Zuckerman and Sbeglia *Evolution: Education and Outreach* (2024) 17:18
https://doi.org/10.1186/s12052-024-00210-3

Evolution: Education and Outreach

RESEARCH Open Access

Measuring undergraduate evolution learning using the CANS: psychometric strengths and limitations

Austin L. Zuckerman[1,2*] and Gena Sbeglia[2,3]

Abstract

Background Evolution continues to be one of the most difficult biological topics to teach, warranting innovative pedagogical tools and assessment strategies for enhancing evolutionary instruction. A major advance in measuring the evolution knowledge of undergraduate students came with the development of the Conceptual Assessment of Natural Selections (CANS). In this study, we use the CANS to measure knowledge and learning of natural selection in a large ($N > 6000$) sample of undergraduate students to expand upon prior validity testing of this instrument and advance knowledge of student evolutionary reasoning. We apply the Rasch measurement framework to examine if the CANS productively measures the intended construct and investigate the patterns of knowledge and learning about evolution among students with different backgrounds and demographic characteristics.

Results While a unidimensional Rasch model demonstrated acceptable reliabilities and fit for most of the CANS items, some items showed problematic fit statistics and were resistant to instruction. The instrument items also did not span the full range of student abilities, which suggests relatively low measurement precision. Our large sample also allowed rigorous tests of multidimensionality, revealing the presence of multiple dimensions or constructs, some of which may not be intentional. These results generated specific item-level recommendations for improving this instrument. Using Rasch measures to examine learning patterns, we found that pre-test evolution knowledge was low but that there were high learning gains by the end of the course. However, some concept categories were found to be more difficult than others, suggesting the need for more attention to these areas by instructors. We also identified pre-test disparities in evolutionary knowledge by socially defined race and biological sex, yet students from all groups achieved comparable learning gains at the end of the course.

Conclusion The CANS holds great potential to generate critical insights about student evolutionary reasoning and provide information about which instructional approaches most effectively mitigate the notable knowledge disparities among students. We leverage the findings of this study to propose tangible ways in which this instrument may be improved in order to better achieve both of these goals.

Keywords Natural selection, Evolution education, Learning, Rasch analysis, Validity, Race, Gender

*Correspondence:
Austin L. Zuckerman
azuckerm@ucsd.edu
[1]Program in Mathematics and Science Education, University of California San Diego, La Jolla, CA, USA
[2]Program in Mathematics and Science Education, San Diego State University, San Diego, CA, USA
[3]Biology Department, San Diego State University, San Diego, CA, USA

Background

Evolution has been recognized as one of the core disciplinary ideas needed to understand the complexities of biological systems at multiple scales and has been identified as a central component of scientific literacy (AAAS 2011). This recognition has promoted the inclusion of evolution as a foundational concept within biology curricula at both the K-12 (NGSS Lead States 2013) and higher education levels (AAAS 2011). However, evolution is one of the most difficult biological topics to teach in part because it is rife with misconceptions that persist even after instruction (Bishop and Anderson 1990; Nehm and Reilly 2007; Andrews et al. 2011). Misconceptions about evolution have been documented across a variety of educational stages, including among secondary school students (Demastes et al. 1995), undergraduate students (e.g. Abraham et al. 2009; Andrews et al. 2011; Gregory 2009; Nehm and Reilly 2007; Petto and Mead 2008; Phillips et al. 2012), and practicing teachers (Nehm and Schonfeld 2007; Ziadie and Andrews 2018). Therefore, there is an urgent need for innovative pedagogical tools and assessment strategies for enhancing evolutionary instruction. A necessary step toward improving evolution education is the implementation of robust assessment tools that can effectively measure the progression toward learning outcomes and proficiency in this core disciplinary idea (Ziadie and Andrews 2018). However, the development of such tools remains a major challenge (Nehm and Mead 2019).

One of the most important topics within the domain of evolution is natural selection. In natural selection, heritable genotypic variants arise from random mutations and encode phenotypes that may confer an advantage in the environment, causing their frequency to increase in a population over time (Speth et al. 2014). As such, there are considered to be three core concepts of natural selection: variation, heredity, and differential survival and reproduction (Nehm et al. 2012). In spite of the longstanding challenges associated with monitoring and improving student understanding of natural selection, a major advance in measuring evolution knowledge in undergraduates came with the development of two easy-to-administer concept inventories: The Conceptual Inventory of Natural Selection (CINS) (Anderson et al. 2002) and the Conceptual Assessment of Natural Selections (CANS) (Kalinowski et al. 2016). Both instruments are closed response assessments with misconception distractors. The CANS was developed to improve upon several weaknesses of the CINS. In particular, the CANS addresses more misconceptions than the CINS and assesses evolution knowledge across multiple biological phenomena (e.g., trait gain in plants vs. trait loss in animals) using a variety of item forms (Kalinowski et al. 2016).

Student reasoning has been found to be impacted by the specific features of an evolutionary phenomenon. The authors of the CANS purposely included various features within the CANS in order to more effectively tap into the way that students reason about evolution. For example, students' evolutionary explanations frequently differ depending on whether the features of the phenomenon involve the gain vs. loss of a trait within an animal vs. plant taxon (e.g., Nehm and Ha 2011). However, evolution works the same way regardless of the trait's polarity or the taxon in which it occurs and thus does not necessitate different evolutionary explanations for phenomena with these features. This example demonstrates a fundamental distinction between novices and experts: novice reasoning tends to be fragmented across phenomena whereas expert reasoning tends to be coherent across phenomena (Kampourakis and Zogza 2008; Opfer et al. 2012). The degree of coherence or fragmentation across phenomena speaks to the structure of knowledge. Evidence indicates that it is more difficult to increase a student's coherence than it is to increase the magnitude of their knowledge (Colton et al. 2018, 2019).

A fundamental step in the development and evaluation of assessment tools is the gathering of various sources of evidence to support the validity of the interpretations made from instrument derived data (American Educational Research Association 2014). These sources of validity evidence allow the determination of how well an instrument measures what it intends to measure. There are many different conceptual frameworks for guiding the sources of evidence that must be gathered (AERA, 2014). Here, we adopt a construct validity framework, which includes a combination of several categories of evidence to support inferences about the predicted constructs (Messick 1995; Campbell and Nehm 2013; AERA, 2014). It is infeasible to capture the full range of evidence needed to establish construct validity from a single study and several gaps in validity evidence exist for the CANS. Furthermore, several of the existing validity findings are inconclusive, in part due to small sample sizes, and thus warrant further study. The purpose of this study is to address the limitations in existing validity evidence for this instrument in order to advance understanding of student reasoning about natural selection. Below, we describe the research aims and questions that guide this study. We then describe the five primary sources of validity evidence within the construct validity measurement framework and summarize the existing evidence for the CANS.

Research aims and questions

In this study, we use the 24-item CANS to measure student learning of natural selection in a large ($N>6000$) sample of undergraduate students enrolled in a gateway

biology course in order to expand upon prior validity testing of this instrument and advance knowledge of student evolutionary reasoning. Our study advances prior work by (1) investigating and broadening internal structure validity and generalization validity for the CANS, (2) including data from a significantly larger sample spanning 14 semesters, which allows both rigorous tests of multidimensionality as well as the replication of findings through time, (3) adopting a more stringent IRT model (Rasch model) as our paradigm of productive measurement, (4) including item level analysis to identify, interrogate, and propose solutions for problematic items, (5) evaluating the suitability of the instrument to effectively tap into student reasoning, and (6) incorporating student background and demographic variables into analyses of student evolution learning using this instrument. We divide this work into two parts, each with a corresponding set of research questions.

In part 1, we investigate whether the CANS productively measures the intended construct by assessing whether the instrument-derived data adhere to well-accepted criteria of robust measurement. Our criteria of robust measurement follow the Rasch model, which theorizes that certain characteristics of the underlying data must be present in order to generate robust measures of a latent construct (Boone 2016; Boone et al. 2014; Borsboom et al. 2003). More information about each of these criteria is provided in the Methods. The research questions are as follows: (RQ1.1) Do items that comprise the CANS display acceptable fit to the expectations of the Rasch model? (RQ 1.2) Does the CANS reliably order items by their difficulties and respondents by their abilities on the latent trait? (RQ1.3) To what extent does the CANS precisely measure the latent trait? (RQ 1.4) Is the structure of the CANS best characterized as unidimensional or multidimensional?

In part 2, we investigate the patterns of knowledge and learning about evolution as measured by the CANS. The research questions for this section are as follows: (RQ 2.1) What are the magnitudes of evolution knowledge and learning gains across 14 semesters of a high-enrollment gateway biology course? (RQ 2.2) How variable are CANS measures across semesters? (RQ 2.3) How variable are CANS measures across different student background characteristics? (RQ 2.4) Which evolution topics within the CANS are most difficult for students? (RQ 2.5) What is the structure (i.e., coherent vs. fragmented) of student evolutionary knowledge across phenomena?

Current validity evidence for the CANS

Within construct validity, the sources of validity evidence include: (i) test content (i.e., content validity), (ii) response processes (i.e., substantive validity), (iii) relationships to other variables (i.e., convergent and/or discriminant validity); (iv) internal structure (i.e., internal structure validity); and (v) validity generalization (i.e., generalization validity). Content validity refers to whether the instrument includes all parts of the intended construct and no irrelevant topics (AERA, 2014). The authors of the CANS specified the content domain of natural selection to include five concept categories–mutation, inheritance, selection, variation, and evolution. The fifth concept, evolution, was designed to assess student understanding of the interaction among the other core topics (Kalinowski et al. 2016). Expert interviews were used to support the relevance of the items to these subtopics (Kalinowski et al. 2016).

Substantive validity refers to whether respondents engage in the expected cognitive processes when answering instrument items (AERA, 2014). To address this source of validity evidence, the authors of the CANS conducted interviews with students and concluded that students interpreted the items as intended (Kalinowski et al. 2016).

Convergent validity refers to the relationships between an instrument's scores and measures originating from other sources that intend to measure the same construct. Several authors have reported that patterns of evolution knowledge and learning were similar when using the CANS as compared to the CINS (Anderson et al. 2002) and the ACORNS (Assessing COntextual Reasoning about Natural Selection, Nehm et al. 2012), which is a constructed response instrument measuring a similar construct (Nehm et al. 2022; Sbeglia and Nehm 2024).

Internal structure validity refers to the degree to which test items represent, or tap into, the intended construct. This source of validity evidence involves analyzing the relationships among items and investigating how they collectively contribute to the measurement of the construct (AERA, 2014). A measurement instrument with strong internal structure validity has items that are related to each other in a theoretically-aligned manner (e.g., items fall on a single dimension and vary in difficulty as hypothesized by theory) and are capable of productively measuring the intended construct (e.g., function the same way for all respondents). The authors addressed this form of validity evidence using Item Response Theory (IRT) in a sample of <300 students to assess the dimensionality and fit of the items. IRT can generate internal structure validity evidence (Boone 2016; Boone et al. 2014; Campbell and Nehm 2013) because it theorizes that certain characteristics of the underlying data must be present in order to generate robust measures of a latent construct (Boone 2016; Boone et al. 2014; Borsboom et al. 2003). More information about each of these criteria are provided in the Methods. The results of Kalinowski et al.'s IRT analysis showed that 18 of the 24 CANS items loaded onto a single dimension, and several

of the remaining items had improved factor loadings when restricted to the set of questions within the primary concept category (the concept categories of the CANS are mutation, inheritance, selection, variation, and evolution). Although this pattern suggests the possibility of multiple constructs, no clear multidimensional pattern was found and more robust multidimensional analyses were not possible due to the small sample size ($n=218$). The authors speculated that the lack of a clear multidimensional pattern could have been due to insufficient sample size, an insufficient number of items, or construct underrepresentation (meaning that the items did not fully tap into some of the constructs).

Generalization validity refers to the degree to which the results of a study can be generalized or extended to other contexts. This source of validity evidence is frequently gathered by analyzing the instrument's performance in various contexts such as different populations, samples, and even groups within samples. Generalization validity evidence for the CANS is minimal. Since the initial development of the CANS, no other study has investigated the instrument's psychometric properties in a new population.

Methods

Study setting

Data collection for this study took place in 14 semesters of a large-enrollment (>250 students/semester) undergraduate introductory biology course at a public, research-intensive university in the United States, classified as "very high research activity" (R1) by the Carnegie Classification of Institutions of Higher Education (McCormick and Zhao 2005). All participants were enrolled in a course that focused on evolutionary concepts as a central topic, including microevolutionary processes and macroevolutionary patterns. The prerequisites of this course included completion of high school biology and college mathematics, but no other prior biology coursework in higher education was required. The course welcomed both major and non-major students, with most enrolled students being in their first or second year at the university.

The course was designed to cover content aligned with five core concepts of biological literacy emphasized in the American Association for the Advancement of Science (AAAS)'s Vision and Change policy document (AAAS 2011). As these data were collected over multiple years, including before, during, and after the COVID-19 pandemic, the instructional style varied across different implementations. Among semesters, this course covered the same content and was taught by a consistent group of instructors who used the same lecture material. However, the format ranged from low to intermediate levels of evidence-based instruction (e.g. active learning, group work), with earlier semesters generally having the least levels of these practices (Nehm et al. 2022).

Participants

Students enrolled in the course were invited to participate in this study, which involved completing an online survey within the first two weeks of the semester (pre-test) and a another survey after the last day of classes (post-test). Students completed the survey asynchronously outside of the scheduled class time and were instructed not to use outside resources, such as the internet, textbook, or other people. Students making a good faith effort received full credit. As part of the survey, participants were prompted to complete the full 24-item CANS instrument and self-report demographic and background information, including biology courses taken, biological sex, socially defined race, English Learner status, PELL eligibility, and college generation status.

Out of the 8599 students enrolled in the course across the fourteen semesters, there was a total of 13,568 consenting responses, 6762 at the pre-test (78.6% participation) and 6806 at the post-test (79.1% participation) with 6483 students completing both surveys. The data were subsequently reduced to exclude participants who (1) received a perfect score on the pre-test or (2) spent less than ten minutes on the survey. The final analyzed data set consisted of 12,876 responses (6500 for the pre-test and 6376 for the post-test). Demographic data for these respondents is summarized in Table 1.

Instrument

We implemented the CANS instrument (Kalinowski et al. 2016) without modification in all 14 semesters in which this study was conducted. The CANS consists of 24 multiple choice items designed to assess students' understanding of five concept categories related to evolution: *variation (three items)*, *selection (five items)*, *inheritance (four items)*, *mutation (four items)*, and how these processes interact in *evolution (eight items)*. The items are presented in four clusters that are focused on specific taxa: anteaters (8 items), bowhead whales (6 items), saguaro cacti (6 items), and mosquitoes (4 items). Each item has a single correct answer and between two and four distractors that address common misconceptions about evolution. Although the authors did not specifically align the items with the misconceptions they address, our interpretations of the items suggest that some of the misconceptions are: need-based reasoning, use and disuse of traits, nature as a selecting agent, and adaptations as exclusively chance-based (Gregory 2009). Higher scores indicate more evolution knowledge (incorrect answers are coded as "0" and correct answers are coded as "1").

In the original study, the five concept categories were conceptualized after developing a concept map around a

Table 1 Sample size and student demographic and background information

	Category	Consenting students	All students[2]
Initial sample size			
	Pre	6762	NA
	Post	6806	NA
Final sample size (after removing problematic responses)	Pre	6497	NA
	Post	6376	NA
Background Variables			
Race/Ethnicity	% American Indian/ Alaska Native	15 (< 1%)	< 1%
	% Asian	6091 (51%)	44%
	% Black/African American	710 (6%)	8%
	% Hispanic of any race	1282 (11%)	15%
	% Native Hawaiian/ Other Pacific Island	15 (< 1%)	< 1%
	% White	3848 (32%)	34%
Biological sex	% Female or non-binary	7555 (60%)	57%
College generation status[1]	% First generation	2814 (41%)	41%
PELL eligibility status[1]	% PELL Eligible	3465 (41%)	40%
Prior Biology	% No prior biology coursework	4507 (35%)	41%

[1]These variables were not gathered in all semesters

[2]Data for all students, including those who did not participate in the study, were available for only 12 of the 14 semesters

network of core concepts (see Fig. 1 in Kalinowski et al. 2016). Four of these concept categories (inheritance, variation, selection, and mutation) were at the center of this concept map and were connected to other core concepts, many of which were also assessed directly within the CANS (e.g. struggle for existence, exponential growth, environmental stress, chance events). The network of concepts converged on a fifth concept category, evolution, which assessed student understanding of the interaction of inheritance, variation, selection, and mutation. The instrument was developed to include other concepts related to natural selection (e.g., population exponential growth) while excluding more advanced evolutionary concepts, such as the molecular basis of evolution, or related topics like extinction (Kalinowski et al. 2016).

Analysis

Part 1: adherence of the CANS to criteria of robust measurement

For an instrument to generate robust measurement of a latent construct, specific characteristics of the instrument-derived data must be present (Borsboom et al. 2003). These characteristics are embodied within the psychometric approaches used to investigate and analyze the data. Different modeling approaches are best suited to different types of response data. Rasch analysis, and Item Response Theory (IRT) more broadly, are considered to be the most appropriate approaches for estimating continuous latent measures from ordinal response data (de Ayala 2019; Hambleton and Jones 1993; Linacre and Wright 1993; Neumann et al. 2011). Like Item Response Theory (IRT), the Rasch model adopts a probabilistic approach for estimating latent measures, establishing that the probability of correctly answering an item is based on one parameter; the person's ability/the item's difficulty (Hambleton and Jones 1993). In IRT, two additional parameters -item discrimination and pseudo guessing- can be added or removed to improve the model fit.

The original validation study of the CANS used a three parameter IRT model in which all three parameters –item difficulty/person ability, item discrimination,

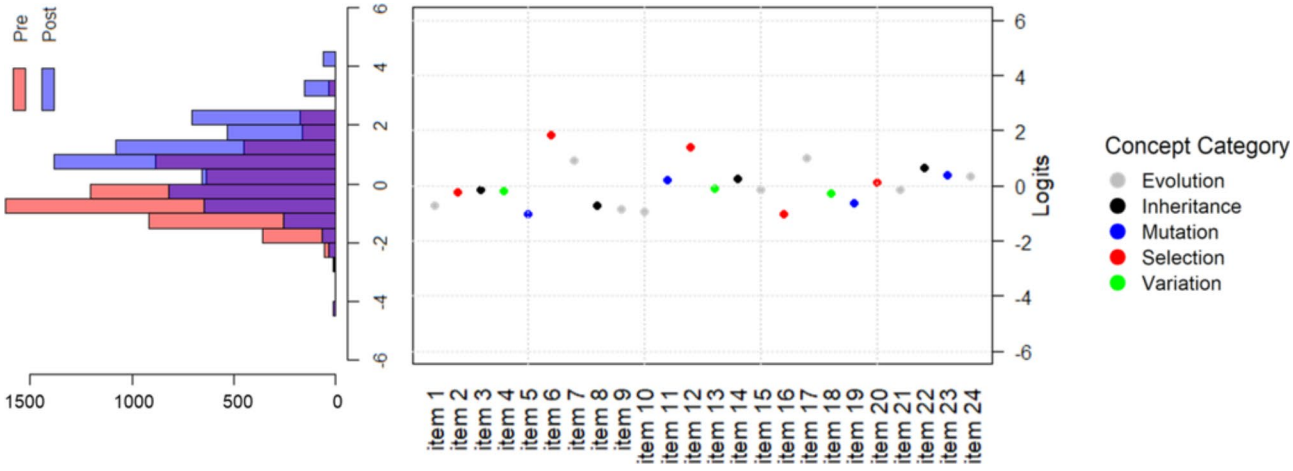

Fig. 1 Wright map showing Rasch-transformed person abilities for pre- and post-test (left) and Rasch-transformed item difficulties (right). The post-test item difficulties were anchored to the pre-test at model estimation

and guessing– were included. Quantitatively, the Rasch model is equivalent to a 1-parameter IRT model in the sense that it includes only one parameter (i.e., item difficulty/person ability). Theoretically, however, Rasch differs from IRT because it makes the strict assumption that no additional parameters are needed for productive measurement and thus are not included (Wright 1977; Boone et al. 2014; Stemler and Naples 2021). According to the Rasch modeling framework, if the data do not fit the Rasch model, then the response patterns of the instrument-derived data are considered to be inconsistent with robust measurement and new data need to be gathered and/or the instrument needs to be improved. This approach is in contrast to the three-parameter IRT model used by Kalinowski et al., which allows the addition and removal of parameters in order to attain the best fit to the data. A benefit of the stricter Rasch approach is that it is more parsimonious and calibrates instruments using an equivalent standard of robust measurement (Romine et al. 2017). Therefore, Rasch modeling was used in this study instead of a three-parameter IRT model because it allowed the evaluation of the CANS using a more robust, stable, and parsimonious benchmark.

One of the benefits of the Rasch model, is that it converts raw instrument data into a continuous, linear scale that can be analyzed using parametric statistics. The raw responses to the CANS items were converted into a continuous, linear scale using a dichotomous Rasch model in the R (v. 4.2-21) Package Test Analysis Modules (TAM) (Robitzsch et al. 2024). Accordingly, the model was specified as a '1PL' (1-parameter logistic) estimation. In contrast to the 3 parameter IRT model used by Kalinowski et al. (2016), item difficulty/person ability is the only parameter included in the Rasch model. To investigate item fit, reliability, person item alignment and dimensionality (see details below), a separate Rasch model was generated for the pre- and post-test using marginal maximum likelihood (MML) estimation. To compare Rasch person abilities between the pre and post test, the post-test item difficulties were anchored to the pre-test at model estimation and a marginal maximum likelihood (MML) estimation was used. Conversely, to compare Rasch item difficulties between the pre and post test, the post test person abilities were anchored to the pre-test at model estimation and a joint maximum likelihood estimation was used due to constraints of the MML function in handling fixed person ability parameters. The equal-interval measures resulting from the Rasch models are on a logit scale and contain information about both the item difficulty and person ability. Given the linear nature of these transformed measures, such measures are appropriate for use in subsequent parametric statistical analyses (Boone et al. 2014). Below we describe how we analyzed pre- and post-test CANS responses from 14 semesters for their adherence to four specific criteria of good measurement: (1) acceptable item fit, (2) acceptable item and person reliability, (3) acceptable person-item alignment (Wright maps), and (4) unidimensionality.

Item fit

To test whether the CANS items display acceptable fit to model expectations, we generated weighted (infit) and unweighted (outfit) mean squares (MNSQ) item fit statistics (RQ.1.1). MNSQ values of 1.0 indicate that the data fits the model as expected, whereas values above 1 may indicate data underfit (due to unmodeled variation) and values below 1 may indicate data overfit (due to redundancy in explained variation). Multiple choice items within a MNSQ fit range between 0.7 and 1.3 are considered productive for measurement but note that less conservative ranges (e.g., 0.5-1.5) have been recommended for Likert scale items Wright and Linacre 1994; Boone et al. 2014; Bond and Fox 2007). Items outside of these ranges do not fit the Rasch model and do not contribute to productive measurement of the latent constructs, which could be due to factors such as inconsistent interpretations of the items or items that represent multiple underlying constructs (Boone 2016). Fit values above 2.0 are interpreted as distorting or degrading to the measurement model (Boone et al. 2014).

Reliability

To test whether the CANS consistently ordered items by their difficulties and ordered respondents by their abilities, we calculated item and person reliability (RQ 1.2). Item reliabilities were calculated to evaluate the item hierarchy, which indicates whether the Rasch model is able to predictably separate items by their difficulties. Item reliability was calculated using the expected posteriori/plausible value reliability (EAP/ PV) index, which estimates whether the hierarchy of item difficulties could be replicated in a population of different individuals with similar person abilities.

Person reliabilities were also calculated to evaluate the extent to which the Rasch model was able to effectively distinguish between persons of different abilities (Bond and Fox 2007). Person reliability was calculated using the WLE separation index, which estimates whether the order of person abilities could be replicated with items of similar difficulties. Reliability values above 0.7 were considered acceptable for both the item and person reliability indices.

Person-item alignment using wright maps

The alignment of an instrument's difficulty to the ability of the sample in which it was administered indicates its level of measurement precision (Boone 2016; Boone et al. 2014). An instrument that is capable of precise

measurement will have varying item difficulties that span the full range of person abilities, allowing for respondents to be differentiated meaningfully based on their abilities. The criterion that person abilities and item difficulties should align is analogous to the expectation that a ruler has a sufficient number and location of tick marks to meaningfully differentiate individuals in the measured population. Therefore, to test whether the CANS is capable of precisely measuring the latent trait (RQ 1.3), item difficulties from a unidimensional Rasch model were plotted against the associated person abilities using a Wright map. Respondents at the top of the Wright map (higher logit measures) are interpreted to have higher abilities on the latent trait (i.e., more evolution knowledge), while respondents at the bottom of the map (lower logit measures) are interpreted as having lower abilities (i.e., less evolution knowledge). Correspondingly, items with higher logit measures are interpreted as being higher in difficulty, while those with lower logit measures are interpreted as lower in difficulty. The Wright map was used to visualize the extent to which the difficulties of items in the CANS spanned the abilities of the respondents. The instrument lacks measurement precision if the item difficulties do not adequately span the respondent abilities (Boone 2016; Boone et al. 2014).

Dimensionality

A measurement instrument must tap into only one construct (i.e., be unidimensional) in order to generate meaningful measures of the amount of that construct a person has (AERA, 2014). Because the CANS was originally developed to cover five major concept categories associated with natural selection (evolution, inheritance, selection, variation, and mutation), Kalinowski et al. initially hypothesized that each of these categories could be distinct constructs (or dimensions), in which case, one would model them as separate unidimensional models. Conversely, it is possible that the entire instrument falls on a single dimension. Kalinowski et al.'s dimensionality analyses of the CANS were limited by low sample sizes and generated inconclusive results regarding whether a unidimensional or a multidimensional model was most appropriate. We used a Rasch framework and a much larger sample size to robustly test the dimensionality of the CANS.

Specifically, two Rasch-based approaches were used to test whether the CANS is best characterized as unidimensional or multidimensional (RQ 1.4). First, we ran a unidimensional Rasch model by treating all 24 items as a single dimension. The dimensionality of the response patterns was then assessed using a principal component analysis (PCA) of the residuals from the unidimensional Rasch model. Because the unidimensional Rasch model assumes that the data had only one dimension, the residuals are expected to have no structure because nearly all variation in the response data should be accounted for by the single modeled dimension. However, if the instrument-derived responses are multidimensional, the residuals from a unidimensional Rasch model are expected to be large and may possibly show a pattern of shared unexplained variance (i.e., shared variance not accounted for by a unidimensional Rasch model), which may indicate the presence of one or more additional and unintended constructs (Boone et al. 2014). If present, signals of patterns in these residuals would be apparent in the PCA's first contrast, with an eigenvalue greater than 2 suggesting potential multidimensionality (Boone et al. 2014). PCA of Rasch residuals is an unsupervised and atheoretical approach to test for signals of multidimensionality, which is advantageous for discovering unexpected dimensionality patterns within the instrument without a priori assumptions.

The second approach of testing for multidimensionality examines hypothesized dimensions using a statistical framework that compares the fit of a unidimensional and multidimensional Rasch model. The hypothesized multidimensional model is grounded in theoretical assumptions of the structure of the construct. The fit of the uni- and multidimensional models may then be compared to each other using likelihood ratio testing (Neumann et al. 2011; Robitzsch et al. 2024). For models that are not significantly different, the more parsimonious unidimensional model is favored. In the case where the null hypothesis is rejected, the models are compared using the Bayesian information criterion (BIC). The BIC parameter is calculated from the log-likelihood of the model and includes a penalty term for the number of parameters, favoring simpler models. The model with the lower BIC is evaluated as having a more optimal fit. We designed the multidimensional model to have five dimensions, one for each concept category in the CANS. Finally, the correlations between the person abilities for each of the five dimensions of this model were examined, with interpretations of correlation coefficient strength informed by Akoglu (2018). A strong correlation between all dimensions would suggest unidimensionality whereas a weak correlation among all dimensions would provide strong evidence of a five-dimensional structure.

Part 2: student knowledge and learning of natural selection

The lme4 package (Bates et al. 2024) was used to generate linear mixed effects models (also called hierarchical linear models) to examine patterns of incoming evolution knowledge and learning in association with students' first exposure to college level evolution instruction. In all models, pre-test person abilities were generated using a unidimensional Rasch model by anchoring the

post-test item difficulties to the pre-test Rasch model at model estimation. The person ability measures (in logits) generated from the unidimensional Rasch models were then included in the GLM model as a continuous outcome variable. Instructional time point (pre-test and post-test), semester, prior biology coursework, biological sex, and socially defined race were included as fixed effects. Respondents who were missing any of these variables were excluded from the final model, and PELL eligibility and college generation status were not included as this information was collected in only a subset of semesters. The pre-test person abilities were designated as the reference for the "time point" variable, "no prior biology" was set as the reference for the prior biology variable, "male" was set as the reference for the biological sex variable, and "White" was set as the reference for the socially defined race variable. Because this analysis involved a repeated measures design (two time points per respondent), student identification was added as a random effect (random-intercept only). The model was fitted using restricted maximum likelihood (REML), and the *p* values were calculated using t-tests with Satterthwaite's method (Satterthwaite 1941). Partial omega squared (^2_p) was used as a measure of effect size to measure the unique variance that each variable contributes to differences in Rasch transformed CANS measures (Lakens 2013). The magnitude of the contribution of instruction on evolution learning (RQ 2.1) can be measured by the effect size of the "time point" variable.

Interaction effects between time point and semester, as well as between time point and each student background variable, were added to examine whether there were disproportionate learning gains by semester (RQ2.2) and student background characteristics (RQ2.3). To determine which concepts were the most difficult for students, the mean Rasch item difficulty for each concept category was calculated and examined (RQ2.4).

Finally, the structure of student knowledge across evolutionary phenomenon (RQ 2.5) was evaluated by comparing the knowledge magnitude and dimensionality across phenomena with the following combinations of features: (1) plant trait gain, (2) plant trait loss, (3) animal trait gain, and (4) animal trait loss. These phenomena were chosen because items in the instrument were specifically designed to include trait polarity (trait gain/loss) and taxon (animal, plant). When the knowledge magnitude differs significantly across phenomena, reasoning can be considered fragmented (i.e., not coherent). If reasoning patterns for one phenomenon (e.g., plant trait gain) are not predictive of those for another (e.g., animal trait loss) across respondents, then the data are multidimensional, and reasoning can be considered *differently* fragmented among respondents. If appropriate items exist within an instrument, a multidimensional model in which each phenomenon is designated as a separate dimension can be estimated and compared to a unidimensional model using a likelihood ratio test.

Results

Part 1: adherence of the CANS to well-accepted criteria of robust measurement

RQ 1.1: do items that comprise the CANS display acceptable fit to model expectations?

Using a unidimensional model, 19 items had acceptable infit and outfit MNSQ fit statistics at the post test, and five items (i.e., 5, 6, 12, 19, and 20) had unacceptable outfit (the infit was within acceptable ranges). Items 6, 12 and 20 were underfitting the Rasch model with an outfit MNSQ above 1.3, whereas items 5 and 19 were overfitting the Rasch model with an outfit MNSQ below 0.7. Items 5 and 19 were also among the easier items at this time point (Table 2). Although all five items are considered unproductive for measurement, they were not degrading to measurement because their outfit MNSQ statistics were below 2 (Table 2). At the pre-test, only three items (i.e., 6, 12, and 20) were misfitting, with both outfit and infit MNSQ values above 1.3, suggesting model underfit. The outfit MNSQ measures for items 6 and 12 were above 2, indicating that they were degrading to measurement. These two items also had the highest item difficulties (Table 2) and the lowest proportion of respondents answering them correctly in the post-test (29% for item 6 and 37% for item 12) (Table S-3; Table S-4).

RQ 1.2: Does the CANS reliably order items by their difficulties and respondents by their abilities on the latent traits?

Using a unidimensional Rasch model, the EAP/PV item separation reliability was 0.80 for the pre-test and 0.81 for the post-test. The WLE reliabilities were 0.78 for both the pre- and post-tests. Therefore, all reliability statistics were above the 0.7 threshold at both time points, indicating that items and persons could be ordered consistently along the latent trait. These findings suggest acceptable reliability of measurement for the CANS.

RQ 1.3: to what extent does the CANS precisely measure the latent trait?

The Wright map for the unidimensional Rasch model shows that item difficulties were clustered around the middle of the distribution of respondent abilities (Fig. 1). Therefore, the item difficulties did not span the full distribution of respondent abilities. Students at the highest and lowest ability levels were not precisely measured by the CANS instrument.

Table 2 Item difficulties, and weighted (infit) and unweighted (outfit) MNSQ fit statistics of the unidimensional CANS model

Item	CANS Concept Category	Item difficulty	Outfit MNSQ (Post-test)	Infit MNSQ (Post-test)	Outfit MNSQ (Pre-test)	Infit MNSQ (Pre-test)
Item 1	Evolution	-0.70	0.86	0.91	0.88	0.91
Item 2	Selection	-0.21	1.09	1.06	1.13	1.11
Item 3	Inheritance	-0.13	1.04	1.03	0.98	0.98
Item 4	Variance	-0.20	1.23	1.15	1.26	1.21
Item 5	Mutation	-1.02	*0.61*	0.79	0.97	1.03
Item 6	Selection	1.84	*1.64*	1.25	**2.32**	*1.73*
Item 7	Evolution	0.92	1.06	1.04	0.72	0.77
Item 8	Inheritance	-0.71	0.75	0.85	0.85	0.90
Item 9	Evolution	-0.87	0.68	0.81	0.81	0.87
Item 10	Evolution	-0.92	0.86	0.92	0.77	0.83
Item 11	Mutation	0.23	0.83	0.86	0.70	0.75
Item 12	Selection	1.38	*1.43*	1.23	**2.37**	*1.89*
Item 13	Variance	-0.12	1.24	1.16	1.25	1.20
Item 14	Inheritance	0.23	1.06	1.05	1.01	1.00
Item 15	Evolution	-0.17	0.83	0.89	0.85	0.87
Item 16	Selection	-1.02	0.73	0.86	0.78	0.85
Item 17	Evolution	1.00	1.09	1.05	0.82	0.83
Item 18	Variance	-0.26	1.01	0.99	1.10	1.08
Item 19	Mutation	-0.64	*0.63*	0.76	0.88	0.92
Item 20	Selection	0.14	*1.37*	1.27	*1.55*	*1.42*
Item 21	Evolution	-0.14	0.86	0.91	0.87	0.89
Item 22	Inheritance	0.63	0.99	0.99	0.88	0.89
Item 23	Mutation	0.39	0.82	0.85	0.72	0.76
Item 24	Evolution	0.35	1.25	1.20	1.10	1.05

The *underlined* values reflect outfits that are outside of the acceptable fit range (0.7–1.3 mean squares). The **bolded** values reflect outfits that are considered degrading to measurement

RQ 1.4: is the structure of the CANS best characterized as unidimensional or multidimensional?

The residuals of the unidimensional Rasch model had an eigenvalue of the first contrast that was greater than 2 (2.34), indicating that there may be unexplained variation in the unidimensional model, possibly due to the unintended inclusion of additional constructs. We plotted the item difficulties against the first contrast of the Rasch residuals to examine clustering among items within the five major concept categories covered in the CANS (Fig. 2). There was little evidence of clustering of these Rasch residuals by the five concept categories that were hypothesized by Kalinowski et al. to represent individual sub constructs within a multidimensional model. A possible exception may be the categories of mutation and selection, which showed some evidence of clustering in the Rasch residuals.

A five-dimensional Rasch model was then constructed based on the five concept categories. However, we do not report the MNSQ fit statistics of this model due to inconsistencies each time the model was generated, which may be partially attributed to the low person and item reliabilities in many of the dimensions. Despite these inconsistencies, we report that when the fit of this model was compared to the fit of the unidimensional model using maximum likelihood testing, the five-dimensional model had a significantly better fit (lower BIC) than the unidimensional model at the post-test ($\chi2=2564.3$, df=14, $p<0.001$) but only some concept categories had acceptable item reliabilities (Table 3). A significant difference was not observed between the unidimensional and five-dimensional models for the pre-test data ($\chi2 = -31403.9$, df=22, $p>0.05$), suggesting that the more parsimonious one-dimensional model was a better fit at this time point (Table S1). For both time points, the correlations between the five dimensions ranged between strong and weak (Figure S1), mirroring the inconclusive dimensionality results reported throughout this manuscript.

We then examined the fit and reliability statistics of the five individual unidimensional models to assess whether this theoretically grounded model demonstrated acceptable measurement statistics. Similar to the five-dimensional model, we report that the weighted MNSQ fit statistics values for the five one dimensional models were inconsistent across model estimations; in some estimations, all items had acceptable fit, whereas items in the inheritance and variation concept categories demonstrated misfit in other iterations. Furthermore, the EAP/

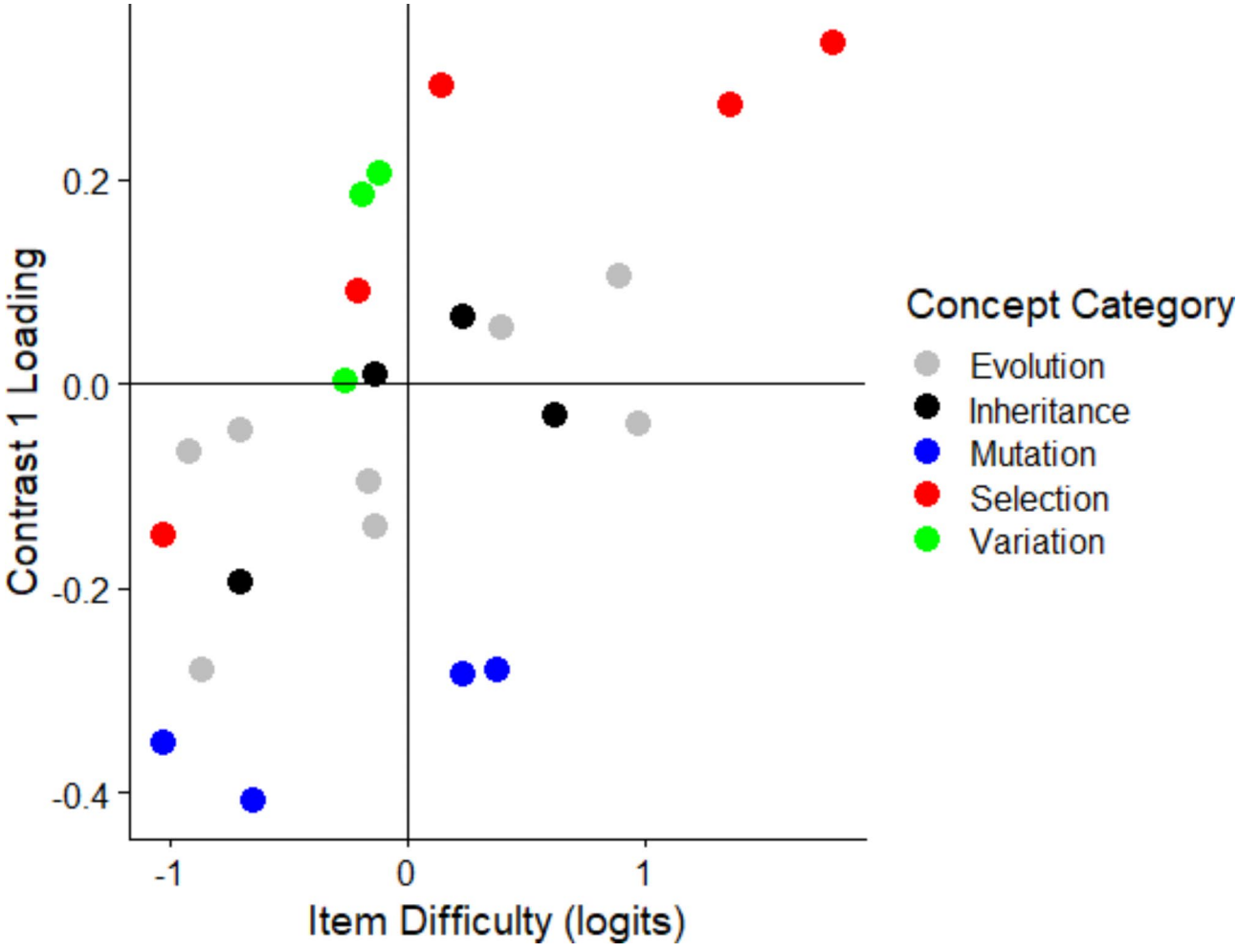

Fig. 2 First contrast of the PCA of Rasch residuals and item difficulties for the unidimensional Rasch model at the post-test model estimation. Items were classified into the five concept categories that the CANS was conceptualized to encompass. There is limited evidence of clustering of residuals across these categories, with the possible exception of shared structure among the items that address mutation and selection

Table 3 Item and person reliabilities for each concept category modeled as a separate unidimensional rasch model or a single five-dimensional model (post-test survey only)

		CANS Concept Category				
Type of Model	Reliability measure	Evolution	Mutation	Inheritance	Selection	Variance
Separate Uni-dimensional Models	*EAP/PV*	0.68	0.68	0.44	0.19	0.37
	WLE	0.51	0.26	0.04	0.00	0.00
Five Dimensional Model	*EAP/PV*	0.81	0.82	0.81	0.47	0.49
	WLE	0.53	0.19	0.07	0.02	0.00

PV item separation and WLE person separation reliabilities were consistently below the acceptable threshold (Table 3).

Part 2: evolution knowledge and learning

RQ 2.1 What are the magnitudes of CANS pre-test measures and learning gains across 14 semesters of a high-enrollment gateway biology course?

Students entered the course with low levels of evolution knowledge as measured by the CANS. Using raw composite CANS scores to allow for comparability with other studies, the median at pre-test was 10/24 correct responses ($\bar{x}$=10.9±4.9, ~45% correct). There was a substantial increase in the raw composite CANS scores from the pre-test to the post-test, with a median of 16/24 correct responses following instruction in the post-test ($\bar{x}$=15.3±4.9, ~64% correct). The Rasch-transformed person ability measures mirrored these findings, with a significant increase in mean logit measures from pre-test ($\bar{x}$= -0.2±1.04) to post-test ($\bar{x}$=0.72±1.12). A linear

Table 4 Summary of regression results for the CANS

	Variable	Differences
Pre-test Knowledge	Biological Sex	Male > Female and non-binary
	Race	White > Asian, Black/African American, Hispanic of any race
	Prior Biology Coursework	1 or more courses > 0 courses
	Semester	Semester 1 > Semester 7, Semester 8, Semester 10, Semester 11, Semester 13
Learning Gains	Instruction (pre to post)	Post > pre
	Biological Sex	n.s.
	Race	n.s.
	Prior Biology Coursework	0 courses > 1 or more courses
	Semester	Semester 1 > Semester 7, Semester 9, Semester 10 Semester 1 < Semester 6

Table 5 Partial omega squared (ωp^2) to measure effect size (with the following cutoffs: small = 0.01, medium = 0.06, large = 0.14) for instructional time point, semester, biological sex, prior biology coursework, and race

Variable	Partial omega squared (ωp^2)
Instructional Time point	0.46
Semester	< 0.01
Biological Sex	0.03
Prior Biology Coursework	0.04
Race	0.04

mixed effects model was conducted using the Rasch-transformed person ability measures (see Table S5 for a conversion guide that can be used on CANS datasets from other samples). The results showed significant and large gains in CANS measures when controlling for demographics, background variables, and semester, (β = 0.45, t = 19.01, $p < 0.001$) (Table 4; Table S6), paralleling the increase in raw scores. The effect size of instruction on CANS measures was large ($\omega p^2 = 0.46$) (Table 5). These results collectively indicate that students had relatively low knowledge of core concepts related to natural selection upon course entry, but experienced significant and meaningful gains in natural selection knowledge following instruction.

In alignment with the general finding that CANS measures increased on the post-test, most items were less difficult for students at the post-test and did indeed show a higher proportion of correct responses following instruction (Table S-3; Table S-4). However, there are notable exceptions that could indicate problems with the items as well as issues related to instruction. Specifically, item 6 had only a marginal increase in the proportion of correct responses and maintained a high item difficulty measure on the post-test, and item 12 had a slight decrease in the proportion of correct responses and an increase in item difficulty on the post-test. Both items also had poor fit on the pre-test. Item 20 also experienced an increase in the item difficulty from pre- to post-test and item 13 experienced almost no change in item difficulty over this time period.

RQ2.2 How variable are CANS measures across semesters?

Both pre-test knowledge of natural selection and post-test learning gains differed across semesters, but the semester explained little of the variance in the Rasch-transformed CANS measures ($\omega p^2 < 0.01$) (Tables 4 and 5; Table S6). In comparison to the earliest semester in which the data were gathered (Semester 1), three semesters displayed significantly lower pre-test knowledge of natural selection at the 0.01 significance level: Semester 7 (β = -0.03), Semester 9 (β = -0.06), and Semester 10 (β = -0.03). Learning gains also varied from the inaugural semester, with significantly lower learning gains in Semester 7 (β = -0.03), Semester 9 (β = -0.06), and Semester 10 (β = -0.03), and significantly higher learning gains in Semester 6 (β = 0.03). These patterns in pre-course knowledge and post-test learning gains are shown in Fig. 3 and Figure S2 and demonstrate that there is no consistent pattern in knowledge or learning that is linked to temporal progression through the 14 semester study period.

RQ 2.3. How variable are CANS measures across different student background characteristics?

Student background and demographic characteristics explained a significant amount of variance in the Rasch transformed CANS measures. Controlling for these variables, male identifying students had significantly higher knowledge of natural selection at pre-test than female and non-binary identifying students (β = -0.12, t = -10.66, $p < 0.001$) (Table 4; Table S6). The unique variance explained by biological sex was small ($\omega p^2 = 0.03$) (Table 5). The interaction effect between biological sex and instructional time point was nonsignificant, indicating that students had similar evolution learning gains regardless of biological sex, although this means that the disparities present at the pre-test remained.

At course entry, students who identified as White had significantly higher knowledge of natural selection in the pre-test than students who identified as Asian (β = -0.14, t=-11.14, $p < 0.001$), Black or African American (β = -0.13, t = -10.759, $p < 0.001$), or Hispanic of any race (β = -0.12, t = -9.59, $p < 0.001$). The unique variance explained by race was small ($\omega p^2 = 0.04$), but had a larger effect than biological sex. Similar to biological sex, students of all socially defined races had similar learning gains from pre- to post-test as indicated by the insignificant interaction effects between each racial group and time point.

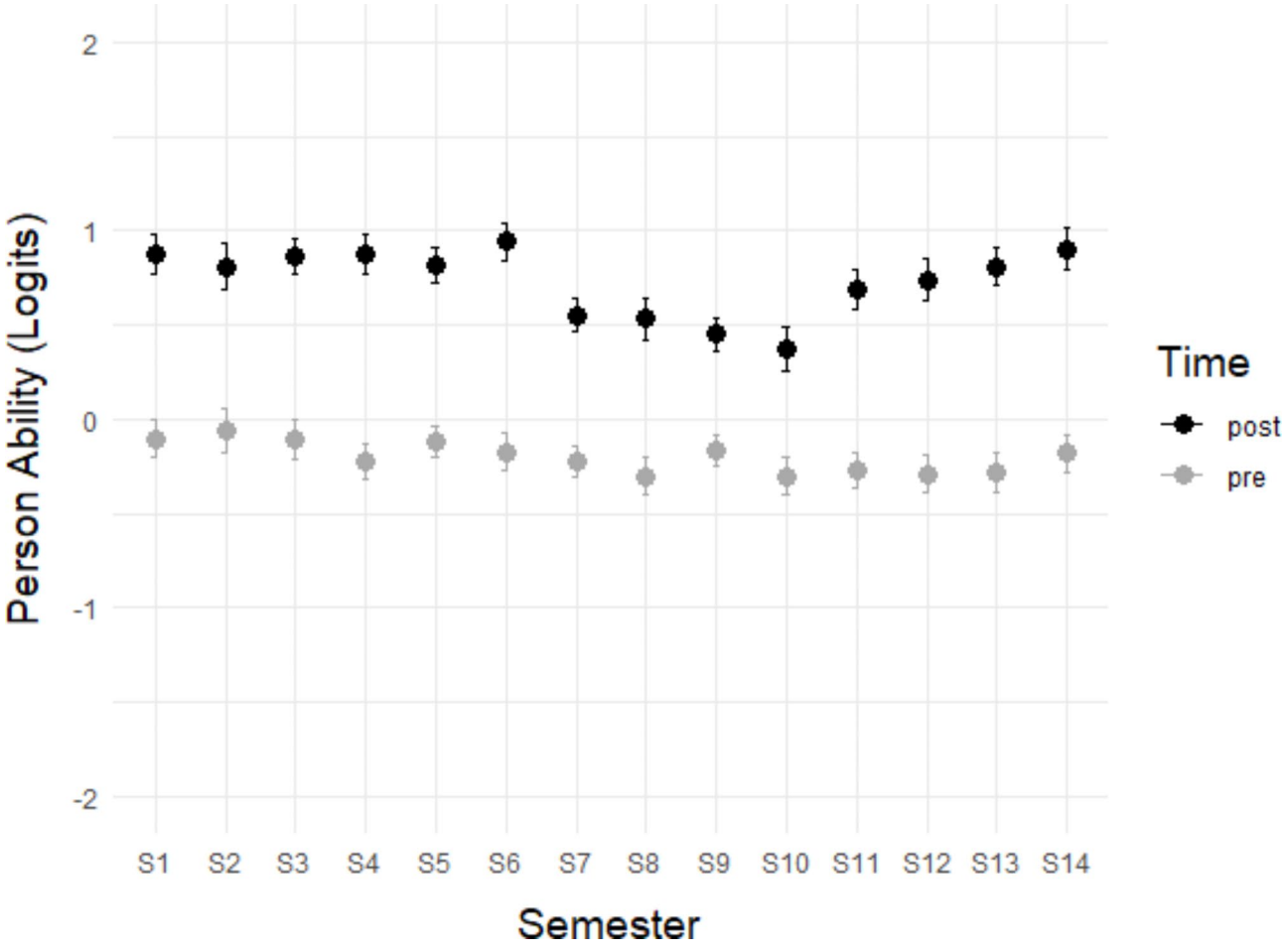

Fig. 3 Mean Rasch person ability measures for the CANS at the pre- and post-test, disaggregated by semester

Students who entered the course after completing one or more prior biology courses had significantly higher knowledge of natural selection ($\beta = 0.18$, $t = 16.11$, $p < 0.001$) than students without prior coursework. The unique variance explained by prior coursework was small ($\omega p^2 = 0.04$). In contrast to biological sex and socially defined race, there was a significant interaction effect between prior biology coursework and instructional time point ($\beta = -0.06$, $t = -5.56$, $p < 0.001$), with students that had one or more prior biology courses experiencing less learning gains than students without prior coursework. In other words, while students who entered the course without prior coursework had lower knowledge of natural selection, they experienced disproportionately higher learning gains that mitigated this knowledge debt by the end of the course.

RQ 2.4 which evolution topics within the CANS are most difficult for students?

Of the five concept categories, items that focused on mutation were generally easiest for students (mean item difficulty=-0.26), as evidenced by the high proportion of students answering correctly in the post-test and the low item difficulties (Fig. 2). The items that were most difficult for students were in the evolution (mean item difficulty=-0.07) and selection (mean item difficulty=0.43) concept categories. Specifically, the most difficult items within these categories (with less than 50% of students answering correctly at the post-test) probed for student understanding of competition in an ideal environment (item 6), trait loss (item 7), exponential growth using graphical intuition (item 12), and the role of individuals' responses to the environment in evolutionary change (item 17).

RQ 2.5: What is the structure (i.e., coherent vs. fragmented) of student evolutionary knowledge across phenomena?

It was not possible to meaningfully assess the structure of knowledge using the CANS because the items representing the four phenomena –plant trait gain, plant trait loss, animal trait gain, animal trait loss– were unbalanced (Table 6). Specifically, trait polarity and taxon do not appear equally in all five concept categories that the CANS authors conceptualized to define the construct

Table 6 Distribution of items across taxa for two theoretical constructs related to the structure of students' evolutionary knowledge

	Plant	Animal
Concept Category		
Evolution	✓ (2 items)	✓ (6 items)
Mutation	✓ (1 item)	✓ (3 items)
Inheritance	**x (0 items)**	✓ (4 items)
Selection	✓ (2 items)	✓ (3 items)
Variation	✓ (1 item)	✓ (2 items)
Trait Polarity		
Trait Gain	✓ (1 item)	✓ (2 items)
Trait Loss	**x (0 items)**	✓ (2 items)

Anteater, bowhead whale, and mosquito were classified into 'animal' whereas saguaro cacti was classified under 'plant.' Bold text indicates there were no items for the construct within the taxon

of natural selection knowledge; there is only 1 item that addresses mutation in plants and no items that address inheritance in plants (Table 1). There are also no items that address trait loss in plants, meaning that no items could be included in the plant trait loss dimension. Overall, plants are represented in substantially fewer items than animals and trait loss is represented in substantially fewer items than trait gain. As a result, it was not possible to test for dimensionality in alignment with these features.

Discussion

Our study sought to expand the validity evidence for the CANS by investigating whether the instrument productively measured the intended construct using a more stringent 1-parameter IRT model and a new and much larger (>6000) sample of undergraduate students that spans 14 semesters of a gateway biology course. Unlike the original publication, our sample allowed rigorous tests of multidimensionality, replication of findings through time, and ample statistical power for examining knowledge and learning of natural selection among students with different backgrounds and demographic characteristics.

Measurement of Natural Selection understanding using the CANS

Our psychometric analyses assessed dimensionality, item fit, reliability, and person-item alignment of the CANS. Overall, our results showed a similar, but more complete picture about the psychometric properties of the CANS as compared to the original study. In the original study, Kalinowski et al. (2016) reported high person separation reliability and unidimensional factor loading for a majority (18/24) of the items. However, they reported that six items had poor factor loading in a unidimensional IRT analysis (items 4, 6, 12, 13, 18, and 20). Similarly, in the present study, a one-dimensional structure generated sufficient reliabilities and acceptable item fit for most (i.e., 19/24) items, but several of the same items identified by Kalinowski et al. were found to have problematic fit statistics in this study as well. Specifically, like Kalinowski et al., items 6, 12, and 20 were identified as misfitting (additionally, items 5 and 19 were also identified as misfitting in our study). However, despite the high reliabilities and generally strong item fit, our analysis of Rasch residuals of the 1-dimensional model generated an eigenvalue of the first contrast that was >2, suggesting the presence of additional, possibly unintended dimensions or constructs. Kalinowski et al.'s finding that six items in two content categories had low loadings on a unidimensional IRT analysis were interpreted by the authors to suggest a similar conclusion about the presence of multiple dimensions.

The finding that there may be multiple dimensions in the CANS is not particularly surprising and was even hypothesized a priori by Kalinowski et al. as one possible structure for this instrument. In particular, the five concept categories were each expected to measure a distinct topic, suggesting the possibility of a five dimensional structure for the CANS. To test this hypothesized structure, Kalinowski et al. fit five separate unidimensional IRT models (using a 2 or 3-parameter IRT model) and their results showed that two of the previously low loading items (12 and 20) in the *selection* category became high loading. The other four items (4,6, 13, and 18) remained low loading. Therefore, Kalinowski et al.'s analyses suggest that the five-dimensional modeling approach may have improved the fit of the response data to some extent.

Our much larger sample size allowed us to conduct a more robust multidimensionality test. We report that the multidimensional model had significantly better fit to the response data than the one dimensional model for the post-test (but not the pre-test). However, the reliabilities for the *variation* and *selection* concept categories were also very low and the items had inconsistent fit indices (Table 7). The low reliabilities could be due to an insufficient number of items within these categories and to possible problems with the items themselves, which we discuss more in the next section. Overall, there were dimensionality problems reported for both the one-dimensional and five-dimensional model in this study, indicating that dimensionality weaknesses of the CANS reported by the instrument's authors are not sufficiently addressed by large sample sizes or more robust dimensionality analyses.

We also analyzed the precision with which the items measured the target population by comparing the alignment of the item difficulties to the person abilities. Rasch item difficulties and person abilities plotted on the Wright map showed that the items did not span the full

Table 7 Recommended improvements for the CANS

Categories	Finding	Recommendation
Item fit	Items 6, 12, and 20 were misfitting on the pre or the post-test survey	Modify or remove these items as described in Table 8.
Item difficulty	Some content categories (e.g., variation, mutation) have too little variation in their item difficulties and do not match the abilities of the respondents.	Write new items that broaden the item difficulty.
Item redundancy	Items 4, 13, and 18 are all about the same topic (variation) and have similar item difficulties.	Modify these items to vary in difficulty.
Content validity	Some items (e.g., items 6, 12, and 20) may be adding construct irrelevant variation.	Remove or modify these items.
Pre-post learning patterns	Items 13 and 20 did not change following instruction despite explicit attention to this topic during the course.	Modify items; see specific item recommendations in Table 8.
Reliability	Low reliability in the variation and selection content categories when modeled as five dimensions.	Remove or modify misfitting items and add more items.
Item representation and balance	The features of trait polarity and taxon are not balanced among the items and poorly operationalized into phenomena	Add items about trait gain, trait loss, inheritance selection, mutation, variation, and evolution in plants to balance it with the animal items.

range of abilities for the undergraduate students in this sample, which suggests relatively low measurement precision for students with the lowest and highest ability levels. This pattern is especially pronounced for three of the concept categories –*variation*, *mutation*, and *inheritance*. The items in the variation category align with a very narrow range of student abilities, and the mutation and inheritance categories are in need of more difficult items. Future revisions to the CANS should modify existing items or add new items to encompass a broader range of difficulties for these categories, and possibly remove items that target the same topic yet have similar item difficulties.

Overall, both Kalinowski et al. (2016) and this study showed evidence that neither a one-dimensional nor five dimensional model was clearly the most appropriate structure for the CANS, and our much larger sample size and more robust dimensionality analysis did not resolve this issue. Furthermore, the instrument may have low precision of measurement for some students, especially for certain topics. Finally, multiple items displayed problematic fit statistics, some of which were also identified as poor fitting by Kalinowski et al. Below we detail the problematic items (Tables 8 and 7) and propose possible explanations and solutions for their poor fit.

Potentially problematic items in the CANS

Several CANS items were identified as misfitting in this study, some of which may be candidates for modification or removal (Tables 8 and 7). Specifically, items 6, 12, and 20 were found to underfit the Rasch model with MNSQ values above 1.3, suggesting guessing and careless mistakes (Bond and Fox 2007). Although item 20 was unproductive for measurement, items 6 and 12 were found to be more problematic because their outfit MNSQ outfit value was above 2, indicating they were likely degrading to the measurement model at the pre-test (see Table S2 for the text of these problematic items). Kalinowski et al. identified these same three items as problematic. In addition, items 12 and 20 were also resistant to instruction in our sample as evidenced by increasing item difficulties from pre to post test (Table S4).

Items 5 and 19 were also misfitting in our study but had MNSQ values at the other end of the acceptable fit range (i.e., < 0.7), indicating a different concern. Specifically, low MNSQ values indicate model overfit and that the responses are overly predictable and redundant (Bond and Fox 2007). In fact, items 5 and 19 assess the same concept and misconceptions using the same item format but differ in the taxon specified. Specifically, item 5 contextualizes natural selection within an animal and item 19 conceptualizes within a plant (See Table S2 for the item text). Parallel items like these can be extremely useful for understanding the role of phenomena in student evolutionary reasoning. Therefore, because their fits were acceptable at the pre-test and not degrading to measurement at post test and because their overfit is likely due to the parallel nature of the items, we do not recommend removing or modifying these items. We therefore focus our analysis of problematic items below on the misfitting items 6, 12, and 20 as well as on one other item (Item 13) that fit the Rasch model but was found to be effectively resistant to instruction despite specific emphasis on the topic during instruction (Table S4).

Item 6 is from the selection content category was found to be misfitting on both the pre- and post-test (pre outfit: 2.17, post outfit: 1.60). This item prompts students to evaluate descriptions of what life is probably like for anteaters who live in a reserve with a stable population size (see Table S2 for the text of the item). The item seeks to draw out reasoning about competition, which has been argued to be a relevant concept in evolutionary reasoning, but not necessarily a core concept (Nehm and Ridgway 2011; Opfer et al. 2012). In other words, normative evolutionary reasoning *could* include the concept of competition, but absolutely requires only three concepts: variation, heredity, and selection. Therefore, the inclusion of competition within a measure of natural selection understanding may introduce construct-irrelevant variation in the response data. In fact, the gateway biology

Table 8 Summary of problematic items as indicated by the current study's analysis and Kalinowski et al. (2016)'s IRT analysis

Item	Psychometric Pattern				Recommended Improvements
	IRT Factor Loading[1]	Unidimensional Rasch Model Fit [2]	Resistance to Instruction[2]	Construct Irrelevant Variation[2]	
4	*Poor factor loading*	Acceptable	*Not resistant. Item difficulty decreased.*	Absent	No specific changes recommended by Kalinowski et al. and our study did not find problems with this item. The item requires reading of an introductory paragraph at the beginning of the section for full context. It may be beneficial to provide this information directly in the question itself.
5	*Acceptable factor loading*	*Unacceptable. Unproductive for measurement likely due to redundancy with item 19.*	*Not resistant. Item difficulty decreased.*	Absent	No improvements recommended
6	*Poor factor loading*	*Unacceptable. Degrading to measurement*	*Not resistant. Item difficulty decreased.*	*Present*	Address possible construct irrelevant variation related to the nature of science or remove the item from the instrument.
12	*Poor factor loading*	*Unacceptable. Degrading to measurement*	*Resistant. Item difficulty increased but topic not addressed*[3]	*Present*	Address possible construct irrelevant variation related to the nature of science or remove the item from the instrument.
13	*Poor factor loading*	Acceptable	*Resistant. Item difficulty remained almost constant.*	Absent	Change text in answer option C from "different genes" to "different alleles" or "different versions" of the same gene.
18	*Poor factor loading*	Acceptable	*Not resistant. Item difficulty decreased.*	Absent	No specific changes recommended by Kalinowski et al. and our study did not find problems with this item.
19	*Acceptable factor loading*	*Unacceptable. Unproductive for measurement likely due to redundancy with item 5.*	*Not resistant. Item difficulty decreased.*	Absent	No improvements recommended
20	*Poor factor loading*	*Unacceptable. Unproductive for measurement.*	*Resistant. Item difficulty increased.*	Absent	Define seedlings in the introductory paragraph and the question stem.

[1]Finding from Kalinowski et al. 2016

[2]Finding from the current study

course in which this study took place did not explicitly address the role of competition in evolutionary change, which may explain why students in this sample showed almost no knowledge gains on this topic at post-test (Table S3).

Another possible reason for this item's misfit is that students may be using knowledge about the nature of science, not just about evolution, in their reasoning. Specifically, the item prompt asks students to evaluate possible *descriptions* of what life is like for these anteaters and one of the answer options is "It is impossible to know without actually observing anteaters in the reserve", which was selected by 36% of students at the pre-test. In the Nature of Science literature, *descriptions* are inextricably connected to the concept of *observations* (Lederman and Abd-El-Khalick 2002; Lederman et al. 2002). Asserting that no *description* can be made because of a lack of observational evidence is arguably legitimate reasoning that aligns with fundamental principles of the nature of science. Items that tap students' nature of science knowledge may also be generating construct-irrelevant variation.

Item 12 is also from the selection content category and was found to be misfitting on the pre-test (outfit: 2.22). This item prompted students to predict what the population growth of bowhead whales would be expected to look like in graphical form (see Table S2 for the full item). The question seeks to draw out reasoning about exponential growth using graphical representations of population size change. Because graphical reasoning, to our knowledge, is not considered part of the construct of natural selection understanding, this item may also be adding construct-irrelevant variation. In addition, the item's focus on the topic of exponential growth may also be inserting construct-irrelevant variation. Although all populations have the potential for exponential growth, such a growth pattern (and indeed any growth at all) is not a necessary feature of natural selection. Again, the gateway biology course in which this study took place did not explicitly address the role of exponential growth in evolutionary change, which may explain why students did not show knowledge gains on this topic at post-test.

Item 13 had no fit issues at either the pre- or post-test, but, though technically not resistant to instruction, the item difficulty decreased by only 0.02 logits despite explicit instruction on the topic it aims to address. This item prompted students to identify the cause of swimming speed variation in bowhead whales, which was

intended to assess students' understanding of the role of *both* the environment and genes in generating phenotypic variation (see Table S2 for the full item). Notably, the phenotype of focus in this item is a behavioral trait (i.e., swimming speed), which may invoke different evolutionary reasoning as compared to a physical trait (e.g., fin length) (Nehm and Ridgway, 2011). The distribution of answer options selected by respondents (Table S4) did not meaningfully change from pre- to post-test despite explicit instruction about the role of both the environment and genes in influencing behavioral and physical phenotypes; Approximately two-thirds of students selected the answer option that was designated as correct by Kalinowski et al. at both the pre- and post-test. This "correct" answer specified that the variation arising from both environmental factors (i.e. nutrition and exercise, option B) *and* genetic factors (i.e., genes, option C) contributed to phenotypic variation in a whale species (option D, encompassing both Options B and C, was therefore designated as correct). However, the text of answer option C, "there will be notable differences among the whales because each whale has *different genes*", is inconsistent with normative genetic reasoning. Individuals of the same species do not have different genes, rather they have different alleles (i.e., versions) of the same genes. Therefore, option D is technically incorrect as written. We recommend modifying the item to state that differences among the bowhead whales arise because of different *alleles* or *versions* of the same gene. The distinction between genes and alleles was explicitly addressed in the gateway biology course in which this study took place, which may explain why the item was resistant to instruction in this population.

Item 20 had a fit index that was deemed unproductive for measurement at the pre-test (outfit:1.55). This item prompts students to reason about the role of chance in evolution, specifically about the extent to which chance plays a role in seedling production. This question assumes that students know what a seed and a seedling is, yet many studies have demonstrated that K-12 students, college students, and pre-service teachers hold many misconceptions about plants, including topics related to seeds, seedlings, and fruit (Wynn et al. 2017; Yangin et al. 2014; See Wynn et al. 2017 for a review). Therefore, again, this item may be tapping into a construct (e.g., plant biology knowledge) other than the one intended by the authors and thus may be generating construct-irrelevant variation. The fit of this item may have been within acceptable levels at the post-test because students were introduced to plant biology topics during instruction. To improve fit of this item at the pre-test, we recommend adding information about seedlings to the introductory text and to the item itself.

Advancing the measurement of natural selection understanding

The CANS was developed to improve upon the weaknesses of previous instruments that aim to measure knowledge of natural selection. In particular, the authors sought to include more misconceptions in the distractors, more item forms, and multiple evolutionary contextual features (e.g., trait gain vs. loss, animal vs. a plant). However, there are some opportunities for improvement within the CANS that go beyond the item-level modifications recommended above.

First, the authors do not provide a rationale for their claim that multiple question forms are needed to accurately assess student understanding of natural selection, nor is it evident that the CANS instrument contains the necessary item forms to achieve whatever benefit the authors may have intended. There certainly are possible rationales for why different item forms may be desirable. For example, closed response items ask students to retrieve knowledge whereas open response items ask students to construct knowledge (Tofade et al. 2013). However, the CANS is composed of only closed response items with misconception distractors. Therefore, if multiple item forms are indeed important to accurately assess student understanding of natural selection, more work is needed to articulate whether the CANS achieves this goal and if not, what kinds of modifications to the instrument might be needed.

Second, the inclusion of various contextual features in the CANS instrument is an important advance in the measurement of evolution understanding because it aligns with student reasoning about evolution. In particular, the CANS instrument includes two features –trait polarity (trait gain vs. loss) and taxon (animal vs. plant)-- both of which have been found to impact students' evolutionary reasoning (Nehm and Ha 2011). However, although the authors' selection of these features appropriately reflects their importance for tapping into student thinking, the way in which these features were operationalized into phenomena within the CANS does not. In particular, there are no or too few items that capture each of the four phenomena – plant trait gain, plant trait loss, animal trait gain, animal trait loss– that can emerge from these features, suggesting that the CANS may underrepresent the construct. For a tool to facilitate accurate inferences about student thinking, it must intentionally tap into that thinking (e.g. NRC, 2001; AERA et al., 2014), but the lack of appropriate and balanced representation of phenomena within the items suggests that the CANS does not appear to effectively achieve this standard. This limitation of the CANS is problematic because a necessary step toward improving evolution education is the implementation of robust assessment tools that can effectively measure the progression from the novice-like

(i.e., fragmented) to expert-like (i.e., coherent) knowledge structures (Ziadie and Andrews 2018). Unfortunately, the unbalanced presentation of phenomena within the CANS mirrors a broader weakness in how phenomena are considered in biology education more generally (e.g., in textbooks, Abreu and Nehm 2024). Future revisions of the CANS should aim to achieve more appropriate balance and representation of phenomena among items (Table 7).

Patterns of evolution knowledge and learning: implications for evolution instruction

Although natural selection is a required topic in K-12 science curricula (NGSS Lead States 2013), the results of this and other studies (e.g. Andrews et al. 2011; Beggrow and Sbeglia 2019; Bishop and Anderson 1990; de Lima and Long 2023; Gregory 2009; Harding et al. 2021; Nehm and Reilly 2007) show a notable lack of proficiency in this topic among undergraduate students. We report that students had low pre-test CANS scores in our population (average score ~ 45%), which was similar to Kalinowski et al.'s original study (average score ~ 47%). Both populations showed large learning gains from pre- to post-test (our study: ~64% average at post-test; Kalinowski et al.: ~71% average at post-test). We also reported variation in pre-test knowledge and post-test learning gains by semester, with some semesters showing higher pre-test measures and higher learning than others. However, these patterns do not show evidence of incrementally improved learning gains through time, which is not particularly surprising due to changes in course modality and implementation amidst the COVID-19 pandemic that occurred in the middle of the study's sampling period.

Despite the overall large learning gains, some concept categories appear to be more challenging than others. In particular, Rasch item difficulties indicated that the mutation items were generally the easiest on average and the evolution and selection items were generally the hardest. The evolution category, by its nature, requires students to integrate multiple evolutionary concepts such as variation, heredity, and selection, which would be expected to be challenging. This finding supports the importance of designing curricula that encourages students to practice integrating these evolutionary ideas. For example, concept mapping has been widely used in evolution education to help students recognize that concepts do not exist in isolation (Okebukola 1990).

Finally, there has been a growing suite of literature documenting the role of student variables in STEM learning and career outcomes, which highlights the systemic inequities experienced by students from historically marginalized groups (e.g. Asai 2020; Chang et al. 2014; Estrada et al. 2016; Sbeglia and Nehm 2024; Whitcomb and Singh 2021). In this current study, we used Rasch measures to show disparities in evolutionary knowledge by socially defined race and biological sex at course entry, but these disparities did not impact the rate of evolution learning; students from all groups showed comparable learning gains throughout the semester. These findings underscore that disparities in knowledge did not increase through time, but rather were maintained with high magnitudes of overall learning. As a necessary consequence of maintaining disparities, students who began the course with more evolution knowledge (in this case, male and White students) continued to have the highest Rasch measures on the CANS at post-test. These results mirror the patterns reported in Sbeglia and Nehm (2024), but with an expanded data set (14 semesters instead of six) and an IRT measurement approach (as opposed to Classical Test Theory). However, although the mitigation of disparities should be the ultimate goal of instructional reform, the significance of the pattern reported here –maintenance of disparities with a high degree of learning– should not be underappreciated. Research suggests that many gateway college courses generate low levels of learning and may actually exacerbate disparities as evidenced by the finding that students from historically excluded communities are disproportionately "weeded out" of STEM degree pathways right around the time they are taking introductory courses (e.g., Hatfield et al. 2022; Nissen et al. 2021; Riegle-Crumb et al. 2019).

Conclusion

Educators require robust measurement instruments to assess student knowledge so that they know where students begin when they enter courses and how far instruction helps them advance. Unfortunately, such longitudinal learning data are rarely gathered, with most institutions instead focusing on static measures such as exam scores and course grades (e.g., Denaro et al. 2022), which provides insufficient information about teaching efficacy. As one of the few robust instruments available to measure the construct of natural selection understanding, the CANS holds great potential to provide some of the data needed to generate critical insights about student evolution knowledge, learning challenges, and progress, as well as information about which instructional approaches work best and are able to mitigate the notable knowledge disparities among students. The findings of this study offer insights into student evolutionary reasoning as well as tangible ways in which this instrument may be improved. In particular, we provide robust evidence that the dimensionality weaknesses of the CANS reported by the instrument's authors are not sufficiently addressed by large sample sizes or more robust dimensionality analyses. Rather, there are clear problems with several of the instrument's items, some of which may be tapping into additional and unintended constructs. There also seems to be limitations in the operationalization of

the five evolution content categories, which have signatures of functioning as distinct dimensions but lack a sufficient number of items that target a broad enough ability range. Finally, although our analysis of the instrument offered insights about student evolution learning challenges (e.g., that some content categories were more difficult for students than others), the unbalanced presentation of phenomena among the items is a substantial limitation in our ability to better understand the role of phenomena in student reasoning. Overall, the modifications recommended here could improve the effectiveness of the CANS as a tool for assessing students' conceptual understandings of evolution and for helping instructors monitor student learning of this core disciplinary idea.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12052-024-00210-3.

Supplementary Material 1

Acknowledgements
We thank Professor John True and Ross Nehm for their support and commitment throughout this project. We also thank Ross Nehm for his comments and suggestions for improving this manuscript. Finally, we thank two anonymous reviewers for their feedback on improving this manuscript.

Author contributions
GCS and ALZ both contributed to the conceptualization of the study and the writing and final approval of the manuscript. GCS collected and organized the data. ALZ performed all of the psychometric analyses, produced the figures, and wrote the methods and results.

Funding
Support for data collection was provided by National Science Foundation grant no. TUES-1322872, and analysis 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.

Data availability
Available upon reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
The study was approved by the university's institutional review board (protocol no. 504271) 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.

Competing interests
The authors declare no competing interests.

Received: 6 July 2024 / Accepted: 22 September 2024
Published online: 26 October 2024

References

AAAS. Vision and change in undergraduate biology education: A call to action. Washington, D.C; 2011.
Abraham JK, Meir E, Perry J, Herron JC, Maruca S, Stal D. Addressing undergraduate student misconceptions about natural selection with an interactive simulated laboratory. Evo Edu Outreach. 2009;2(3):393–404.
Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018;18(3):91–3.
American Educational Research Association, editor. Report and recommendations for the reauthorization of the institute of education sciences. Washington, D.C: American Educational Research Association; 2014. p. 60.
Anderson DL, Fisher KM, Norman GJ. Development and evaluation of the conceptual inventory of natural selection. J Res Sci Teach. 2002;39(10):952–78.
Andrews TM, Kalinowski ST, Leonard MJ. Are humans evolving? A classroom discussion to change student misconceptions regarding natural selection. Evo Edu Outreach. 2011;4(3):456–66.
Asai DJ. Race matters. Cell. 2020;181(4):754–7.
Bates D, Maechler M, Bolker B, Walker S, Christensen RHB et al. lme4: Linear Mixed-Effects Models using Eigen and S4. 2024. https://cran.r-project.org/web/packages/lme4/index.html
Beggrow EP, Sbeglia GC. Do disciplinary contexts impact the learning of evolution? Assessing knowledge and misconceptions in anthropology and biology students. Evo Edu Outreach. 2019;12(1):1.
Bishop BA, Anderson CW. Student conceptions of natural selection and its role in evolution. J Res Sci Teach. 1990;27(5):415–27.
Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Lawrence Erlbaum Associates. 2007.
Boone WJ. Rasch analysis for instrument development: why, when, and how? LSE. 2016;15(4):rm4.
Boone W, Staver J, Yale M. Rasch analysis in the human sciences. Dordrecht: Springer; 2014.
Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev. 2003;110(2):203–19.
Campbell CE, Nehm RH. A critical analysis of assessment quality in genomics and bioinformatics education research. CBE—Life Sci Educ. 2013;12(3):530–41.
Chang MJ, Sharkness J, Hurtado S, Newman CB. What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups. J Res Sci Teach. 2014;51(5):555–80.
Colton J, Sbeglia G, Finch S, Nehm RH. A quasi-experimental study of short- and long-term learning of evolution in misconception-focused classes. Paper presented at: American Educational Research Association (AERA) International Conference; 2018; New York, NY.
Colton J, Sbeglia GC, Finch S, Nehm RH. Differential impacts of active-learning approaches on demographic groups: Implications for efficacy studies in biology education. Paper presented at: CELT Teaching and Learning Colloquium; 2019; Stony Brook, NY.
de Ayala RJ. Item response theory and Rasch modeling. In: Hancock GR, Stapleton LM, Mueller RO, editors. The reviewer's guide to quantitative methods in the social sciences. 2nd ed. New York: Routledge/Taylor & Francis Group; 2019. pp. 145–63.
de Lima J, Long TM. Students explain evolution by natural selection differently for humans versus nonhuman animals. CBE—Life Sci Educ. 2023;22(4):ar48.
Demastes SS, Settlage J Jr., Good R. Students' conceptions of natural selection and its role in evolution: cases of replication and comparison. J Res Sci Teach. 1995;32(5):535–50.
Denaro K, Dennin K, Dennin M, Sato B. Identifying systemic inequity in higher education and opportunities for improvement. PLoS ONE. 2022;17(4):e0264059.
Estrada M, Burnett M, Campbell AG, Campbell PB, Denetclaw WF, Gutiérrez CG, et al. Improving underrepresented minority student persistence in STEM. CBE—Life Sci Educ. 2016;15(3):es5.
Gregory TR. Understanding natural selection: essential concepts and common misconceptions. Evo Edu Outreach. 2009;2(2):156–75.
Hambleton RK, Jones RW. An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. Educ Measure Issues Pract. 1993;12(3):38–47.
Harding RLS, Williams KR, Forcino FL, Dees J, Pennaz M, Momsen JL. What do students know about evolution by natural selection after a non-majors geology course? An analysis of student responses to open-ended questions. J Geosci Educ. 2021;69(3):253–64.
Hatfield N, Brown N, Topaz CM. Do introductory courses disproportionately drive minoritized students out of STEM pathways? PNAS nexus. 2022;1(4):pgac167.
Kalinowski ST, Leonard MJ, Taper ML. Development and validation of the conceptual Assessment of Natural selection (CANS). CBE—Life Sci Educ. 2016;15(4):ar64.

Kampourakis K, Zogza V. Students' intuitive explanations of the causes of homologies and adaptations. Sci Educ. 2008;17:27–47.
Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4:863.
Lederman N, Abd-El-Khalick F. Avoiding de-natured science: activities that promote understandings of the nature of science. In: McComas WF, editor. The nature of science in science education: rationales and strategies. Dordrecht: Springer Netherlands; 2002. pp. 83–126.
Lederman NG, Abd-El-Khalick F, Bell RL, Schwartz RS. Views of nature of science questionnaire: toward valid and meaningful assessment of learners' conceptions of nature of science. J Res Sci Teach. 2002;39(6):497–521.
Linacre M, Wright B. Constructing linear measures from counts of qualitative observations. In: Fourth International Conference on Bibliometrics. Berlin; 1993.
McCormick AC, Zhao CM. Rethinking and reframing the Carnegie classification. Change. 2005;37(5):51–7.
Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741–9.
National Research Council. Knowing what students know. Washington, D.C.: National Academies Press; 2001.
Nehm RH, Abreu E. How do biology textbooks align evolutionary phenomena with causal-mechanistic explanations? Paper presented at the fourth international workshop in genetics and evolution education. 2024; Karlstad University, Sweden.
Nehm RH, Ha M. Item feature effects in evolution assessment. J Res Sci Teach. 2011;48(3):237–56.
Nehm RH, Mead LS. Evolution assessment: introduction to the special issue. Evo Edu Outreach. 2019;12(7):1–5.
Nehm RH, Reilly L. Biology majors' knowledge and misconceptions of natural selection. Bioscience. 2007;57(3):263–72.
Nehm RH, Ridgway J. What do experts and novices "see" in evolutionary problems? Evo Edu Outreach. 2011;4:666–79.
Nehm RH, Schonfeld IS. Does increasing biology teacher knowledge of evolution and the nature of science lead to greater preference for the teaching of evolution in schools? J Sci Teach Educ. 2007;18(5):699–723.
Nehm RH, Beggrow EP, Opfer JE, Ha M. Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach 2. 2012;74(2):92–8.
Nehm RH, Finch SJ, Sbeglia GC. Is active Learning Enough? The contributions of misconception-focused instruction and active-learning dosage on student learning of evolution. Bioscience. 2022;72(11):1105–17.
Neumann I, Neumann K, Nehm R. Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. Int J Sci Educ. 2011;33(10):1373–405.
NGSS Lead States. Next Generation Science standards: for States, by States. Washington, D.C.: National Academies; 2013.
Nissen JM, Her Many Horses I, Van Dusen B. Investigating society's educational debts due to racism and sexism in student attitudes about physics using quantitative critical race theory. Phys Rev Phys Educ Res. 2021;17(1):010116.
Okebukola PA. Attaining meaningful learning of concepts in genetics and ecology: an examination of the potency of the concept-mapping technique. J Res Sci Teach. 1990;27(5):493–504.
Opfer JE, Nehm RH, Ha M. Cognitive foundations for science assessment design: Knowing what students know about evolution. J Res Sci Teach. 2012;49(6):744–77.
Petto AJ, Mead LS. Misconceptions about the evolution of complexity. Evo Edu Outreach. 2008;1(4):505–8.
Phillips BC, Novick LR, Catley KM, Funk DJ. Teaching tree thinking to college students: it's not as easy as you think. Evo Edu Outreach. 2012;5(4):595–602.
Riegle-Crumb C, King B, Irizarry Y. Does STEM stand out? Examining racial/ethnic gaps in persistence across postsecondary fields. Educ Res. 2019;48(3):133–44.
Robitzsch A, Kiefer T, Wu MTAM. Test Analysis Modules. 2024. https://cran.r-project.org/web/packages/TAM/index.html
Romine WL, Walter EM, Bosse E, Todd AN. Understanding patterns of evolution acceptance—A new implementation of the measure of Acceptance of the theory of evolution (MATE) with midwestern university students. J Res Sci Teach. 2017;54(5):642–71.
Satterthwaite FE. Synthesis of variance. Psychometrika. 1941;6(5):309–16.
Sbeglia GC, Nehm RH. Building conceptual and methodological bridges between SSE's diversity, equity, and inclusion statement and educational actions in evolutionary biology. Evolution. 2024;78(5):809–20.
Society for the Study of Evolution (SSE). Diversity statement. 2017. https://www.evolutionsociety.org/content/diversity-statement.html
Speth EB, Shaw N, Momsen J, Reinagel A, Le P, Taqieddin R, et al. Introductory biology students' conceptual models and explanations of the origin of variation. CBE—Life Sci Educ. 2014;13(3):529–39.
Stemler S, Naples A. Rasch measurement v. item response theory: knowing when to cross the line. Pract Assess Res Eval. 2021;26(1):11.
Tofade T, Elsner J, Haines ST. Best practice strategies for effective use of questions as a teaching tool. Am J Pharm Educ. 2013;77(7):155.
Whitcomb KM, Singh C. Underrepresented minority students receive lower grades and have higher rates of attrition across STEM disciplines: a sign of inequity? Int J Sci Educ. 2021;43(7):1054–89.
Wright BD. Solving measurement problems with the Rasch model. J Educ Meas. 1977;14(2):97–116.
Wright BD, Linacre M. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8(3):370.
Wynn AN, Pan IL, Rueschhoff EE, Herman MAB, Archer EK. Student misconceptions about plants – A first step in building a teaching resource. J Microbiol Biol Educ. 2017;18(1):18.1.40.
Yangin S, Sidekli S, Gokbulut Y. Prospective teachers' misconceptions about classification of plants and changes in their misconceptions during pre-service education. JBSE. 2014;13(1):105–17.
Ziadie MA, Andrews TC. Moving evolution education forward: a systematic analysis of literature to identify gaps in collective knowledge for teaching. CBE—Life Sci Educ. 2018;17(1):ar11.

Publisher's note