Evolution: Education and Outreach

RESEARCH Open Access

Measuring evolution learning: impacts of student participation incentives and test timing

Gena C. Sbeglia[1,3*] and Ross H. Nehm[1,2]

Abstract

Background: Policy documents like *Vision and Change* and the *Next Generation Science Standards* emphasize the importance of using constructed-response assessments to measure student learning, but little work has examined the extent to which administration conditions (e.g., participation incentives, end-of-course timing) bias inferences about learning using such instruments. This study investigates potential biases in the measurement of evolution understanding (one time point) and learning (pre-post) using a constructed-response instrument.

Methods: The constructed-response ACORNS instrument (Assessment of COntextual Reasoning about Natural Selection) was administered at the beginning of the semester, during the final exam, and at end of the semester to large samples of North American undergraduates (N = 488–1379, 68–96% participation rate). Three ACORNS scores were studied: number of evolutionary core concepts (CC), presence of evolutionary misconceptions (MIS), and presence of normative scientific reasoning across contexts (MODC). Hierarchical logistic and linear models (HLMs) were used to study the impact of participation incentives (regular credit vs. extra credit) and end-of-course timing (final exam vs. post-test) on inferences about evolution understanding (single time point) and learning (pre-post) derived from the three ACORNS scores. The analyses also explored whether results were generalizable across race/ethnicity and gender.

Results: Variation in participation incentives and end-of-course ACORNS administration timing did not meaningfully impact inferences about evolution understanding (i.e., interpretations of CC, MIS, and MODC magnitudes at a single time point); all comparisons were either insignificant or, if significant, considered to be small effect sizes. Furthermore, participation incentives and end-of-course timing did not meaningfully impact inferences about evolution *learning* (i.e., interpretations of CC, MIS, and MODC changes through time). These findings were consistent across race/ethnicity and gender groups.

Conclusion: Inferences about evolution understanding and learning derived from ACORNS scores were in most cases robust to variations in participation incentives and end-of-course timing, suggesting that educators may have some flexibility in terms of when and how they deploy the ACORNS instrument.

Keywords: Assessment, Learning evolution, Undergraduates, ACORNS, Testing conditions

Introduction

Evolution is a foundational component of life science education (AAAS 2011), yet a large body of work indicates that it remains a challenging concept for students to learn (Bishop and Anderson 1990; Gregory 2009). Numerous studies have documented common "misconceptions" and cognitive biases that students hold about

*Correspondence: gsbeglia@sdsu.edu

[3] Department of Biology, San Diego State University, San Diego, USA
Full list of author information is available at the end of the article

evolution (e.g., need-based trait change (i.e., teleological reasoning), use-disuse inheritance; Gregory 2009; Kampourakis 2020; Kelemen 2012).Concerningly, many misconceptions persist after evolution instruction (Andrews et al. 2011; Nehm and Reilly 2007). Students also struggle with identifying the salient features of real-world evolutionary problems (Nehm and Ridgway 2011). The polarity of evolutionary change–trait gain vs. loss–also poses particular challenges for undergraduates throughout the world (e.g., USA, China, Germany, Korea, Indonesia); students have been found to use more misconceptions on trait loss assessment tasks (Nehm and Ha 2011; Ha et al. 2019; Nehm 2018). Finally, students often develop idiosyncratic mental models tied to specific learning examples and lack coherent causal frameworks (Nehm 2018). Determining which instructional approaches are most effective at fostering evolution learning and reducing misconceptions requires assessment tools capable of generating valid and reliable inferences about student thinking (Mead et al. 2019; Nehm and Mead 2019).

Several different assessment tools have been developed for undergraduate evolution educators to measure learning (reviewed in Furrow and Hsu 2019; Mead et al. 2019), and they include both closed-response formats (e.g., multiple choice, true false; Kalinowski et al. 2016) and constructed-response (e.g., written) formats (e.g., Nehm et al. 2012a). Two-tier formats (e.g., a closed response item with an associated open-response explanation for the initial response) have been developed for some areas of biology but not for evolution. There are advantages and disadvantages of different assessment formats (reviewed in Nehm 2019). Closed-response tools are easy to administer and grade, and they require little time and few resources (Klymkowsky et al. 2003). One disadvantage is that education researchers have questioned whether closed-response tools are actually measuring deep disciplinary understanding, or if they are instead measuring surface-level reasoning and test taking strategies (Huffman and Heller 1995; Smith and Tanner 2010). By design, closed-response assessments require students to select (vs. generate) an answer, which may limit students from communicating what they actually think (Smith and Tanner 2010). Another major disadvantage of closed-response assessments is that they are poorly suited for measuring a variety of problem solving and communication skills relevant to authentic scientific practice (e.g. explanation, argumentation, modeling; NRC 2012).

Real-world scientific problems require the problem solver to weigh the relevance and importance of information and to assemble diverse sources of information into coherent and logical structures (Nehm and Schonfeld 2010; Nehm et al. 2012b; Haudek et al. 2012). Yet research shows that students often lack proficiency in these open-ended and ill-structured tasks (Haudek et al. 2012). Students are often most proficient at retaining large amounts of isolated bits of information and choosing from sets of pre-structured, clearly-laid out statements (NRC 2012). Authentic performance tasks like explanation, communication, model building, and argumentation are central components of real-world scientific problem solving, and are core learning objectives in the *Next Generation Science Standards* and *Vision and Change* (see NRC 2012; AAAS 2011). In order to address these learning objectives, these standards documents emphasize that more robustly validated constructed response assessments are needed to measure performance tasks in science (e.g., explanation, argumentation, modeling).

Despite their advantages, widespread adoption of constructed-response assessments in undergraduate settings has been limited because of the prohibitive costs associated with scoring student responses (see Nehm, Ha, Mayfield 2012 for a detailed discussion of these limitations). Over the past decade, these costs have been reduced substantially because of inexpensive artificial intelligence tools (such as machine learning) capable of automatically and accurately scoring student responses (Moharreri et al. 2014; see also www.beyondmultiplechoice.org). It is likely that technological advances will continue to foster the development of more constructed response assessments for evolution and other science domains. The only constructed-response instrument designed to measure evolution knowledge that can be automatically scored is the ACORNS (Assessment of COntextual Reasoning about Natural Selection; Nehm et al. 2012a).

The ACORNS was developed by enhancing and standardizing Bishop and Anderson's (1990) questions. The ACORNS instrument prompts students to generate evolutionary explanations for patterns of biological difference across different lineages (e.g., plants vs. animals), trait polarities (e.g., gain vs. loss of a trait), taxon familiarities (penguin vs. prosimian), scales (within- vs. between-species), and trait functions (e.g., claws vs. fur color). These items provide faculty with a range of contexts that can be used to understand student thinking about evolutionary processes. An example item is: "How would biologists explain how a species of cactus without spines evolved from a species of cactus with spines?" The skeletal structure of an ACORNS item permits substitution of the features (underlined above). That is, "How would [A] explain how a [B] of [C] [D1] [E] evolved from a [B] of [C] [D2] [E]?" This skeleton is fleshed out with specific features: A = perspective (e.g., you, biologists), B = scale (e.g., species, population), C = taxon (e.g., plant, animal, bacteria, fungus), D = polarity (e.g., with, without), and E = trait (e.g., functional, static). Dozens of

Table 1 Selected prior studies of the properties, potential biases, and validity of ACORNS instrument score inferences

ACORNS study	Authors	Key finding
Grounding the design of the assessment in well-established cognitive principles	Opfer et al. (2012)	The ACORNS aligns with three core cognitive principles central to scientific reasoning following NRC (2001) recommendations
Correspondence of written explanation scores to clinical oral interviews with undergraduates	Beggrow et al. (2014)	More than 100 students' interview scores were compared to ACORNS scores and found to have greater correspondence than to a multiple-choice evolution assessment
Analysis of potential English Learner (EL) bias in written tasks	Ha and Nehm (2016)	Scoring of EL ACORNS spelling errors did not show bias using the EvoGrader scoring tool
Examination of potential gender bias in written tasks	Federer et al. (2016)	DIF analyses found minimal gender bias in ACORNS written tasks
Study of how the order of items impacts student performance	Federer et al. (2014)	Recommendation that two ACORNS items differing in surface features have the least order and test fatigue effects
Analysis of ACORNS-like responses and interpretation bias for lexically ambiguous wording (e.g., "adapt")	Rector et al. (2013)	The vast majority of scoring interpretations were corroborated after follow-up questioning, although some misinterpretation errors were documented
Correspondence of automated scoring of ACORNS responses using machine learning models to trained raters	Moherrari et al. (2014)	The EvoGrader automated scoring tool provides accurate and consistent scoring of answers, eliminating human rater inconsistencies across individuals and through time

additional examples of ACORNS items may be found at www.evograder.org. Students generate written explanations to the question in an unconstrained format, which permits them to determine which ideas *they* view as most relevant to an evolutionary explanation.

Many studies of the ACORNS instrument have been conducted (Table 1); they have explored the role of gender in response patterns, the impact of item order on response patterns, the scoring and interpretation of student language, and validity and reliability inferences (See Table 1 and Additional file 1: Table S1 for citations additional details about existing validity and reliability evidence for this instrument). Existing gaps in the literature on constructed-response assessments in general and for the ACORNS in particular include the impact of test administration conditions on static measures of understanding and longitudinal measures of learning (DeMars 2000).

Prior work on testing conditions and assessment scores

Robust assessment tools are backed by multiple types of reliability and validity evidence; this evidence is used to support claims about instrument quality (see articles in Nehm and Mead 2019; Mead et al. 2019). While necessary, these forms of reliability and validity evidence may be insufficient for guiding instrument administration decisions. For example: Does it matter when a course post-test is administered (e.g., last week of classes, during the final exam, after the final exam)? Should participation incentives be offered to students for completing the assessment (e.g., extra credit, regular credit, no credit)? At present, remarkably little evidence-based guidance is available to help inform test administration decisions, particularly for constructed-response instruments.

Prior work on closed-response assessments (e.g., multiple-choice, true–false) in domains other than evolution have pointed to potential measurement biases due to the administration conditions used by researchers (DeMars 2000; e.g., Couch and Knight 2015; Ding et al. 2008; Duckworth et al. 2011; Smith et al. 2012; Wolf et al. 1996; Wolf and Smith 1995). For example, researchers have reported a significant impact of participation incentive (e.g., incentive absent/no course credit vs. incentive present/course credit) on multiple-choice and true–false assessment scores (e.g., Ding et al. 2008; Duckworth et al. 2011; Wise and DeMars 2005; Wolf et al.1996; Wolf and Smith 1995). Conversely, some researchers have found no impact of other testing conditions on these scores, such as the end-of-course assessment timing (e.g., last day of the course vs final exam; Smith et al. 2012) and the test setting (e.g., in-person vs. out-of-class; Couch and Knight 2015). Across domains, far fewer studies have investigated and isolated the impacts of test administration conditions on assessment scores using constructed-response items, which have become increasingly central to biology education research and practice (AAAS 2011; NRC 2012; Beggrow et al. 2014; Nehm et al. 2012b). The one evolution-focused study identified in our literature review that examined effects of administration conditions on a constructed-response item examined only assessment timing, and reported significant differences in the use of normative ideas about evolution at two different end-of-course timepoints, but no differences in the use of misconceptions about evolution at these two timepoints (Nehm and Reilly 2007).

Prior studies of bias associated with test administration conditions typically lack longitudinal designs (indeed, many lack the requisite data [i.e., a pre-test] for such approaches). For this reason, it is unclear if the significant differences reported by prior, non-longitudinal work (i.e., Nehm and Reilly 2007) would extend to meaningful differences in the overall *magnitudes of change* over time. In educational contexts, analyzing patterns and magnitudes of change from, for example, pre- to post-instruction is essential for identifying whether and to what extent learning occurred (and for whom). Overall, two gaps in the literature on constructed response instruments motivated our work: (i) the extent to which test administration conditions impact assessment scores and (ii) the extent to which test administration conditions impact inferences about student learning.

Research questions

This study examines the measurement of evolution learning in two semesters of an introductory biology course using the ACORNS instrument. Two test administration conditions are studied: student participation incentives (i.e., regular credit vs. extra credit) and end-of-course timing (i.e., final exam vs. post-course). Our research questions address the impact of these conditions on inferences about (i) evolution understanding (i.e., knowledge, misconceptions, and normative reasoning) and (ii) increases in understanding (i.e., learning) over time. Three research questions are addressed:

> **RQ1: (1.1)** Do parallel ACORNS items administered at the *same* end-of-course assessment time point (i.e., during the final exam) but with *different* participation incentives (i.e., regular credit vs. extra credit) produce statistically comparable knowledge and misconception patterns? **(1.2)** Do these findings generalize across gender and race/ethnicity groups?
>
> **RQ2: (2.1)** Do parallel ACORNS items administered at *different* end-of-course assessment time points (i.e., final exam vs. end of semester) produce statistically comparable knowledge, misconception, and normative reasoning patterns? **(2.2)** Do these findings generalize across gender and race/ethnicity groups?
>
> **RQ3: (3.1)** Does the timing of the ACORNS end-of-course assessment impact inferences about magnitudes of student learning? **(3.2)** Does this pattern generalize across gender and race/ethnicity groups?

Materials and methods

Study context

This study took place in two semesters of a high-enrollment introductory biology course at a large, public, doctorate-granting university in the United States. This course is focused on evolution and is taken by biology majors and nonmajors early in their academic careers (typically the first two years). There is no lab section associated with this "lecture" course and the prerequisites include high school biology and freshman-level mathematics. The course content is divided into six units aligned with *Vision and Change* (AAAS 2011). Evolution is a central organizing theme throughout the course. All assessments, including exams, occurred online. For each unit of the course, students were assigned recorded lectures and online mastery quizzes (i.e., the quizzes could be taken multiple times). Students engaged virtually in collaborative learning group work for each of the six units. The evolution unit focused on patterns and processes, sources of variation, natural selection, sexual selection, genetic drift, speciation, and misconceptions about evolution. Using the widely-adopted Classroom Observation Protocol for Undergraduate STEM (COPUS, Smith et al. 2013), at least 21% of the evolution unit involved active learning.[1]

Instrument and measures

The Assessment of Contextual Reasoning about Natural Selection (ACORNS; Nehm et al. 2012a) instrument was used to measure students' understanding of evolution in this study. Following recommendations from prior research (see Table 1), students were asked to complete two ACORNS items at each assessment time point that were identical except for the taxon, trait, and trait polarity. Specifically, one of the items focused on plant trait loss (i.e., orchid leaf loss, lily petal loss, rose thorn loss) and one item focused on animal trait gain in a familiar and unfamiliar context (i.e., snail poison gain and Strepsirrhine tapetum lucidum gain)[2] (Additional file 1: Table S2). All items were at the species scale and were asked from the perspective of a biologist. Prior work has indicated that using two ACORNS items that display maximum differences in surface features (e.g., animal vs. plant, trait gain vs. plant loss) generated the most robust inferences over the shortest period of time (Table 1). Validity and reliability inferences, along with many other factors, have been examined in a series of prior studies of the ACORNS (Additional file 1: Table S1, Table 1).

Student explanations produced in response to the ACORNS may utilize many different normative (scientifically accurate) and/or non-normative (naïve or scientifically inaccurate) ideas about evolution. Non-normative

[1] Several authors have used COPUS behaviors characteristic of student-centered teaching as measures of active learning. See Stains et al. 2018 and Sbeglia et al. 2021 for details.

[2] Some students received an animal loss item at some of the time points instead of an animal gain item, but these responses were not included in this study because they were not parallel across time.

ideas common to student thinking have been thoroughly documented in the evolution education literature (see Gregory 2009and Kampourakis 2020 for detailed reviews of these concepts). For example, inappropriate teleological thinking, 'use-disuse inheritance' models, and 'adaptation equals acclimatization' are common in undergraduate samples in general and our sample in particular (see Nehm 2018 and Additional file 1: Table S3 for examples of student responses from our sample). In our study, these three ideas are referred to using the colloquial term "Misconceptions" (MIS). The overall misconception score for each student response was coded as either present (1) or absent (0). We used presence/absence scoring to avoid modeling an ordinal scale with zero-inflated data, which was the pattern of the final exam and post-test in this sample. This binary coding approach is consistent with a perspective on evolution education in which the goal is to help students reach zero misconceptions. Normative or "correct" ideas most central to adaptive evolutionary causation (variation, heritability, differential survival/reproduction) are referred to as "Core Concepts" (CC) and scored as 0–3 for each response. Finally, as a measure of the degree to which students provide consistent normative reasoning across contexts (a measure of reasoning strategy consistency across surface features [see Opfer et al. 2012]), Model Consistency (MODC) was calculated.[3] MODC was scored as present (1) or absent (0) for each *pair* of student responses administered at a given time point (e.g., at the pre-test). Specifically, a student was scored as having a consistent scientific model if they used only CCs (and no MISs) in *both* of their ACORNS explanations (i.e., an ACORNS item about trait loss in a plant and an ACORNS item about trait gain in an animal) in a given two-item assessment. A student was scored as having the absence of a consistent scientific model if they used any MISs or no CCs. For additional scoring details and rubrics for these variables, see Nehm et al. (2010).

In order to provide consistent and independent scoring of students' written responses to the ACORNS items, core concepts and misconceptions were scored using the EvoGrader machine learning system (for details on scoring, comparisons to human raters, and reliability, see Moharreri et al. 2014). EvoGrader scoring has been shown to be as valid and more reliable (or consistent) than trained human raters, who often drift in their scoring and make small errors due to fatigue (Moharreri et al. 2014).

Sample

Student participants in two semesters completed the ACORNS items at three time points: within the first week of the course, during the online final exam, and after the course ended (see the next section for further details). At all assessment time points, students were administered two ACORNS items, one about evolution in a plant and one about evolution in an animal. More specifically, all 1434 students enrolled in the course were administered a plant item about the *loss* of a trait at all three time points. For this item, the number of participating students across time points ranged from 993–1379 (69–96% participation) depending on the analysis (see Additional file 1: Table S4 for details). Furthermore, a random half of the 1434 enrolled students (~716[4]) received an animal item about the *gain* of a trait at all three time points (the other half of the enrolled students received a mixture of animal gain and animal loss items throughout the semester and their responses on the animal item were therefore not included in this study). For this item, the number of participating students ranged from 488–685 (68–96% participation) depending on the analysis (see Additional file 1: Table S4 for details). For both the plant gain and animal loss items, the different analyses in the study have slightly different sample sizes (as shown in Additional file 1: Table S4) because they focus on different assessment time points, which differed in their participation patterns. In other words, not all students offered participation did so at all three time points.

Students reported several demographic and background variables at the beginning of the semester, including prior biology coursework (i.e., no prior biology, AP biology, one introductory biology course, two or more biology courses), gender (i.e., male, not male), and race/ethnicity (i.e., White, Asian, Underrepresented minority [URM]). URM students included those who identified as Black/African American, American Indian/Alaska, Native, Hispanic of any race/ethnicity, or Native Hawaiian/Other Pacific Islander. Students who identified as a URM made up ~22% of the sample and those identifying as male made up ~41% of the sample (see Additional file 1: Table S4 for details). This study was approved by the University's Institutional Review Board (IRB2019-00412) and classified as "Not Human Subjects Research." The procedures outlined in this study were in accordance

[3] The use of the word "model" in this previously-defined variable is unfortunate given the many meanings of the word model in science and science education. As mentioned earlier, this variable reflects a general approach to evolutionary reasoning (using exclusively scientific or normative concepts to explain multiple phenomena).

[4] The total number of students administered the animal gain item (716) is estimated because the testing platform we used randomly assigns items to students in real time during the exam, which makes it impossible to know which version of the animal item (animal gain or animal loss) a student who did *not* participate *would have* randomly received. As a result, we made the assumption that about half the class each semester would be randomly assigned an animal gain item and half assigned an animal loss item had everyone taken the final exam. In line with this assumption, the number 716 reflects half of the total number of students enrolled in each semester.

with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975.

ACORNS administration and testing conditions

The ACORNS items were administered online without the ability to backtrack (students could not see the second ACORNS item without completing their response to the first). All assessments were completed outside of a physical classroom space (e.g., students' personal computers). Students were informed that each assessment was closed-book and that they should not use any outside sources (e.g., other students, class notes, internet).

The two ACORNS items (one animal gain, one plant loss) were administered at three time points during the semester that we will refer to as "pre-test", "final exam", and "post-test." The pre-test was a voluntary assessment that occurred within the first week of the semester (prior to any evolution content). Course extra credit was provided for completing the pre-test. The post-test was a voluntary assessment that occurred during the two weeks following the final exam (i.e., during final exam week but before course course grades were posted). Course extra credit was also provided for completing the post-test. For both the pre- and post-test, students were informed that they would receive the extra credit if they provided complete and thoughtful answers.

The final exam occurred between the pre- and post-test and after all instruction was complete. For each student, one of the ACORNS items on the final exam was designated as regular credit (i.e., counted as a required part of the final exam score) and the second was designated as extra credit (i.e., counted as additional credit on the final exam). Specifically, half the class was randomly assigned a *regular credit* plant item and an *extra credit* animal item and the other half of the class was randomly assigned an *extra credit* plant item and a *regular credit* animal item. Students were presented with the regular credit item first and could not see the extra credit item until they submitted their answer for the regular credit item. Regular and extra credit items were worth the same number of course points. The ACORNS items were the only evolution-focused items on the final exam and, like the pre- and post-tests, students were not told in advance that they would be asked any questions about evolution.

The ACORNS administration conditions outlined above allowed for an investigation into the impact of two specific testing conditions. The first test administration condition analyzed (condition 1) was the participation incentive offered to students (regular credit vs. extra credit). The second test administration condition analyzed (condition 2) was the timing of the end-of-course assessment (final exam vs. post-test). The methods for each research question are described below.

Statistical analysis

RQ1. (Do parallel ACORNS items administered at the *same* end-of-course assessment time point (i.e., during the final exam) but with *different* participation incentives (i.e., regular credit vs. extra credit) produce statistically comparable knowledge and misconception patterns and do these findings generalize across demographic groups?) To address RQ1, evolution CC and MIS scores generated from extra credit and regular credit ACORNS items administered at the same assessment time point (i.e., the final exam) were compared. Separate regression models were run for the plant loss item and the animal gain ACORNS item using model formulations that were appropriate for the nature of the response data. Specifically, hierarchical logistic regressions were used for MIS (a binary outcome variable) and hierarchical linear regression was used for CC (treated as a continuous outcome variable) via the R package lme4 (Bates et al. 2021). Because CC could be conceptualized as an ordinal outcome variable, it was also modeled using an ordinal logistic regression via the R package ordinal (Christensen 2019). The outcomes of the analyses using these two characterizations of the variable did not differ (see the results section).

The incentive condition of the item (regular credit = 0, extra credit = 1) and semester were modeled as fixed effects (level 2 predictors). Student ID was modeled as a random effect (level 1 predictor). Partial omega squared (Lakens 2013) and odds ratios (Chen et al. 2010) were used as effect sizes where appropriate. Because the two incentive conditions (extra credit and regular credit) occurred at the same testing time point, this design effectively controlled for most other administration conditions (e.g., the timing of the assessment, the setting of the assessment, the prior knowledge at the time of the assessment).

To evaluate whether the results from this analysis generalized across demographic groups, the above models were modified to include an interaction effect between incentive condition and (a) gender (not male = 0, male = 1) and (b) race/ethnicity group (White = 0, Asian = 1, URM = 2). Interaction effects were also modeled between incentive condition and prior biology, which was treated as a control variable in this analysis. See Additional file 1: Table S4 for the exact sample sizes for these analyses.

RQ2. (Do parallel ACORNS items administered at *different* end-of-course assessment time points (i.e., final exam vs. end of semester) produce statistically comparable knowledge, misconception, and normative reasoning and do these findings generalize across demographic groups?) To address RQ2, the three measures of understanding (CC, MIS, and MODC) generated from ACORNS items administered at two end-of-course time points–the final exam (week 14) and the post-test (weeks 15–16)–were compared. Separate mixed models were run for plant loss and animal gain ACORNS items. Specifically, hierarchical logistic regressions were used for MODC and MIS, and hierarchical linear regression was used for CC. For each regression, student ID was modeled as a random intercept and time point (final exam = 0, post-test = 1) was modeled as a fixed effect. As in RQ1, CC was also modeled using an ordinal logistic regression (the results did not differ between these statistical approaches). Incentive condition (regular credit = 0, extra credit = 1) and semester were modeled as controls. As above, partial omega squared and odds ratios were used as effect sizes where appropriate. Importantly, because the two assessment time points in this analysis occurred within two weeks or less of each other, and were at least seven weeks after the evolution unit, differences in ACORNS measures at these two end-of-course time points likely do not represent differences in understanding. Therefore, the focus of this particular analysis was not to make inferences about evolution learning from one time point to the next, but rather about how the chosen end-of-course assessment time point (i.e., final exam vs. post-test) impacts inferences about the magnitude of students' evolution understanding at the end of an introductory course.

To evaluate whether the results from this analysis generalized across demographic groups, the above models were modified to include an interaction effect between end-of-course time point and (a) gender (not male = 0, male = 1) and (b) race/ethnicity group (White = 0, Asian = 1, URM = 2). Interaction effects were also modeled between end-of-course time point and prior biology as a control variable. See Additional file 1: Table S4 for the exact sample sizes for these analyses.

RQ3. To address RQ3 (Does the timing of the ACORNS end-of-course assessment impact inferences about magnitudes of student learning and does this pattern generalize across demographic groups?), the magnitude of change in ACORNS scores were compared between the beginning of the semester (i.e., pre-test) and the two end-of-course assessment time points: the final exam and the post-test. Therefore, the two time spans of interest were (1) pre-test to final exam (called "time span 1") and (2) pre-test to post-test (called "time span 2"). As above, hierarchical logistic regressions were used for MODC and MIS and hierarchical linear regression was used for CC. Additionally, CC was modeled using an ordinal logistic regression but, as was the case for RQ1 and RQ2, the results did not differ between these statistical approaches. The two time spans of interest were modeled as separate fixed effect variables in the same model using the following dummy coding: time span 1: pre-test = 0, final exam = 1, post-test = 0; time span 2: pre-test = 0, final exam = 0, post-test = 1. For each regression, student ID was modeled as a random intercept and the incentive condition (0 = regular credit, 1 = extra credit) and semester were modeled as control variables. Partial omega squared and odds ratios were used as effect sizes where appropriate. As described above, the two end-of-course assessment time points occurred within a couple of weeks of one another and the interval between them did not contain evolution instruction. Therefore, while the change in scores from the beginning of the semester to each end-of-course time point likely reflects evolution learning, differences in a student's *magnitude* of change between time span 1 and time span 2 does not. Rather, magnitude differences between the two time spans reflect differences in the inferences about a student's learning that a researcher could potentially make depending on when they chose to administer the end-of-course assessment.

To evaluate whether the results from this analysis generalized across demographic groups, the above models were modified to include an interaction effect between (a) time span 1 and gender (not male = 0, male = 1) and (b) time span 2 and gender, (c) time span 1 and race/ethnicity group (White = 0, Asian = 1, URM = 2), (d) time span 2 and race/ethnicity group. Interaction effects were also modeled between each time span and prior biology as a control variable. See Additional file 1: Table S4 for the exact sample sizes for these analyses.

Results

RQ1: (Do parallel ACORNS items administered at the *same* end-of-course assessment time point (i.e., during the final exam) but with *different* participation incentives (i.e., regular credit vs. extra credit) produce statistically comparable knowledge and misconception patterns and do these findings generalize across demographic groups?). There was no significant difference between the regular and extra credit incentive condition for either ACORNS CC (Fig. 1) or MIS (Fig. 2) scores for both the plant loss and animal gain items (see Table 2 for detailed results[5]). Furthermore, there was no significant interaction effect between the incentive condition and the race/ethnicity or gender of the respondents for either item (Additional file 1: Table S7[6]). Therefore, the

[5] The results for RQ1.1 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S6).

[6] The results for RQ1.2 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S8).

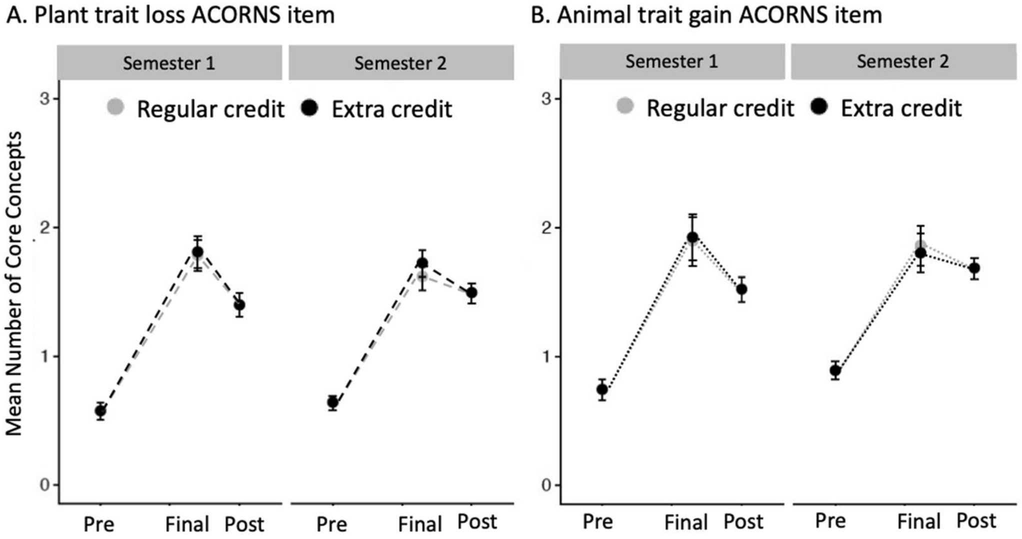

Fig. 1 Mean number of core concepts (CC) at the pre-test, final exam, and post-test for the plant item (**A**) and animal item (**B**) in both the regular and extra credit participation incentive conditions. The students in these plots reflect those who completed the assessments at all three time points. In the interest of space, please note that the x-axis of the graphs is not to scale and thus, does not represent the chronological time distance between each test administration time point. The pre-test and final exam were 13–14 weeks apart and the final exam and post test were less than two weeks apart

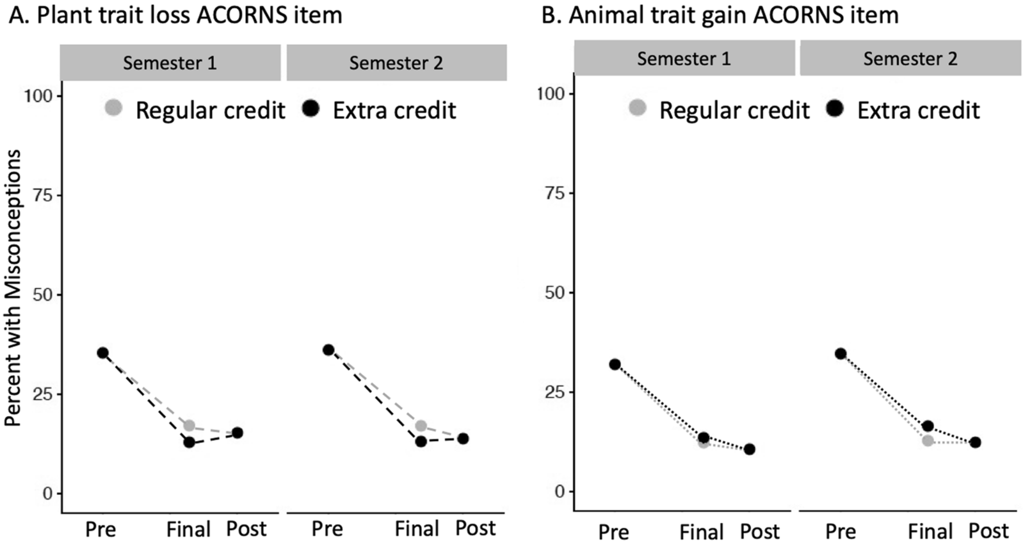

Fig. 2 Percent of respondents with at least one misconception (MIS) at the pre-test, final exam, and post-test for the plant item (**A**) and animal item (**B**) in both the regular and extra credit participation incentive conditions. The students in these plots reflect those who completed the assessment at all three time points. In the interest of space, note that the x-axis of the graphs is not to scale and thus, does not represent the chronological time distance between each test administration time point. The pre-test and final exam were 13–14 weeks apart and the final exam and post-test were under two weeks apart

incentive condition did not appear to impact ACORNS scores for any race/ethnicity or gender group.

RQ2: (Do parallel ACORNS items administered at *different* end-of-course assessment time points (i.e., final exam vs. end of semester) produce statistically comparable knowledge, misconception, and normative reasoning and do these findings generalize across demographic groups?) For both the plant loss and animal gain ACORNS item, there was no significant difference in MIS (Fig. 2) or MODC (Fig. 3) scores between the final exam and post-test. There was a significant difference in CC scores between the final exam and

Table 2 Results for comparison between regular and extra credit participation incentive conditions (RQ1.1)

Score	Condition	Plant Loss	Animal Gain
Core Concepts (CC)	Extra Credit (Final exam) vs. Regular credit (Final exam)	$B = 0.074$, $\beta = 0.071$, $\omega^2_p = 4.88e-04$ (NS)	$B = -0.020$, $\beta = -0.019$, $\omega^2_p = -0.001$ (NS)
Misconceptions (MIS)		$B = -0.313$, $OR = 0.73$ (NS)	$B = 0.106$, $OR = 1.11$ (NS)

Odds ratio (OR): small = 1.68 (0.59), medium = 3.47 (0.29), large = 6.7 (0.15) (Chen et al. 2010)

Partial omega squared (ω^2_p): small = 0.01, medium = 0.06, large = 0.14 (Lakens 2013)

NS not significant

* $p \leq 0.01$, $^{**}p \leq 0.001$, $^{***}p < 0.0001$

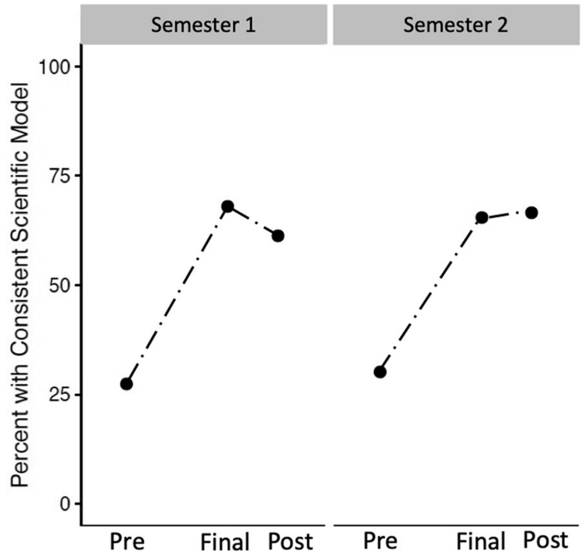

Fig. 3 Percentage of respondents with a consistent scientific model at the pre-test, final exam, and post-test. The students in these plots reflect those who completed the assessment at all three time points. In the interest of space, note that the x-axis of the graphs does not represent the chronological time distance between each test administration time point. The pre-test and final exam were 13–14 weeks apart and the final exam and post test were under two weeks apart

the post-test but the effect size was considered to be small (Fig. 1, Table 3[7]). In addition, there were no significant interaction effects between the end-of-course time point (i.e., final exam vs. post-test) and the race/ethnicity or gender of respondents for any of the measures of evolution understanding (i.e., CC, MIS, or MODC) (Additional file 1: Table S9[8]).

RQ3: (Does the timing and participation incentive condition of the ACORNS end-of-course assessment impact inferences about magnitudes of student learning and does this pattern generalize across demographic groups?) For both the plant loss and animal gain items, there was a significant and large *increase* in CC scores (Fig. 1[9]) and moderate *increase* in MODC scores (Fig. 3) from the pre-test to both the final exam and the post-test (Table 4). Additionally, there was a significant and moderate *decrease* in MIS scores (Fig. 2) from the pre-test to both the final exam and the post-test (Table 4). These patterns were replicated in both semesters studied. The effect size of the changes in ACORNS scores from the beginning to the end of the semester were similar regardless of the chosen end-of-course time point. Therefore, the measurement of evolution learning was similar regardless of the chosen end-of-course assessment time point. Furthermore, there was no significant interaction effect between the end-of-course assessment time point and the race/ethnicity or gender of the respondent (Additional file 1: Table S10[10]).

Discussion

Formative and summative assessments are fundamental components of student-centered teaching and learning (NRC 2001, 2007), and faculty administering these assessments are frequently faced with choices about when and how to collect data from students. Educators and administrators in higher education settings would therefore benefit from evidence-based guidelines to inform these choices. Although many studies have investigated the impact of testing conditions on assessment scores, few have done so using constructed-response items.

[7] The results for RQ2.1 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S6).

[8] The results for RQ2.2 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S8).

[9] The results for RQ3.1 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S6).

[10] The results for RQ3.2 did not differ when CC was modeled as as ordinal variable (Additional file 1: Table S8).

Table 3 Results for the comparison of evolution knowledge scores between the final exam and the post-test (RQ2.1)

Score	Timing	Plant loss	Animal gain
Core concepts (CC)	Final exam vs Post-test	$B = -0.352$***, $\beta = -0.168$, $\omega^2_p = 0.05$ (small)	$B = -0.253$***, $\beta =$ -0.117, $\omega^2_p = 0.03$ (small)
Misconceptions (MIS)		$B = 0.483$, $\beta = 0.664$, OR = 1.62 (NS)	$B = -0.557$, $\beta = -0.790$, OR = 0.57 (NS)
Model Consistency (MODC)		$B = -0.155$, $\beta = -0.167$, OR = 0.86 (NS)	

Odds ratio (OR): small = 1.68 (0.59), medium = 3.47 (0.29), large = 6.7 (0.15) (Chen et al. 2010)

Partial omega squared (ω^2_p): small = 0.01, medium = 0.06, large = 0.14 (Lakens 2013)

NS not significant

* $p \leq 0.01$, **$p \leq 0.001$, ***$p < 0.0001$

Table 4 Comparison of evolution learning from the pre-test to the final exam vs. the pre-test to the post-test (RQ3.1)

Score	Timing	Plant loss	Animal gain
Core concepts (CC)	Pre-test vs Final exam	$B = 1.214$***, $\beta = 0.534$, $\omega^2_p = 0.25$ (large)	$B = 1.071$***, $\beta = 0.451$, $\omega^2_p = 0.19$ (large)
Misconceptions (MIS)		$B = -1.536$***, $\beta = -1.758$, OR = 0.22 (medium)	$B = -1.226$***, $\beta = -1.452$, OR = 0.29 (medium)
Model Consistency (MODC)		$B = -1.536$***, $\beta = -1.758$, OR = 0.22 (medium)	
Core Concepts (CC)	Pre-test vs Post-test	$B = 0.905$***, $\beta = 0.398$, $\omega^2_p = 0.25$ (large)	$B = 0.839$***, $\beta = 0.353$, $\omega^2_p = 0.21$ (large)
Misconceptions (MIS)		$B = -1.373$***, $\beta = -1.571$, OR = 0.25 (medium)	$B = -1.547$***, $\beta = 1.833$, OR = 0.21 (medium)
Model Consistency (MODC)		$B = -1.373$***, $\beta = 1.571$, OR = 0.25 (medium)	

Odds ratio (OR): small = 1.68 (0.59), medium = 3.47 (0.29), large = 6.7 (0.15) (Chen et al. 2010)

Partial omega squared (ω^2_p): small = 0.01, medium = 0.06, large = 0.14 (Lakens 2013)

* $p \leq 0.01$, **$p \leq 0.001$, ***$p < 0.0001$

It therefore remains an open question as to whether implementation biases characteristic of closed-response assessments generalize to more contemporary item types (cf. AAAS 2011; NRC 2012). The Next Generation Science Standards and Vision and Change emphasize that undergraduate educators should move away from recognition-based approaches to science assessment and towards more authentic performance tasks such as model building, explaining, and arguing using evidence (AAAS 2012; NRC 2012). As more biology instructors heed these guidelines, questions about how to administer constructed-response assessments like the ACORNS in ways that minimize bias become increasingly important. Ultimately, understanding how best to foster student understanding of core biology topics such as evolution will rely on both the quality of assessment tools and the administration procedures that minimize bias.

In this study, two test administration conditions that instructors routinely employ–participation incentives (in this case, regular credit vs. extra credit) and end-of-course time point (final exam vs post-test)–were studied using the ACORNS instrument. A quasi-experimental design in which students were randomly assigned to a treatment condition was used. The findings indicated that the variations in these two administration conditions did not meaningfully impact inferences about evolution *understanding*; all differences between conditions were either insignificant or, if significant, considered to be small effect sizes. Furthermore, these administration conditions did not meaningfully impact inferences about evolution *learning* in terms of reasoning approach, increases in core concepts, or declines in misconceptions. Importantly, these findings were consistent across race/ethnicity and gender groups.

Prior work on the impact of testing conditions on closed-response assessment scores have produced results that both align with and diverge from those presented here. For example, Smith et al. (2012) reported that scores on a True/False genetics assessment did *not* differ between the biology majors who took it on the last day of the course and those who took it as part of the final exam (with no relevant interceding instruction). However, students were told that they would receive extra credit only if they scored 100% on the assessment. Although this work conflated incentive-related and timing-related testing conditions, the conclusions were similar to those in the present study; the end-of semester time point did not meaningfully impact instrument scores.

Ding et al. (2008) tested a variety of incentive conditions on a multiple-choice physics assessment using a cross-sectional design. In contrast to Smith et al., Ding et al. found that some conditions they tested *were* associated with different instrument scores. In particular, the extra credit and regular credit incentive conditions produced significant differences. However, the Ding et al. (2008) study has several design limitations, notably (i) a lack of controls for pre-test measures as well as background variables (all of which may differ among the students in cross-sectional study designs), and (ii) the conflation of multiple testing conditions (i.e., timing and incentives). Thus, the Ding et al. study lacks many of the controls used in the current study.

The only study in our literature review that tested administration condition effects on a constructed-response evolution test found remarkably similar results to those presented in this study. Specifically, Nehm and Reilly (2007) administered a constructed-response item about evolutionary change at two end-of-semester time points (one week apart: as an extra credit item on a post-test and as an extra credit item on a final exam). Responses were scored for seven key concepts of evolution (three of which overlapped with the ACORNS core concepts used in this study) and six misconceptions (three of which overlapped with the ACORNS misconceptions used in this study). In alignment with the findings reported in the present study, Nehm and Reilly found that the number of misconceptions did not differ between administration time points, but students used significantly fewer evolutionary concepts in the post-test as compared to the final exam. As in our study, the size of this difference (i.e., average of 0.5 key concepts) was relatively small.

As one might predict, the assessment time points used for each of our research questions were associated with different participation rates (96% of students completed the final exam items [RQ1], whereas 85% completed both the final exam *and* post-test items [RQ2], and 76% completed all three assessments [RQ3]). Although these three participation rates are generally very high for introductory biology settings, Ding et al. (2008) also reported a lower participation rate for some assessment conditions. Because reduced participation coincided with significantly different assessment scores in their study, the authors concluded that the different conditions attracted differently motivated "fractions" of the class. Although the present study was not designed to investigate student motivation, our findings do not align with this conclusion because none of the analyses (regardless of participation rate) resulted in meaningful differences between conditions. More specifically, performance on the ACORNS items in both the incentive and timing conditions led to similar inferences about the magnitudes of evolution learning in the course. Additionally, because the percentages of URM and male students who completed each assessment were similar (as shown in Additional file 1: Table S5 [RQ1 vs. RQs 2–3]), it does not appear that participation motivation in our sample was strongly related to gender, race/ethnicity, or assessment outcomes.

Directions for future work on testing conditions in biology education

Three areas of work on testing conditions in biology education would benefit from further attention: (1) anchoring research in relevant conceptual and theoretical frameworks, (2) conceptualizing the array of possible testing condition dimensions and studying them independently, (3) implementing longitudinal study designs, and (4) analyzing a broader array of assessment types and time points. These four points are discussed below.

Most of the studies of testing conditions in biology contexts that we reviewed were not anchored in explicit conceptual or theoretical frameworks (cf. Nehm 2019; see also Sbeglia et al. 2021). In other educational fields, in contrast, the impact of test incentives on assessment scores have been grounded in motivation-related perspectives, such as variations of the Expectancy-Value Framework (for examples, see Eccles 1983; Duckworth et al. 2011; Wigfield and Eccles 2000; Wise and DeMars 2005; for an exception in biology education see Uminski and Couch 2001). There are other categories of testing conditions beyond test incentives that could also introduce construct-irrelevant variation (e.g., "noise") to biology assessment scores that have not been explicitly situated within appropriate frameworks (e.g., assessment timing, assessment administration within an exam or independent of exams). Future work should ground empirical studies within theoretical models that seek to explain (rather than only test for) testing condition outcomes.

Many testing conditions that have the potential to impact assessment scores have not been clearly defined in biology education, which may explain why prior work has conflated distinct categories of conditions instead of isolating salient dimensions within categories. For example, participation incentive is a category of assessment condition that can differ along many axes (e.g., regular credit vs. extra credit; no incentive vs. small incentive vs. large incentive; scored for accuracy vs. scored for completion). The present study focused on only one dimension of incentive condition–regular credit vs. extra credit (and controlled for these other dimensions). The development of a matrix of possible conditions would allow more complete testing and prevent weak research designs (e.g., conflating testing conditions).

Studying how a broader array of testing dimensions impact student performance would be valuable because interaction effects among conditions are likely. Some participation incentives, for example, have been proposed as contributors to the perceived "stakes"[11] of a test (low vs. high) which in turn influence students' test-taking motivation[12] (Cole et al. 2008; Ding et al. 2008; Duckworth et al. 2011; Wise and DeMars 2005). Test taking motivation, in turn, may impact assessment scores and learning inferences (Wise and DeMars 2005). Although studying these interactions in controlled settings may be possible, some of these conditions may not apply to real classroom settings. For example, a testing condition with a required post-course test does not align with standard university practices (e.g., requiring students to complete an assessment after a course has been completed). Nevertheless, many more testing conditions need to be investigated.

Future studies of test administration conditions should include the collection of datasets that permit longitudinal analyses of how these conditions impact inferences about changes in response to instruction. For many educators in undergraduate settings, the goal of pre-post assessment is to understand how instruction impacts learning objectives. At present, it is unclear if prior findings about the impacts of test administration conditions from static datasets translate longitudinally. Our findings suggest that we should not necessarily expect that they will. For example, although the number of ACORNS CC scores differed significantly between the two end-of-course time points, these time points generated similar magnitudes of pre-post change; for both end-of-course time points, assessment scores indicated that significant and large magnitudes of learning occurred in the two semesters.

Finally, it is important to emphasize that faculty inferences about student understanding can be derived from formative and/or summative assessments. These assessment artifacts can vary widely (e.g., take-home assignments, in-class writing tasks). This study examined only more traditional testing approaches for summative assessment purposes (i.e. Did students learn evolution in this course?). Analyzing a broader array of assessment approaches (formative, summative) and types (assignments, in-class tasks) would be a valuable direction for future research. In addition, many studies in higher education tend to focus on beginning and end-of-course time points (pre-post). Yet numerous assessment events occur throughout a course, and future work should therefore not be restricted to traditional summative testing time points.

Study limitations

This study focused on two test administration conditions and their impacts on ACORNS scores: test participation incentive and end-of-course timing. Several limitations apply to these study conditions.

Participation incentives

The analyses for RQ1 focused on one dimension of participation incentive–the extra credit vs. regular credit dimension–and controlled for the size of the incentive (i.e., the amount of credit given was held constant). However, another dimension of participation incentive condition–the extra credit scoring procedure (e.g., graded for completion vs. graded for accuracy)–was not explicitly mentioned to students. Furthermore, because

[11] A low stakes assessment is one in which "there are typically no consequences associated with student performance, and many students perceive no personal benefit from the assessment testing experience" (Wise and DeMars 2005, p. 2).

[12] The process whereby one gives their "best effort to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test." (Wise and DeMars, 2005, p. 2).

only the final exam included both an extra credit and regular credit incentive for the ACORNS items (the pre-test and post-test included only an extra credit incentive for these items), the study design was unbalanced for this assessment condition, which precluded us from answering questions about how the participation incentive interacted with end-of-semester timing (or impacted inferences about pre-post learning). Answering these questions would require a "regular credit" post-test, which was not realistic or appropriate in our–and perhaps most–instructional settings. Simply put, requiring students to complete an assessment after a course has been completed would be unusual. The participation incentive was included in our analyses as a control variable (for RQ2 and RQ3) to account for potential impacts. Evolutionary knowledge outcomes for both research questions were not significantly associated with the participation incentive (which aligns with the finding that participation incentive did not impact final exam scores).

End-of-course timing

Many studies of the impact of test administration conditions on assessment scores conflate multiple conditions (e.g., Ding et al. 2008, Smith et al. 2012). The present study was designed to tease apart two test administration conditions: participation incentive and end-of-course assessment timing. This design goal focused on the ACORNS items themselves (i.e., two participation incentives for the ACORNS were studied while controlling for test timing and two test time points for the ACORNS were studied while controlling for participation incentive). However, at each time point, the two ACORNS items were situated within a broader assessment and this broader assessment may have inadvertently conflated the participation incentive and the timing. Specifically, the final exam as a whole was a *required* assessment (in which ACORNS items were randomly assigned as either extra credit or regular credit). In contrast, the post-test as a whole was a purely *voluntary* assessment. Therefore, although the design of the ACORNS items themselves effectively controlled for participation incentive condition across these time points (and vice versa), the design of the broader assessments did not. Whether the participation incentive of the broader test impacts scores even when it differs from the incentive of the ACORNS items themselves (e.g., an extra credit ACORNS item within a required final exam) is not clear but taking our findings in combination with those of Nehm and Reilly (2007) suggests that it might be. Specifically, the results of these two studies collectively suggest that administering voluntary ACORNS items within a required test (e.g., a final exam) vs. administering them within a voluntary assessment may indeed impact ACORNS scores; although both studies employed opposite assessment sequencing for their voluntary and required assessments (Nehm and Reilly administered their voluntary assessment *before* their required test whereas in this study, it was administered *after* the required test), these two studies nonthe-less found that students scored consistently lower on the voluntary assessment. This consistent pattern of student performance reported in both studies may therefore be better explained by the participation incentive of the broader assessment rather than by the participation incentive of the items themselves or by the timing/squencing of test administration. Regardless, both studies found that the few differences between administrations were small or not significant, resulting in similar inferences about the magnitude of evolution learning.

Interpreting effect sizes

Although the benchmarks for small, medium, and large effects are generally well accepted for many effect size measures, interpretation frameworks differ and exactly how to use these standards to draw inferences varies in the literature. For example, published effect size benchmarks have been conceptualized as minimum values for each level of effect (e.g. a medium effect size benchmark of 0.6 could imply that only values above this benchmark be classified as a medium effect; e.g., Olejnik and Algina 2000). Conversely, effect sizes can be interpreted based on which published benchmark an effect size value is closest to (e.g. Olejnik and Algina 2000). Our prior work uses the former interpretation framework (e.g., Sbeglia and Nehm 2018), which we maintain in the present study so as not to bias interpretation from study to study. Regardless, these interpretation discrepancies in the literature more broadly indicate that authors and readers should be careful in how definitively they position effect size claims.

Conclusion

Evolution is a core concept of biological and scientific literacy, and for this reason assessing the impact of undergraduate coursework on learning outcomes is paramount (AAAS 2011). Variations in incentive conditions and end-of-course assessment timing did not meaningfully impact ACORNS scores or inferences about student learning for any race/ethnicity or gender group in this study. Therefore, the measurement of evolution understanding *and* evolution learning using the ACORNS was generally robust to these conditions and suggests that educators may have some flexibility in terms of when and how they design their ACORNS assessment conditions. As the types of assessments utilized in higher education evolve in their sophistication, so too must studies of how testing conditions impact inferences about student learning outcomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12052-022-00166-2.

Additional file 1. Supplementary Information.

Acknowledgements
We thank the Howard Hughes Medical Institute Inclusive Excellence science education award to Stony Brook University (Nehm, PI) for partial funding of this work. The authors alone are responsible for all content in this article.

Author contributions
Study conceptualization: RHN and GCS, Data collection: GCS and RHN, Data analysis: GCS, Data interpretation: GCS and RHN, Writing of manuscript: GCS and RHN. All authors read and approved the final manuscript.

Funding
This study was funded by the Howard Hughes Medical Institute Inclusive Excellence science education award.

Availability of data and materials
Should our manuscript be accepted, we will archive the data used in this study in our university's library data archiving system, which is a publicly accessible repository.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Ecology and Evolution Stony, Brook University, Stony Brook, NY 11794, USA. [2]Program in Science Education, Stony Brook University, Stony Brook, NY 11794, USA. [3]Department of Biology, San Diego State University, San Diego, USA.

Received: 31 December 2021 Accepted: 2 May 2022
Published online: 20 June 2022

References

American Association for the Advancement of Science. Vision and change in undergraduate biology education: a call to action. Washington, DC: Directorate for Biological Sciences; 2011.

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST. Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sci Educ. 2011;10:394–405.

Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, Dai B. lme4: Linear mixed-effects models using 'eigen' and S4. R package. 2021.

Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ. Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance? J Sci Educ Technol. 2014;23:160–82.

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

Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological Studies. Commun Stat. 2010. https://doi.org/10.1080/03610911003650383.

Christensen RHB. Ordinal: Regression models for ordinal data. R package. 2019.

Cole JS, Bergin DA, Whittaker TA. Predicting student achievement for low stakes tests with effort and task value. Contemp Educ Psychol. 2008;33:609–24.

Couch B, Knight J. A comparison of two low-stakes methods for administering a program-level biology concept assessment. J Microbiol Biol Educ. 2015;16:178–85.

DeMars CE. Test stakes and item format interactions. Appl Meas Educ. 2000;13:55–77.

Ding L, Reay NW, Lee A, Bao L. Effects of testing conditions on conceptual survey results. Phys Rev Spec Top Phy Educ Res. 2008. https://doi.org/10.1103/PhysRevSTPER.4.010112.

Duckworth AL, Quinn PD, Lynam DR, Loeber R, Stouthamer-Loeber M. Role of test motivation in intelligence testing. PNAS. 2011;108:7716–20.

Eccles J. Expectancies, values, and academic behaviors. In: Spence JT, editor. Achievement and achievement motives. San Francisco: Freeman; 1983. p. 75–146.

Federer MR, Nehm RH, Opfer J, Pearl D. Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations. Res Sci Educ. 2014;45:4.

Federer MR, Nehm RH, Pearl D. Examining gender differences in written assessment tasks in biology: A case study of evolutionary explanations. CBE Life Sci Educ. 2016;15:1.

Furrow RE, Hsu JL. Concept inventories as a resource for teaching evolution. Evol Educ Outreach. 2019;12:2.

Gregory TR. Understanding natural selection: Essential concepts and common misconceptions. Evol Educ Outreach. 2009;2:156–75.

Ha M, Nehm RH. The impact of misspelled words on automated computer scoring: a case study of scientific explanations. J Sci Educ Technol. 2016;25:358–74.

Ha M, Wei X, Wang J, Nehm RH. Chinese pre-service biology teachers' evolutionary knowledge, reasoning patterns, and acceptance levels. Int J Sci Educ. 2019;41:628–51.

Haudek KC, Prevost LB, Moscarella RA, Merrill J, Urban-Lurain M. What Are They Thinking? Automated Analysis of Student Writing about Acid-Base Chemistry in Introductory Biology. CBE Life Sci Educ. 2012. https://doi.org/10.1187/cbe.11-08-0084.

Huffman D, Heller P. What does the force concept inventory actually measure? Phys Teach. 1995;33:138–43.

Kalinowski ST, Leonard MJ, Taper ML. Development and validation of the Conceptual Assessment of Natural Selection (CANS). CBE Life Sci Educ. 2016;15:4.

Kampourakis K. Understanding Evolution (2nd ed., Understanding Life). Cambridge: Cambridge University Press. 2020. Doi: https://doi.org/10.1017/9781108778565

Kelemen D. Teleological minds: How natural intuitions about agency and purpose influence learning about evolution. In: Rosengren K, Brem SK, Evans EM, Sinatra GM, editors. Evolution challenges: Integrating research and practice in teaching and learning about evolution. Oxford: Oxford University Press; 2012.

Klymkowsky MW, Garvin-Doxas K, Zeilik M. Bioliteracy and teaching efficacy: what biologists can learn from physicists. Cell Biol Educ. 2003;2:155–61.

Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. Front Psychol. 2013. https://doi.org/10.3389/fpsyg.2013.00863.

Mead LS, Kohn C, Warwick A, Schwartz A. Applying measurement standards to evolution education assessment instruments. Evol Educ Outreach. 2019;12:5.

Moharreri K, Ha M, Nehm RH. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. Evol Educ Outreach. 2014;7:15.

National Research Council. Knowing what students know. Washington, D.C.: National Academies Press; 2001.

National Research Council. Knowing what students know: the science and design of educational assessment. Washington, D.C.: National Academies Press; 2001.

National Research Council. Taking science to school: learning and teaching science in grades K-8. Washington, D.C.: National Academies Press; 2007.

National Research Council. Framework for science education. Washington, D.C.: National Academies Press; 2012.

Nehm RH. Understanding undergraduates' problem solving processes. J Microbiol Biol Educat. 2010;11:119–22.

Nehm RH. Chapter 14: Evolution. In: Reiss M, Kampourakis K, editors. Teaching biology in schools. New York and London: Routledge; 2018. p. 164–77.

Nehm RH. Biology education research: Building integrative frameworks for teaching and learning about living systems. DISER. 2019;1:15.

Nehm RH, Schonfeld I. The future of natural selection knowledge measurement. J Res Sci Teach. 2010;47:358–62.

Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48:237–56.

Nehm RH, Mead L. Evolution assessment. Introduction to the special issue. Evol Educ Outreach. 2019;12:7.

Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection. Bioscience. 2007;57:263–72.

Nehm RH, Ridgway J. What do experts and novices "see" in evolutionary problems. Evol Educ Outreach. 2011;4:666–79.

Nehm RH, Beggrow E, Opfer J, Ha M. Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. Am Biol Teach. 2012a;74:92–8.

Nehm RH, Ha M, Mayfield E. Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. J Sci Educ Technol. 2012b;21:183–96.

Olejnik S, Algina J. Measures of effect size for comparative studies: Applications, Interpretations, and Limitations. Contemp Educ Psychol. 2000;25:241–86.

Opfer J, Nehm RH, Ha M. Cognitive foundations for science assessment design: Knowing what students know about evolution. J Res Sci Teach. 2012;49:744–77.

Rector M, Nehm RH, Pearl D. Learning the language of evolution: Lexical ambiguity and word meaning in student explanations. Res Sci Educ. 2013;43:1107–33.

Sbeglia GC, Goodridge JH, Gordon LH, Nehm RH. Are faculty changing? How reform frameworks, sampling intensities, and instrument measures impact inferences about student-centered teaching practices. CBE Life Sci Educ. 2021. https://doi.org/10.1187/cbe.20-11-0259.

Sbeglia GC, Nehm RH. Measuring evolution acceptance using the GAENE: influences of gender, race, degree-plan, and instruction. Evol Educ Outreach. 2018;11:18.

Smith MK, Jones FHM, Gilbert SL, Wieman CE. The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. CBE Life Sci Educ. 2013. https://doi.org/10.1187/cbe.13-08-0154.

Smith JI, Tanner K. The problem of revealing how students think: Concept inventories and beyond. CBE Life Sci Educ. 2010. https://doi.org/10.1187/cbe.09-12-0094.

Smith M, Thomas K, Dunham M. In-class incentives that encourage students to take concept assessments seriously. J Coll Sci Teach. 2012;42:57–61.

Stains M, Harshman J, Barker MK, Chasteen SV, Cole R, DeChenne-Peters SE, et al. Anatomy of STEM teaching in American universities: a snapshot from a large scale observation study. Science. 2018;359:1468–70.

Uminski C, Couch BA. GenBio-MAPS as a case study to understand and address the effects of test-taking motivation in low-stakes program assessments. CBE Life Sci Educ. 2001. https://doi.org/10.1187/cbe.20-10-0243.

Wigfield A, Eccles J. Expectancy-value theory of achievement motivation. Contemp Educ Psychol. 2000;25:68–81.

Wise SL, DeMars CE. Low examinee effort in low-stakes assessment: problems and potential solutions. Educ Assess. 2005;10:1–17.

Wolf LF, Smith JK. The consequence of consequence: Motivation, anxiety, and test performance. Appl Meas Educ. 1995;8:227–42.

Wolf LF, Smith JK, DiPaolo T. The effects of test specific motivation and anxiety on test performance. Paper presented at the annual meeting of the National Council on Measurement in Education, New York. 1996.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.