this period for all outcome measures (CINS, $B = 0.97$, $\beta = 0.22$, $\omega^2_p = 0.02$, $p < .001$; core concepts, $B = 0.04$, $\beta = 0.20$, $\omega^2_p = 0.02$, $p < .001$; misconceptions, OR = 0.64, $\beta = -0.90$, $p < .001$; MODC, OR = 1.88, $\beta = 1.06$, $p < .001$; see figure 7a–d). However, although evolution knowledge continued to increase significantly from the posttest to the delayed posttest (as reported above), the growth during this period was *not* conditional on the amount of MFI. This finding is visualized by the parallel posttest to delayed posttest slopes for all doses of the percentage of MFI in figure 7e–g. See model AM in supplemental table S6a–d for additional detail.

The percentage of AL ∩ MFI did not significantly contribute to evolution learning above and beyond the separate contributions of AL and MFI (the LRTs between model AM and AMI were not significant for any outcome variables). Therefore, the growth in evolution knowledge through time was not conditional on the amount of intersection between AL and MFI. See model AMI in supplemental table S6a–d for additional detail.

The results for the four evolution understanding measures (CINS, ACORNS core concepts, ACORNS misconceptions, ACORNS MODC) were similar in that they showed that AL and MFI contributed to evolution learning. However, ACORNS core concepts and CINS showed different significance patterns for the percentage of AL depending on whether the percentage of MFI was included in the model. Specifically, in model AM (which included the percentage of MFI and the percentage of AL), the growth in evolution knowledge through time was conditional (although with a very small effect size) on the amount of AL for ACORNS core concepts ($B = 0.01$, $\beta = 0.14$, $\omega^2_p = 0.005$, $p < .01$; supplemental figure S3a). This was not the case for the CINS (supplemental figure S3d). Therefore, ACORNS core concepts appears to have accounted for more unique variation in the growth of normative evolution knowledge than the CINS.

Table 3. Effect size (standardized coefficients, odds ratios, and partial omega squared) of instruction for each evolution understanding measure in each course iteration (generated from model C)

Measure	Course iteration	Pre to post (time 1)			Post to delayed post (time 2)		
		β	OR	ω^2_P	β	OR	ω^2_P
CINS	A	0.19	–	0.18	–	–	–
	B	0.36	–	0.4	–	–	–
	C	0.43	–	0.48	–	–	–
	D	0.39	–	0.46	–	–	–
Core concepts	A	0.29	–	0.35	0.07	–	0.03
	B	0.46	–	0.55	0.10	–	0.06
	C	0.55	–	0.63	0.09	–	0.06
	D	0.64	–	0.73	0.07	–	0.07
Misconceptions	A	−0.6	0.57	–	−0.5	0.63	–
	B	−1.41	0.30	–	−0.56	0.62	–
	C	−2.5	0.13	–	−0.70	0.57	–
	D	−3.37	0.08	–	−0.48	0.70	–
MODC	A	1.29	3.62	–	0.28	1.32	–
	B	1.95	7.58	–	0.44	1.58	–
	C	3.22	24.52	–	0.60	1.81	–
	D	3.97	47.17	–	0.70	1.96	–

Note: partial omega squared (ω^2_P): small = 0.01, medium = 0.06, large = 0.14 (Lakens 2013); odds ratio (OR): small = 1.68 (0.59), medium = 3.47 (0.29), large = 6.7 (0.15) (Chen et al. 2010); β is the standardized coefficient. All effect sizes are significant at $p < .001$

Table 4. Likelihood ratio test results.

Model comparison	Instrument	Bayesian information criteria (BIC)	Chi squared	df	p
Control model versus AL-only model	CINS	BIC_{modelC} = 25,890, BIC_{modelA} =25,850; ΔBIC = 40	56.258	2	< .001
	ACORNS core concepts	BIC_{modelC} =21,530, BIC_{modelA} = 21,334; ΔBIC = 196	222.76	3	< .001
	ACORNS misconceptions	BIC_{modelC} = 8095.4, BIC_{modelA} = 7982.2; ΔBIC = 113.2	140.46	3	< .001
	ACORNS MODC	BIC_{modelC} = 4963.6, BIC_{modelA} = 4842.0, ΔBIC = 121.6	146.86	3	< .001
AL-only model versus AL+MFI model	CINS	BIC_{modelA} = 25,850, $BIC_{modelAM}$ = 25,831; ΔBIC = 19	34.87	2	< .001
	ACORNS core concepts	BIC_{modelA} = 21,334, $BIC_{modelAM}$ = 21,323; ΔBIC = 11	38.69	3	< .001
	ACORNS misconceptions	$deviance_{modelA}$ = 7845.6, $deviance_{modelAM}$ = 7821.9, Δ deviance = 23.7	23.66	3	< .001
	ACORNS MODC	BIC_{modelA} = 4842.0, $BIC_{modelAM}$ = 4839.2, ΔBIC = 2.8	28.04	3	< .001

Active learning, misconceptions, and biology teaching

Although AL has become a focus in biology education in recent years, misconceptions in general and evolution misconceptions in particular have been issues of educational concern for at least a century (Osbern 1922). Constructivist perspectives on learning emphasize that effective teaching requires engaging with student ideas about how the world works (NRC 2001). Nevertheless, remarkably little is known about whether and to what extent misconceptions are explicitly targeted in undergraduate biology education or whether addressing them is worth the risk of reinforcement and perpetuation (cf. Lewandowsky et al. 2012). To our knowledge, the only large-scale study that asked biology instructors if they addressed evolution misconceptions was Andrews and colleagues (2011). In a survey of 33 instructors (out of 88 invited), they found that many instructors in this subsample addressed misconceptions in some way. However, the amount of instructional time devoted to misconceptions was not measured, and the data gathered were based on self-reports (versus empirical observations). Intentional study designs that investigate the impacts of varying amounts of both AL and MFI on biology learning outcomes are needed. Such designs may benefit from the conceptualization of classroom activities as occurring along a gradient (see figure 1), as has been done in this study, as opposed to presence or absence.

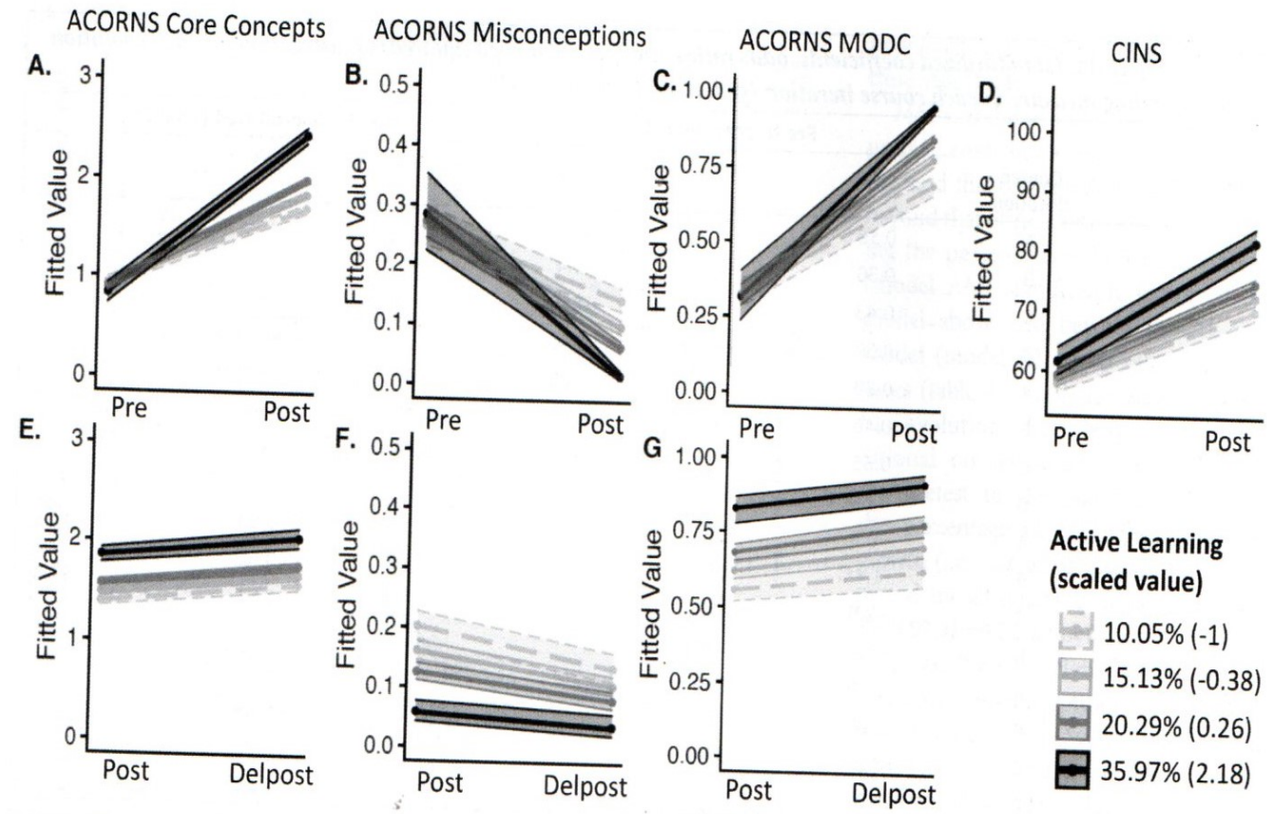

Figure 6. Marginal effects plots of model A. The slope of the fitted values from the pretest to the posttest for each active learning (AL) dosage is shown in panels (a)–(d), and the slope of the fitted values from the posttest to the delayed posttest for each AL dosage is shown in panels (e)–(g). The shaded areas around the lines represent the 95% confidence intervals.

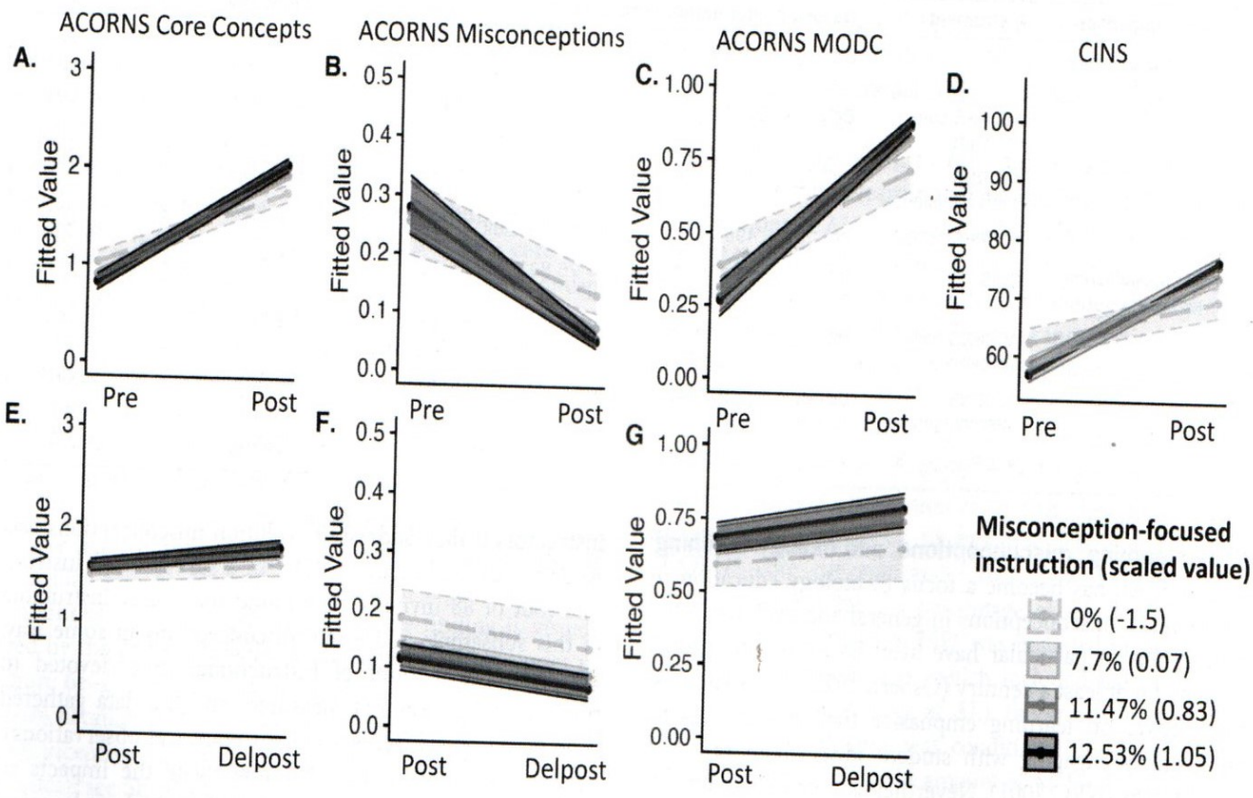

Figure 7. Marginal effects plots of model AM. The slope of the fitted values from the pretest to the posttest for each misconception-focused instruction (MFI) dosage is shown in panels (a)–(d), and the slope of the fitted values from the posttest to the delayed posttest for each MFI dosage is shown in panels (e)–(g). The shaded areas around the lines represent the 95% confidence intervals.

The students in all of the course iterations experienced large evolution knowledge gains that were not instructor dependent. Furthermore, higher doses of AL were significantly associated with larger increases in evolution knowledge (measured using the CINS and ACORNS core concepts) and abstract reasoning about evolution (ACORNS MODC) and with larger declines in evolution misconceptions (ACORNS misconceptions). For all evolution knowledge measures, this pattern occurred only from the pretest to the posttest, and after this point, all AL dosages were associated with a similar (though still significant) rate of increase in evolution understanding. Although no explicit evolution instruction occurred between the posttest and the delayed posttest, the significant rate of change between these two timepoints could be due to the incorporation of evolution examples in the remaining units of the course (e.g., plant evolution).

In their influential study, Freeman and colleagues (2014) compared student outcomes in AL with traditional courses and reported a standardized mean difference (Hedges's *g*) of approximately 0.30 in postcourse knowledge scores in biology classrooms. To compare findings from Freemen and colleagues (2014), which was limited to the presence or absence of AL, with those of the present study, the course iteration with the lowest-intensity AL dosage (iteration A) was compared with other iterations (iterations B, C, and D). The resulting Hedges's *g* values from the present study were in line with Freeman and colleagues (2014; A versus B = 0.13, A versus C = 0.25, A versus D = 0.46). Freeman and colleagues (2014) also reported an OR for failing a course of 1.95 in traditional versus AL classrooms, which is similar to our reported effect of AL on whether or not students have misconceptions (OR = 0.49, equivalent to approximately 2.00) and consistent scientific models (OR = 2.11). In the current study, we did not use course failure versus passing as an outcome variable because we focused on only one unit of the course.

As was described above, the results were clear that the intensity (or dose) of AL (up to approximately 36% of instructional time) was associated with gains in normative ideas and reasoning abstraction and with the loss of misconceptions. The results were also clear that higher doses of MFI (up to approximately 12.5%) were significantly associated with these same outcomes: larger increases in normative ideas (CINS, ACORNS core concepts) and abstract reasoning (ACORNS MODC) and larger declines in evolution misconceptions (ACORNS misconceptions). Importantly, this impact was above and beyond the contribution of AL alone. Although these findings are encouraging, recent work suggests that the impacts of AL may continue to increase at even higher intensities than were addressed in this study. Theobald and colleagues (2020), for example, reported that very-high-intensity AL conditions (more than 60% AL) were associated with larger reductions in achievement gaps between minoritized and nonminoritized students. Furthermore, although no prior work has evaluated how the dosage of MFI relates to learning outcomes, the impacts of MFI may also continue to increase at even higher intensities. Indeed, implementing higher intensity AL and MFI conditions (to fill the empty space in figure 1) and disaggregating results by student backgrounds are essential next steps.

In addition to the amount of MFI, the type and quality of misconception treatment as well as the specific misconception content addressed may also contribute to learning outcomes, but these topics were beyond the scope of this study. Indeed, all misconceptions (supplemental table S1) and treatment types (e.g., video, active, collaborative work, or individual work; supplemental table S4) were treated as equal even though it is likely that they differ in fundamental ways (see Introduction). However, all treatments attempted to promote cognitive dissonance in students, which is known to be an essential feature of effective MFI (Guzzetti et al. 1993).

Few studies have documented the frequencies, types, and content of MFI in undergraduate biology classrooms, and evidence-based frameworks outlining the salient dimensions of MFI are lacking. For this reason, this study focused on whether the amount of MFI was associated with learning outcomes, but future work is clearly needed to evaluate the contributions of treatment types, misconception content, and quality (e.g., promoting cognitive dissonance). Arguably, by including the percentage of AL $\cap$ MFI in model AMI, we accounted for one possible dimension of MFI type (i.e., use of AL) that could possibly contribute to learning outcomes. However, adding this variable did not significantly improve the model according to LRTs, and the percentage of MFI generally remained a significant contributor to evolution learning outcomes.

The results from this study show remarkable consistency with Andrews and colleagues' (2011) finding that addressing misconceptions was positively associated with evolution learning gains. Unfortunately, we were unable to compare the size of the effect of their misconception classes (i.e., addressed or not) with ours for a few reasons. First, Andrews and colleagues (2011) used only class level (versus student level) variables in their models, making the resulting coefficients incomparable. Second, the two studies quantified instructor use of misconceptions differently. Andrews and colleagues (2011) used instructor self-reports to document whether they addressed misconceptions and whether they explained why those misconceptions were incorrect. Our study used archived Echo videos to quantify the amount of MFI and confirm that all misconception-related activities (supplemental table S1) did indeed take place. Despite similar findings about the role of MFI in evolution learning outcomes, Andrews and colleagues (2011) and the present study were not aligned in terms of the roles that AL played in these outcomes. Specifically, the present study reported a significant contribution of AL, and Andrews and colleagues (2011) did not. See section 5 of the supplemental material for an expanded discussion of this point.

Future work on the role of AL and MFI in biology learning must consider several aspects of study design. First, AL and MFI were correlated in the present study, and experimental

designs that independently vary AL and MFI will be important next steps for deepening our understanding of how to most effectively affect learning in introductory courses. Second, student attendance patterns may affect study results. Although we did not document attendance on a per-student basis, the course in this study had a relatively high overall attendance (more than 80% as measured by clicker data); courses with lower attendance may generate different findings. Third, future work must also consider the choice of instruments used to measure learning. This study used multiple instruments that adopt different perspectives on knowledge that may affect the inferences one might make about the factors contributing to learning (cf. Freeman et al. 2014). For example, ACORNS misconceptions and MODC measure different aspects of evolution learning that are not entirely captured by ACORNS core concepts or the CINS and show subtle differences in how they interact with instructional interventions such as AL. Finally, our results suggest that researchers who include both AL and MFI in the same model (e.g., Andrews et al. 2011) could draw different conclusions about the contribution of AL in evolution learning depending on whether they use the CINS or ACORNS core concepts as their measure of evolution knowledge.

This study has implications for teaching, learning, and research in biology domains in addition to evolution. Over the past century, biology educators have published hundreds of studies (e.g., sources in https://archiv.ipn.uni-kiel.de/stcse) documenting numerous student misconceptions in cell biology, photosynthesis and respiration, genetics, inheritance, gene expression, speciation, phylogenetics, genetic drift, matter and energy, and ecology (e.g., Driver et al. 1994, Baum et al. 2005, Wilson et al. 2006, Gregory 2008, Novick and Catley 2008, Halverson et al. 2011, Kinlock et al. 2020, Catley et al. 2013). This work has been extended in recent decades by biology education researchers through the development of many concept inventories designed to measure both misconceptions and normative ideas for a given biology topic (see Nehm 2019 for a review). These research-based assessment tools exist for many biology topics and could be used to help more faculty become aware of student misconceptions and guide the design of AL experiences that effectively engage students in thinking more deeply about their alternative perspectives on how the living world works. Biologists and biology educators should work together to better understand the roles that the intensity, type, and quality of MFI have in enhancing AL outcomes in biological domains beyond evolution.

So, is active learning using normative scientific ideas enough?

Our results demonstrate that explicitly addressing misconceptions produced significant and meaningful evolution learning above and beyond AL alone. However, the presence and extent of MFI in prior AL studies remains underspecified. Indeed, the most influential meta-analysis demonstrating the benefits of AL across STEM disciplines (Freeman et al. 2014) did not include MFI as a variable. This raises the question of whether variation in AL efficacy (e.g., Andrews et al. 2011, Freeman et al. 2014's figure 1a, Theobald et al. 2020's figure 2) is related to the presence and intensity of MFI during learning experiences. Explicit attention to both high-intensity AL and MFI is likely to enhance learning outcomes in many areas of biology education.

Acknowledgments

Support for instrument development, data collection, and response scoring was provided by National Science Foundation grant no. TUES-1322872, and implementation was supported by a Howard Hughes Medical Institute Inclusive Excellence grant. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or HHMI. We thank professor John True for his support and commitment throughout this four year project. We also thank two anonymous reviewers for providing many helpful comments that strengthened the depth and clarity of the manuscript.

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

Anderson DL, Fisher KM, Norman GJ. 2002. Development and evaluation of the conceptual inventory of natural selection. Journal of Research in Science Teaching 39: 952–978.

Andrews TM, Leonard MJ, Cogrove CA, Kalinowski ST. 2011. Active learning *not* associated with student learning in a random sample of college biology courses. CBE–Life Science Education 10: 329–435.

Ausubel DP. 1968. Educational Psychology: A Cognitive View. Holt, Rinehart, and Winston.

Bates D et al. 2022. Package "lme4": Linear Mixed-Effects Models using "Eigen" and S4. R Project. https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Baum DA, Smith SD, Donovan SSS. 2005. The tree-thinking challenge. Science 310: 979–980.

Beardsley PM, Bloom MV, Wise SB. 2012. Challenges and opportunities for teaching and designing effective K–12 evolution curricula. Pages 287–310 in Rosengren KS, Brem SK, Evans EM, Sinatra GM, eds. Evolution Challenges. Oxford University Press.

Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ. 2014. Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? Journal of Science Education and Technology 23: 160–182.

Bertolini R, Finch SJ, Nehm RH. 2021. Enhancing data pipelines for forecasting student performance: Integrating feature selection with cross-validation. International Journal of Educational Technology in Higher Education 18: 44.

Bishop BA, Anderson CW. 1990. Student conceptions of natural selection and its role in evolution. Journal of Research is Science Teaching 27: 415–427.

Brewer C, Smith D. 2011. Vision and Change in Undergraduate Education: A Call to Action. American Association for the Advancement of Science.

Brown SA, Ronfard S, Kelemen D. 2020. Teaching natural selection in early elementary classrooms: Can a storybook intervention reduce teleological misunderstandings? Evolution: Education and Outreach 13: 1–19.

Caravita S, Halldén O. 1994. Re-framing the problem of conceptual change. Learning and Instruction 4: 89–111.

Catley KM, Phillips BC, Novick LR. 2013. Snakes and eels and dogs! Oh, my! Evaluating high school students' tree-thinking skills: An entry