NGSS-Aligned Test Items to Measure High School Students' Understanding of Evolution

George E. DeBoer, Joseph Hardcastle, Jo Ellen Roseman
AAAS Project 2061

Presented at the 2019 Annual International Conference of NARST
Baltimore, MD
April 3, 2019

Abstract

This is a report on the development and validation of a set of 88 assessment items designed to test high school students' understanding of ideas related to evolution, aligned to learning goals in the NRC *Framework for K-12 Science Education* and *Next Generation Science Standards*. The test items assess students' understanding of ideas about natural selection, common ancestry, speciation, heredity, the molecular basis of genetic transmission, along with the science practices of explanation, quantitative reasoning, and evidence-based argumentation. There are 88 unique test items organized into eight equivalent forms. Four of the forms were pilot tested on students prior to their having instruction on the target ideas, and four forms were pilot tested following instruction. Evidence is presented on the reliability and validity of the tests and their suitability as both pre- and posttests. Both classic test theory and item response theory (Rasch) methods were used in the analyses.

In this paper, we report on the development and validation of a set of 88 assessment items designed to test high school students' understanding of ideas related to evolution that are aligned to learning goals in the NRC *Framework for K-12 Science Education* (National Research Council, 2012) and *Next Generation Science Standards* (NGSS Lead States, 2013). The test items were developed in connection with a curriculum unit, *Evolution: DNA and the Unity of Life*, that was being developed at the same time at the University of Utah's Genetic Science Learning Center (Drits-Esser, et al., 2018). The curriculum unit and test items were both aligned to NGSS learning goals but not intentionally to each other, and they were developed independently by research teams at the two different institutions. The items were intended to be used to test students' understanding of the NGSS ideas, not limited to any specific curriculum.

A variety of assessments have previously been created, including both comprehensive tests of the evolution topic as well as tests that focus on specific subtopics. The most familiar, often-used, and closely examined comprehensive test is the *Conceptual Inventory of Natural Selection* (CINS) by Anderson et al., 2002. Other tests assess more specific topics, such as common ancestry and tree-thinking (Kummer, 2017; Smith, et al. 2013), genetic drift (Price et al., 2014; Smith, et al. 2008), EvoDevo (Perez et al., 2013), and genetic dominance (Abraham, et al., 2014). Those tests were developed largely for undergraduate students, either majors or non-majors in biology, and were not written to be aligned to learning goals for high school students.

The set of items we developed treat the topic comprehensively and are based on ideas outlined in *NGSS* and the NRC *Framework*, including their emphasis on testing key concepts along with the practices of science. Target learning goals include ideas about natural selection, common

ancestry, speciation, heredity, the molecular basis of genetic transmission, and the science practices of explanation, quantitative reasoning, and evidence-based argumentation. The items can be organized to test specific sub-topics or sampled to test the topic more generally. With different test forms, there is also the opportunity to use different forms in pre- and post-testing.

Altogether, there are 88 unique test items organized into eight equivalent test forms. Four of the forms were pilot tested on students prior to their having had instruction on the target ideas (pretest forms), and four forms were pilot tested following instruction (posttest forms). The posttest forms were used both with students in a business-as-usual setting, in which the students received standard instruction on the topics, as well as in a setting where students received an instructional unit developed by education researchers to target the evolution and related ideas. Teachers in the business-as-usual classes were given a list of the target learning goals at the beginning of the study but no instruction on how to accomplish them. Each form has 28 items, nine of which are common across all eight forms. Each of the eight forms has two constructed response (CR) items and 26 multiple choice (MC) items.

Methods

Assessment Development. Development of the assessment items began by reviewing the relevant NGSS learning goals, including performance expectations, evidence statements, disciplinary core ideas, science practices, and related statements from the *NRC Framework*. Sub-ideas were written to clarify what was expected of students under each key idea. Research on student learning was examined to identify common misconceptions, which were then incorporated into the items as distractors. (See DeBoer, et al. for a more detailed description of assessment development procedures that were followed.)

Items were pilot tested with 4,588 middle and high school students throughout the U.S. during the fall of 2015 and spring of 2016. To obtain feedback from students about the items during pilot testing, students were asked to choose the correct answer, explain why the answer was correct, and indicate if any language in the item was confusing. Items were revised based on an analysis of student answer choice selections and their written pilot test feedback.

Revised test forms were used as pre- and posttests during pilot testing of the curriculum during the fall of 2016 and spring of 2017. Four forms were randomly distributed in each classroom, first as pretests and then again as posttests after the unit was completed, with each student receiving a different test. A total of 944 students participated. Most students were in 9^{th} (44%) or 10^{th} (46%) grade with a smaller percentage in 8^{th} (2%), 11^{th} (4%), and 12^{th} grade (4%).

Items were again revised based on an examination of the students' answer choice selections and item reliabilities. Following revision, 88 items were selected for the final version of the test forms. Items were distributed across four pretest forms and four posttest forms as described above. These tests were then used to measure the effectiveness of the curriculum in a randomized control (RCT) study. Students in control (business-as-usual) and treatment (evolution unit) groups were randomly assigned one of the four pretest forms and, following instruction, one of the four posttest forms. After removing students who did not complete the test, the control group comprised 1,104 students and the treatment group 1,165 students. The MC test items were scored dichotomously, and the CR items were scored according to scoring rubrics described below.

Three constructed response items were used in the RCT study only. One item that dealt with common ancestry was used on all pre- and posttest forms; one item that dealt with a natural

selection scenario (changes in finch beak size) was used on just the pretest forms; and one item that dealt with a different natural selection scenario (changes in anole leg length ratio) was used on just the posttest forms. The common ancestry CR item asked students to provide evidence and reasoning to support a scientific claim. The natural selection CR items asked students to construct a scientific argument that included a claim about a natural phenomenon, evidence in the form of scientific data that supported the claim, and reasoning that used appropriate scientific principles and justified why the data counted as evidence for the claim. A scoring rubric was created for the CR items that included between nine and 12 scoring "elements" depending on which item was being scored. Each element was scored dichotomously so that a student could earn up to nine points on one of the CR items and 12 points on the other two CR items. MC and CR data were then combined, which yielded 85 MC data points and 33 CR scoring elements.

Key ideas, sub-ideas, test items, misconceptions, test forms, and RCT study results can be found at assessment.aaas.org under the Evolution Project tab.

Data Analysis. Both Classical Test Theory (CTT) and Item Response Theory (IRT) were used to analyze data. Rasch analysis was used for IRT analysis and was conducted using the software WINSTEPS (Linacre, 2016). Each element of the constructed response rubric was treated as a dichotomously scored item. We used Rasch to measure item difficulty and student ability, and to determine the reliability of the item and student measures. Data from the pre/posttests was stacked so that each instance of a student taking a test was treated as unique, and item difficulties were viewed as being constant.

Items that had poor fit to the Rasch model were examined to determine if there was anything in the item could be responsible for the misfit. Based on that analysis, we eliminated one MC item and three CR scoring elements. To decrease the influence of guessing on the MC item measures we used an approach outlined by Andrich *et al.* (2012) in which student responses with large z-residual values are treated as missing data. For multiple choice items, we replaced student responses that had z-residuals greater than 4, which resulted in 302 responses being replaced. Student responses were removed because they fell far outside the expected range for the student, such as a student who scored very low overall but correctly responded to a very difficult item.

Wright maps (Wilson & Draney, 2002) were used to compare student ability (Rasch student measure) and item difficulty (Rasch item measure). On a Wright map, the distribution of person abilities in logits appears vertically from lowest to highest, and next to it the distribution of item difficulties vertically from easiest to hardest. Almost all scores fall in the range between -4.0 logits and +4.0 logits and most within +/-2.0 logits.

Results

Table 1 summarizes the fit statistics for the pretest and posttest data from all MC and CR items combined. The item and person separation indexes, which indicate the number of levels into which items and individuals can be reliably placed, were high, indicating a wide range of item difficulties and person abilities.

Table 1: *Summary of Rasch Fit Statistics for Multiple-choice & Constructed Response Items*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.03	0.14	0.07	0.36	1.84	0.38

Infit mean-square	0.77	1.38	1.02	0.42	2.24	0.99
Outfit mean-square	0.58	1.65	1.00	0.15	9.78	0.85
Point-measure correlation	0.10	0.62	0.41	-0.21	0.85	0.46
Separation index (Reliability)	17.65 (1.00)			2.66 (0.88)		

Although all items had positive point-measure correlations, seven items were found to have relatively *large* outfits (>1.4) and two items had relatively *low* outfits (<0.6). We also found that when the MC and CR items were modeled together, many of the CR items had low point-measure correlations with the total set of items. If we were trying to optimize the efficiency of a single test, we would eliminate some of those misfitting items, but in this situation, it was informative to keep them in the set to see how they related to the full complement of items.

To explore whether the items were measuring a unidimensional construct, we conduced principal components analysis (PCA) on the Rasch residuals (Wright 1996). The highest PCA eigenvalues were found to be 3.0, 2.4, 2.1, and 1.8, suggesting the possibility that the set of items was measuring multiple dimensions. Simulation studies have found that for sets of items similar in length to ours, eigenvalues greater than two falsify the hypothesis that the variance in the residuals is due to randomness in the data. For the first value, we found that constructed response items loaded highly on the dimension and MC items had low loading on the dimension. When we looked at the content of the MC and CR items, it was clear that the MC items asked students to recognize correct statements based on their evolution content knowledge, but the CR items asked them to use their knowledge to construct an argument. Because of this, we decided to analyze the two types of items separately.

Multiple-choice Items

Table 2 summarizes the fit statistics for multiple choice items alone. All items had positive point-measure correlations, and all items except two had acceptable infit and outfit (<1.4), indicating an overall good fit to the Rasch model. The item separation index was high, indicating a wide range of item difficulties. The student separation index was lower because a smaller number of data points (26 MC items) was used to place students on the student ability scale, although the items do reliably separate students into two distinct ability levels. A PCA analysis on the Rasch residuals of the multiple choice items found no eigenvalues over two, indicating that any extra dimensions in the data are comparable to what we would expect by random chance, and thus the MC items were measuring a unidimensional construct.

Table 2: *Summary of Rasch Fit Statistics for Multiple-choice test*

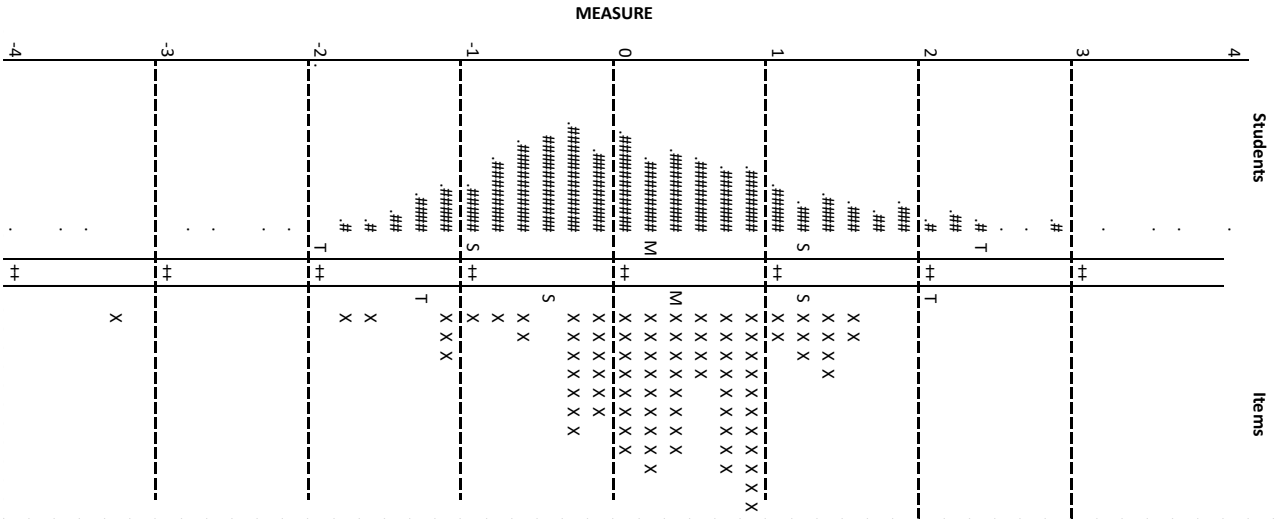
	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.03	0.15	0.07	0.42	1.90	0.44
Infit mean-square	0.77	1.31	0.99	0.46	1.74	1.00
Outfit mean-square	0.59	1.52	0.98	0.078	2.78	0.94
Point-measure correlation	0.15	0.61	0.40	-0.50	0.77	0.37
Separation index (Reliability)	11.52 (0.99)			1.98 (0.80)		

All but one item on the eight test forms had item discrimination indexes (DI) above .30, which is generally considered an acceptable level of discriminability (Garvin & Ebel, 1980). One item was considered marginal, i.e., positive but below .30. Items ranged in difficulty from about -1.9 to +3.6 logits.

On the pretest, student ability (-0.19 logits) was slightly below the average item difficulty of the items (0 logits) while on the posttest, student ability (+0.54 logits) was slightly above the difficulty of the items. This suggests that the test can discriminate between students of different abilities and is not overly difficult for students who *have not* had high school level instruction on this topic or overly easy for students who *have* had that instruction.

Figure 1 shows the Wright map for the multiple-choice items. Some items have redundant difficulties, that is, they appear at the same difficulty level on the map as other items. This could be problematic (at least inefficient) if items of the same difficulty also measured the same concept (true redundancy), but when they measure different topics (e.g., ideas from speciation and common ancestry), it is not undesirable. Although this does not help spread students out on the ability dimension, it does provide information about specific ideas they have. With only a few exceptions, the items with redundant difficulties were testing different ideas.

Figure 1: *Wright map for the Multiple-choice test*



Instructional Sensitivity of the MC Items. If the items are a valid measure of student knowledge of evolution as described in *NGSS* and the *NRC Framework*, the students who received instruction on each topic should demonstrate growth in their understanding and improved scores on the test items. This should be true for the students who received the evolution unit and for the students who received business-as-usual instruction.

Our data show that students who received business-as-usual instruction and students who received the evolution curriculum unit made significant gains on the test items. In addition, the increase in the number of students answering items correctly was significantly higher for students who received the curriculum unit than for students who received business-as-usual instruction. Overall gains, and gains on each topic, appear in Table 4.

Table 4: *Summary of gains in percent correct for each topic assessed*

Topic	Group	Percentage point gains	*Cohen's d*
Argumentation	Treatment	8.5	0.86
	Control	1.6	0.15
Common Ancestry	Treatment	18.4*	0.99
	Control	9.4*	0.46
Heredity	Treatment	18.0*	0.96
	Control	2.8	0.15
Natural Selection	Treatment	15.5*	1.21
	Control	8.1*	0.57
Shared Biochemistry	Treatment	14.5*	0.99
	Control	1.4	0.08
Speciation	Treatment	17.5*	1.8
	Control	6.7*	0.68
Overall	Treatment	17.0*	1.10
	Control	6.2*	0.40

Note: Percentage point gains are the change in the percentage of students who answered the items under each topic correctly, averaged across all the items for that topic.

Statistically significant gains in student performance and substantial effect sizes were found for both control and treatment groups for the topics of Common Ancestry, Natural Selection, and Speciation. Students who received the evolution curriculum unit (treatment group) also showed statistically significant gains for the topics of Heredity and Shared Biochemistry, but the students who received the business-as-usual instruction (control group) did not. This was not unexpected since students in the business-as-usual group had not received formal instruction on those two topics prior to testing. (Teachers were selected for the RCT study who had not yet taught heredity to their students.) Neither group showed statistically significant gains on the MC argumentation items. For the most part, the MC argumentation items assessed students' ability to recognize claims, evidence, and reasoning, which is a skill that was included in the treatment curriculum but probably not explicitly targeted in business-as-usual evolution instruction. While the treatment group improved by 8.5 percentage points on the Argumentation topic, this gain was not statistically significant, likely due to a lack of statistical power.

Constructed Response Items

As previously noted, there were three CR items. One of the items deals with the topic of common ancestry and appeared on all forms of the test, both as a pretest and as a posttest. One item that deals with natural selection appeared only on the pretest and another item that deals with natural selection appeared only on the posttest. The items were different in the content covered and in how they probed students' ability to formulate a scientific argument. The items and scoring rubrics appear in Appendix A.

Rasch Modeling of the CR Items. Table 5 summarizes the fit statistics when the elements of the constructed response items were modeled together as a single test. Each element of the constructed response items was scored dichotomously and treated as a single item in the Rasch

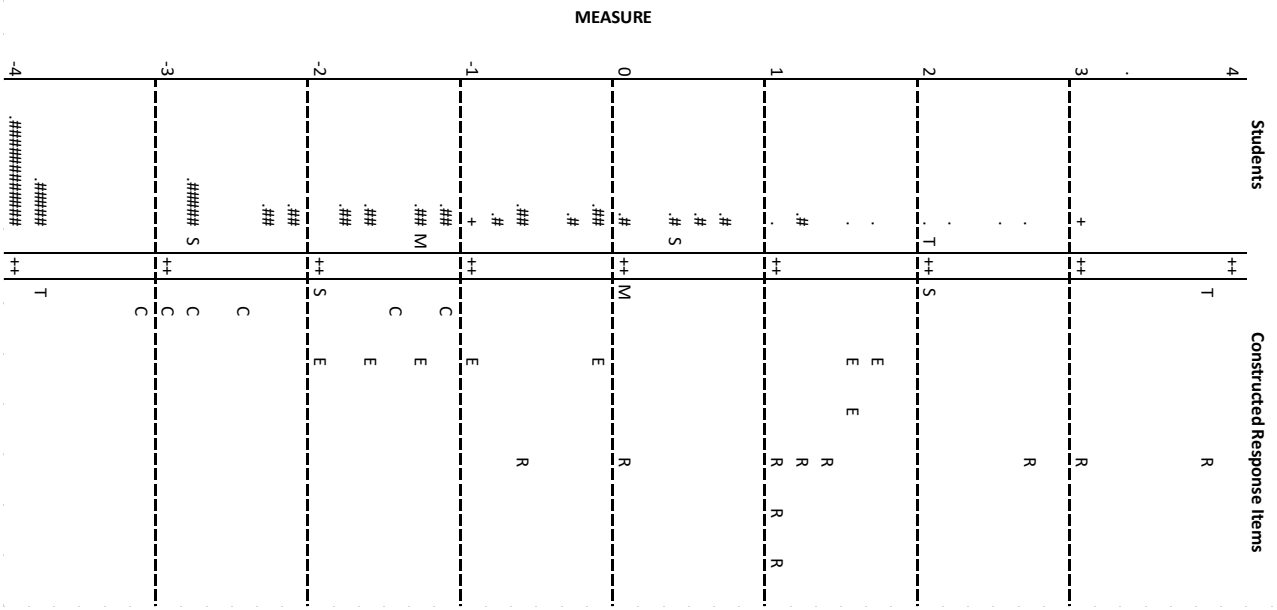
model. When modeled this way, all rubric elements had positive point-measure correlations, and all but three elements had acceptable infit and outfit (<1.4), indicating an overall good fit to the Rasch model. The item separation index was high, indicating a wide range of item difficulties. However, the student separation index was only 1.59, indicating that the CR elements cannot reliably sort students into even two distinct ability levels. Evidence of low reliability of the elements is also seen in a median point-measure correlation of 0.37 for the CR elements, which indicates that, on average, the individual CR elements do not correlate well with the total scores for the constructed response CR items. All constructed response items had item discrimination indexes (DI) above .30.

Table 5: *Summary of Rasch Fit Statistics for Constructed-response Items*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.04	0.13	0.07	0.66	1.90	0.75
Infit mean-square	0.81	1.26	0.96	0.33	2.81	1.00
Outfit mean-square	0.55	1.87	0.92	0.14	9.90	0.80
Point-measure correlation	0.18	0.77	0.53	-0.38	0.88	0.37
Separation index (Reliability)	26.28 (1.00)			1.59 (0.72)		

Figure 2 shows the Wright map for the CR items. Scoring elements that were associated with Claims (C), Evidence (E), and Reasoning (R) are shown on the item side of the map. As seen there, making claims, citing evidence, and providing reasoning were of increasing difficulty to students. While most students were able to make claims, fewer adequately cited evidence, and fewer still provided reasoning based on the appropriate evolution concepts.

Figure 2: *Wright map for the Constructed Response test*



Overall, constructed response items ranged in difficulty from about -3.2 to +3.8 logits. On the pretest, the average student ability on that construct was -2.4 logits, and on the posttest the average student ability was -1.8 logits. On both the pretest and the posttest, the average student

ability was less than the average difficulty of the items. In other words, the constructed response items were difficult for students both before and after they had high school level instruction on the topic of evolution.

Instructional Sensitivity of the Constructed Response Items. To help establish the validly of the CR items, we looked at the gains in performance on the combined CR scoring elements using Rasch to model all data together. The data show that the students who received business-as-usual instruction and students who received the evolution curriculum unit made significant gains in their ability to write arguments using natural selection and common ancestry ideas. Students who received business-as-usual instruction had an average increase in their Rasch measure of 0.19 logit ($p<0.01$) and students who received the evolution unit had an average increase in their measure of 1.04 logits ($p<0.01$).

When we looked at claim, evidence, and reasoning separately for all three items, using percent correct (i.e., average percentage of points earned) as our measure of item difficulty, we found that the treatment group students made gains on all three aspects of the argumentation practice, but the control group students made gains only on their ability to write claims. We also found that the increase in the number of students receiving points on each aspect of the argumentation practice was significantly higher for students who received the curriculum unit than for students who received business-as-usual instruction. Gains per argumentation topic appear in Table 6.

Table 6: *Summary of gains in points earned for the claim, evidence, and reasoning topics in the CR items*

Topic	Group	Percentage point gains	*Cohen's d*
Claim	Treatment	16.6**	0.38
	Control	4.8**	0.11
Evidence	Treatment	10.1**	0.34
	Control	1.3	0.04
Reasoning	Treatment	5.2**	0.27
	Control	-0.7	0.04

Note: Percentage point gains refers to the change in the percentage of students who answered the items under each topic correctly, averaged across all the items for that topic.

Although both groups had statistically significant gains on argumentation overall, Table 6 shows that only students who received the treatment had gains in their ability to include evidence and reasoning in their arguments. Because treatment group students received explicit instruction on the claim, evidence, and reasoning protocol for scientific argumentation, it is not surprising that the treatment students made greater gains. The fact that the items are sensitive to instruction that targets the ideas being tested by the items lends support to the idea that the CR items are a valid measure of students' ability to formulate a scientific argument. Additional information about what could be learned from the student responses to the CR items, including student example responses, appears in Appendix B.

Using Student Results to find out what Students do and do not Know and the Misconceptions they have

The MC and CR items can be used to pinpoint specific aspects of the knowledge students have as well as the ideas that are problematic for them, which can then guide instruction. Using resources on the website, *assessment.aaas.org*, users can cluster items that address sub-ideas they are interested in, perhaps giving students the items before and after instruction. Under the key idea of speciation, for example, it is possible to examine student understanding of two sub-ideas. The first has to do with how you can tell if two populations of organisms are different species: *It is evidence that two populations are different species if they cannot reproduce with each other*. On that sub-idea, using our data, the control group improved by 17.5 percentage points and the treatment group by 26.5 points. On a related sub-idea, that *the alleles of individuals of the same species that reproduce with each other mix, and that this allele mixing affects the traits of their offspring*, the gains were much smaller. There was a 7.5 percentage point gain for the treatment group on that sub-idea and no gain between pretest and posttest for the control group. Clearly, students need more help on the second idea than the first.

At the sub-idea level for the common ancestry topic, there is a statement that *the greater the number of traits two organisms or types of organisms have in common, the more recent ancestor they share (or the more closely related they are*). For this idea, both control and treatment group students improved. The gain was 7.7 percentage points for the control group students and 12.0 percentage points for the treatment group students. On the second sub-idea that *all organisms share a common ancestor*, control group students improved by 12.0 percentage points and treatment group students by 30.8 percentage points. Again, the information can be used to pinpoint student difficulties and to guide instructional decisions.

On the CR items, student responses revealed several significant problem areas for students. One is that they are rarely inclined to discuss the "distribution" of a trait, even when given graphical data that shows how that trait (e.g., beak sizes) is distributed in a population. This is an idea that middle school students are expected to have learned in their mathematics classes, but it is not something that they appear likely to use in the context of a science question like this. That a trait varies in a population, that the proportion of individuals having different levels or values of that trait change over time, and that this is what causes the shifts in the characteristics of a population of organisms, are critical for understanding the process of natural selection. This application of basic mathematical knowledge to science ideas is clearly one that needs more attention.

We also found that students are not inclined to provide mechanisms to explain how phenomena occur. On the item that asked them to cite evidence and provide reasons why two organisms with greater genetic similarity share a more recent common ancestor than two organisms with less genetic similarity, only a very small percentage of students showed they can work through the steps in the process, including the idea that genes are inherited, that mutations in genes lead to changes in heritable traits, and that mutations accumulate over time so that species with fewer differences have a more recent common ancestor. Whether it is too much to expect high school students to construct such arguments is a question worth asking. We did so here to find out how they would respond, which seemed to us to be an important first step.

Conclusions

Fit of the entire set of data to the Rasch model provides evidence for the overall reliability of the test items as a measure of student understanding of evolution and associated key ideas. The intentional and precise alignment of items to learning goals and the items' instructional sensitivity provide evidence of their validity. The fact that the MC and CR items clustered as separate dimensions suggests that each of the two types of items can reveal information about student understanding that the other alone cannot provide.

For the MC content items, there is strong evidence that the items are a valid measure of students' ideas about evolution, natural selection, and related topics. A closer analysis of student performance on some of the sub-ideas provides additional information about what can be learned about student thinking from these items and the instructional sensitivity of those sub-ideas. These observations, both at the key idea level and sub-idea level, suggest that the items can be used to pinpoint what students know, what they do not know, and the misconceptions they have.

Regarding the CR items, the items were shown to be valid measures of ideas targeted in instruction based on their instructional sensitivity. This was true for the treatment group for all three components of the claim, evidence, and reasoning protocol, but only true on the claim portion for the control group.

We have also demonstrated that these CR items are testing a separate dimension from the MC content items, and users will have to decide how they will use the items and how they will score them. When the MC and CR items are modeled together, many of the CR items have low point-measure correlations with the total set of items. This means that students who did relatively well on the MC items did not necessarily do well on the CR items. Why is this if both item types were intended to measure students' understanding of evolution? We believe that it may be because there were not enough ways for students to demonstrate their understanding of scientific argumentation in the context of evolution at lower levels of understanding on the CR items. It is also possible that the items were not explicit enough about what was expected of students. Students who could have written more complete arguments may not have done so because they were not aware of what was expected of them. Our scoring rubric for the CR items gives points for specificity and completeness, which are not assessed in the MC items, either on the content items or the MC argumentation items. For example, on the natural selection CR items, where students are asked to write an argument about whether natural selection could have caused certain changes (shown graphically) in the finch populations at two different points in time (or in the anole's leg length ratio), students get zero points for saying **yes, it could have**, one point for stating that **natural selection could have caused the change**, another point for specifying that **the trait that was changed was the beak size**, and a third point for specifying that **an environmental change (drought or reduction in the number of small seeds) was responsible for the change**. Simply stating that natural selection caused the change and then providing evidence and reasoning for it is not enough to get the maximum number of points on the claim portion of the item. The Rasch analysis tells us that this is expecting something of students that was not measured in the MC argumentation items and suggests that students' ability to identify correctly stated claims, evidence, and reasoning may be different from their ability to write arguments, at least in the way it was measured by our scoring rubric.

The important question is whether this added dimension is something that we want to be measuring. We think it is, and we recommend that the CR and MC items be used together so that

users of the test items learn not only what students know, in the sense of recognizing correct and incorrect statements, but also learn how well students can construct arguments and the extent to which they are inclined to offer complete and thorough explanations for what they see. Users will have to decide for themselves how to allot points to students for the answers they give. Our scoring rubrics, with their emphasis on detailed elements of an argument, are offered as a starting point.

Our work also points to areas where the set of items can be improved. Additional items at the lower end of the difficulty range could further improve the reliability and more accurately measure students' ability there. Adding items that test simpler ideas, perhaps middle school level ideas, could be used to test the lower range of student ability. Similarly, adding simpler argumentation constructed response items or rubric elements based on middle school level ideas and practices could improve the constructed response argumentation items.

Significance

This work should be of interest to researchers currently developing high school level, NGSS-aligned curriculum units and assessments on evolution and natural selection. The work should also be of interest to teachers who need reliable assessment items for measuring students' understanding of evolution.

Acknowledgements

The work reported here was funded by the National Science Foundation through Grant *DRL-1418136* to the University of Utah and subcontract to the American Association for the Advancement of Science. The opinions expressed are those of the authors and do not represent views of the National Science Foundation.

We also acknowledge the work of former staff members—Jean Flannagan, Linda Wilson, Martin Fernandez, and Bernard Koch—who contributed to the item development work.

References

Abraham, J. K., Perez, K. E., & Price, R. M. (2014). The dominance concept inventory: A tool for assessing undergraduate student alternative conceptions about dominance in mendelian and population genetics. *CBE Life Sciences Education*.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*.

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. J*ournal of Educational and Behavioral Statistics*.

DeBoer, G., Herrmann Abell, C., Gogos, A. Michiels, A., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. In J. Coffey, R. Douglas, and C. Stearns (Eds.) *Assessing Science Learning: Perspectives from Research and Practice* (231-252). Arlington, VA: National Science Teachers Association Press.

Drits-Esser D., Homburger S., Malone M., Hawkins A.J., Bass K., Roseman J.E., DeBoer G., Hardcastle J., & Stark L.A. (2018). *Development and pilot testing of an NGSS-aligned unit*

that integrates evolution and heredity. Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.

Garvin, A. D., & Ebel, R. L. (1980). Essentials of Educational Measurement. *Educational Researcher*.

Kummer, T. (2017). *Assessing and Improving Student Understanding of Tree-Thinking*. Brigham Young University.

Linacre, J. M. (2016). Winsteps ® Rasch measurement computer program. Beaverton, Oregon. Retrieved from Winsteps.com

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. (C. on a C. F. for N. K.-12 S. E. S. B. on S. E. D. of B. and S. S. and Education, Ed.). Washington DC: The National Academies Press.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington DC: The National Academies Press.

Perez, K. E., Hiatt, A., Davis, G. K., Trujillo, C., French, D. P., Terry, M., & Price, R. M. (2013). The evodevoci: A concept inventory for gauging students' understanding of evolutionary developmental biology. *CBE Life Sciences Education*.

Price, R. M., Andrews, T. C., Mcelhinny, T. L., Mead, L. S., Abraham, J. K., Thanukos, A., & Perez, K. E. (2014). The Genetic Drift Inventory: A Tool for Measuring What Advanced Undergraduates Have Mastered about Genetic Drift. *CBE—Life Sciences Education*.

Smith, J. J., Cheruvelil, K. S., & Auvenshine, S. (2013). Assessment of student learning associated with tree thinking in an undergraduate introductory organismal biology course. *CBE Life Sciences Education*.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concept Assessment: A New Concept Iventory for Gauging Student Understanding of Genetics. *CBE Life Sciences Education*.

Wilson, M., & Draney, K. (2002). A Technique for Setting Standards and Maintaining Them Over Time. In *Measurement and Multivariate Analysis* (pp. 325–332). Springer, Tokyo.

Appendix A
Items and Scoring Rubrics

Item 1. **Common Ancestry (This item appeared on all test forms, both pre- and posttest.)**

Scientists studying evolution compared the DNA of chimpanzees, gorillas, and orangutans.

The scientists summarized their data in the following table:

Pair of Species Compared	Average Genetic Similarity
Chimpanzee and Gorilla	98%
Chimpanzee and Orangutan	97%
Gorilla and Orangutan	97%

When the scientists published their research, they made the following claim:

"Chimpanzees and gorillas have a more recent common ancestry than chimpanzees and orangutans."

What **evidence** and **reasoning** are the scientists using to make this **claim**?

Your answer should include **evidence** in the form of specific scientific data that supports the scientists' claim, and **reasoning** that uses scientific principles about heredity and common ancestry to justify why the data counts as evidence for their claim.

How the item was scored:

Evidence (3 points)

- **Cites specific data that supports** the claim (**98%** genetic similarity for chimp and gorilla and **97%** genetic similarity for chimpanzee and orangutan) – 1 point
- **Avoids citing data that is irrelevant** to the claim (97% genetic similarity for gorilla and orangutan) – 1 point
- **Summarizes data** (chimpanzee and gorilla have greater **genetic/DNA similarity** than chimpanzee and orangutan) – 1 point

Reasoning (4 points)

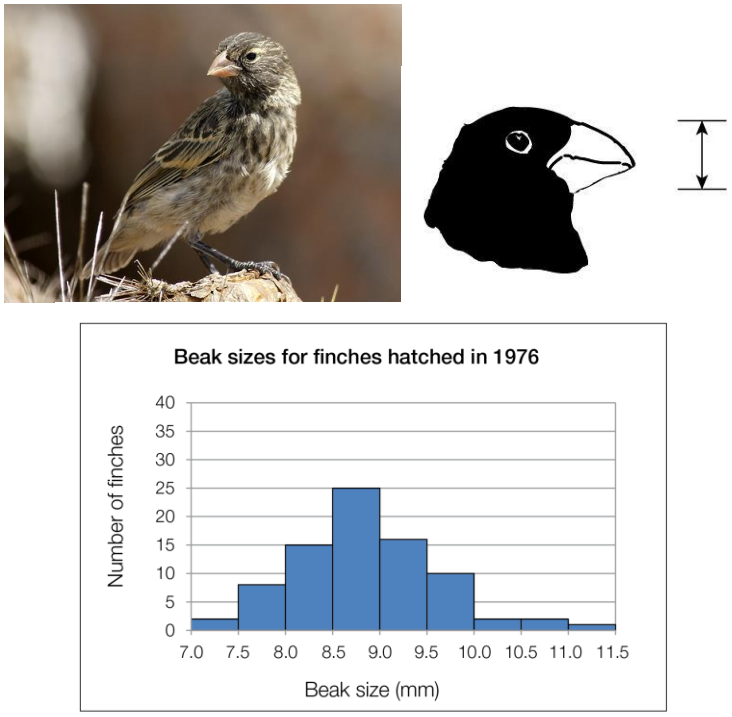
- **States the general principle** that links evidence to claim (i.e., the greater the genetic similarity between 2 species, the more recently their common ancestor existed, or species with more genetic similarity have a more recent common ancestor) – 1 point
- Provides **a mechanism** to explain the link between the claim and the evidence
    - Includes the idea that **genes/DNA are inherited** from parents – 1 point
    - Includes the idea that **changes occur in genes/DNA**, through mutation or recombination) – 1 point
    - Includes the idea that **changes accumulate** over time – 1 point

---

***Item 2*. Natural Selection (This item appeared only on the pretest)**

The ground finch is a species of bird. Seeds are the finches' main source of food. Finches with small beaks can eat only small seeds, but finches with large beaks can eat both small and large seeds. Beak size is inherited.

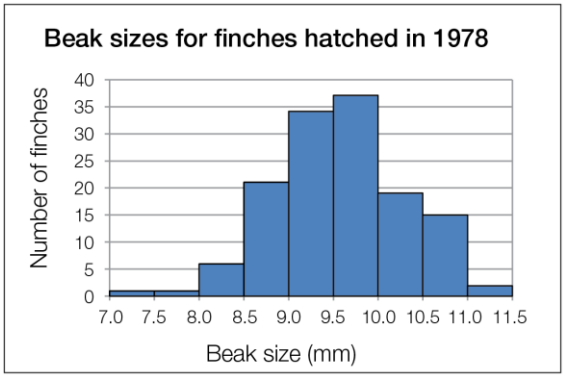
In 1976 scientists collected data on the size of the beaks of a population of finches that hatched that year. The graph below shows the beak size (in millimeters) of the 1976 sample.



Left: An individual female ground finch. Middle: A drawing showing the part of the finches' beaks that the scientists were measuring. Right: Graph showing beak size (in millimeters) of ground finches hatched in 1976.

In 1977, during a long period without rain many plants that produced small seeds died.

To learn how this influenced the next generation of finches; scientists returned in 1978 and measured the beak size of finches that hatched that year. The graph below shows the beak size (in millimeters) of the 1978 sample.



Graph showing beak size (in millimeters) of ground finches that hatched in 1978.

Do you think the process of natural selection caused the changes in the finch populations between 1976 and 1978?

Write your answer in the form of an **argument**. Your argument should include: A **claim** that answers the question, **evidence** in the form of specific scientific data that supports your claim, and **reasoning** that uses appropriate scientific principles and justifies why the data counts as evidence for your claim.

How the Item was Scored:

Claim – 3 points

- States that **Natural Selection** caused the change – 1 point
- Indicates that the trait that changed was the **beak size** – 1 point
- Specifies the **environmental condition** (drought or reduction in small seeds) that is thought to have caused the change – 1 point

Evidence – 3 points

- Specifies the **directionality** of the change (increased beak size) – 1 point
- Cites data from **a single graph** - 1-point
- Compares data from **both graphs** - 1-point

Reasoning – 6 points

Students get points for identifying science ideas that explain how natural selection can be responsible for a change in the distribution of a trait over time.

Reproductive advantage – some traits give an organism a survival advantage in their environment

- **States** the general principle - 1-point (merely saying "survival of the fittest" is not enough to earn a point)
- **Applies** the general principle to link the evidence to the claim – 1 point

Heritability – genes/DNA for a trait are passed from one generation to the next

- **States** general principle - 1-point
- **Applies** general principle to link the evidence to the claim – 1 point

Variability in traits – traits vary within a population of organisms of the same species

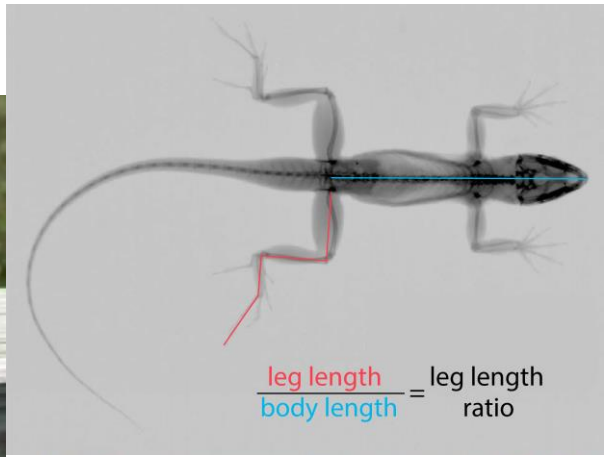
- **States general principle** - 1-point
- **Applies general principle** to link the evidence to the claim – 1 point

---

Item 3. **Natural Selection (This item appeared only on the posttest)**

Anoles are lizards that live in the southeastern United States, South America, and the Caribbean islands. Different anoles vary from each other in many ways. One trait on which anoles vary is their hind-leg length ratio. The hind-leg length ratio is the hind-leg length divided by the body length.

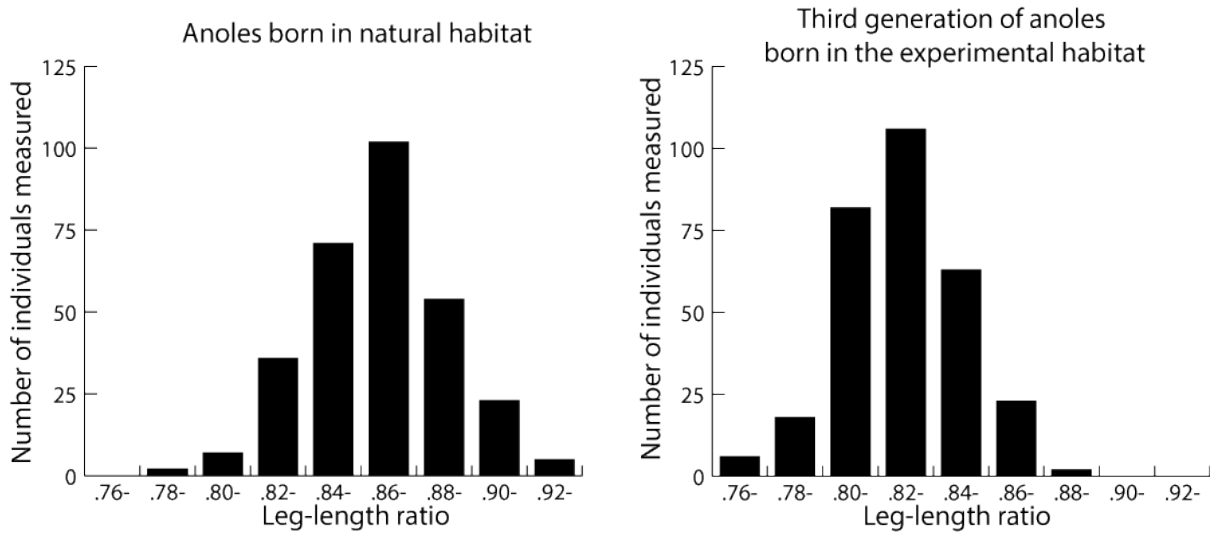
The images below show an anole and an X-ray image showing how scientists measure leg length (red line) and body length (blue line). They use these measurements to calculate the anole's leg-length ratio. Anoles that have smaller leg-length ratios are better at running on thin branches and anoles that have larger leg-length ratios are better at running on thick branches.



Left Image: Photograph of an Anole (Photo by Kristin Winchell.) Right Image: X-ray image showing how the leg-length ratio is calculated.

Adult hind-leg length ratio is a heritable trait. Scientists decided to test whether the leg-length ratio is a trait that undergoes evolution by natural selection. To do this they placed a group of anoles on small islands where there are only bushes with thin branches (no trees) and no other anoles. They called this the experimental habitat.

Each year, the scientists returned to the experimental habitat to measure the leg-length ratio of individuals from each generation of anole lizards. The graphs below show the leg-length ratios of anoles born in their natural habitat and the leg-length ratio of the third generation of anoles born in the experimental habitat.



Left Image: Distribution of Leg-length ratio in Anoles born in the natural habitat. Right Image: Distribution of Leg-length ratio in the third generation of Anoles born in the experimental habitat.

Do you think the process of natural selection caused the change in the leg-length ratio between anoles born in the natural habitat and the third generation of anoles born in the experimental habitat?

Write your answer in the form of an **argument**. Your argument should include: A **claim** that answers the question, **evidence** in the form of specific scientific data that supports your claim, and **reasoning** that uses appropriate scientific principles and justifies why the data counts as evidence for your claim.

How the item was scored

(See scoring rubric for the finch item above.

Appendix B

What we learned from student responses to the CR items

Student responses provided insights into difficulties related to students' ability to state accurate claims, to provide evidence to support their claims, and to include scientific principles that explain the phenomena they observe. We focus here on how we scored the evidence and reasoning components of a scientific argument and what we learned.

How we scored the Common Ancestry item and what we learned

Evidence. The 3-point rubric assigned 1 point for indicating that the data table shows the amount of genetic similarity among pairs of species, 1 point for citing numerical data that provides evidence that chimpanzees and gorillas (98%) had more genetic similarity than chimpanzees and orangutans (97%), and 1 point for NOT citing irrelevant data, e.g., that there is 97% genetic similarity for gorilla and orangutans. Responses that cited all the data in the table, including the data comparing the genetic similarity between gorillas and chimpanzees, did not discriminate between relevant and irrelevant data and did not earn any points for the students.

> Example of not relating evidence to the claim
> …Chimpanzees and Gorillas have a 98% average genetic similarity compared to Chimpanzees and Orangutans or Gorillas and Orangutans which have a 97% average genetic similarity.

Reasoning. The 4-point rubric assigned 1 point for stating the general principle that links the evidence to the claim (i.e., the greater the genetic similarity between 2 species the more recently their common ancestor existed), and 3 additional points for providing three elements of a causal mechanism (genes are inherited from parents, genes change through mutation or recombination, changes accumulate over time)

Rasch analysis showed that linking evidence to claim by stating the general principle was considerably easier (-0.67 logits) than providing a mechanism to explain why the general principle is reasonable (>2 logits). Arguably, the item did not explicitly ask students to provide a mechanism; but some students were inclined to do so.

Example of reasoning that stated the general principle
"a higher percent in genetic similarity indicates a more recent common ancestor."

Example of reasoning that provided a causal mechanism
"…In the course of time, species can inherit mutations, differing genetic traits that change the sequence of DNA within an individual, and eventually, perhaps, a population. It will take a long time before the DNA of a population is remarkably different. The similarity of the DNA of chimpanzees and gorillas suggests this long time has not come to pass, and they're still inheriting genes similar to those of their ancestors."]

How we scored the Natural Selection CR items and what we learned

Two of the three constructed response items assessed students' understanding of and ability to write and justify a claim about natural selection with evidence and reasoning—one in the context of a shift in distribution of finch beak length following a drought that reduced the supply of small seeds, and another in the context of a shift in the distribution of leg length ratio of anoles placed in an experimental habitat with shrubs that had smaller diameter branches. In both scenarios, as can be seen in the items shown above), students were given graphs of the distribution of the trait in the population at two different timepoints. For full description of the rubrics see the Appendix.

Evidence. The evidence component of the rubrics assigned points for indicating the direction of the change and for citing and comparing specific data from the graphs. Many students had difficulty interpreting the graphs as distributions. When describing the direction of the change, some students said that the beak size increases and then decreases, suggesting they were incorrectly interpreting the x-axis of the distribution as change over time. When citing data to provide evidence of the change in the distribution over time, ideal responses would have compared a statistical parameter like mean and range, though lay terms like middle and spread were accepted.

> "… In a chart measuring the finches' beak sizes in 1976, it is noted that the finches' average beak size is approximately 8.5 millimeters. … the chart from 1978 shows that their beak size increased from 8.5 millimeters to 9.5 millimeters..."

> "…Evidence: Common beak size in 1976 was between 8.5 and 9 millimeters. In 1978, beak size increased with the most common size being between 9.5 and 10 millimeters…"

According to the Common Core Math Standards for Grade 6 (CCSS SP3 and 4), students should know that a distribution "can be described by its center, spread, and overall shape" and be able to summarize and describe data distributions using those simple terms. Fewer than 10% of students used terms like center (average, mean, median, or mode) and spread (range), or overall shape.

Reasoning. In terms of reasoning, students were expected to provide a mechanistic explanation for the change over time using ideas about variability (at least two variants must exist for the environment to select individuals with the preferred variation), heritability (genes for the trait must be passed from parents to offspring), and reproductive advantage (one of the variants must produce more offspring). In a formal argument, students would be expected to state each general principle and then use it to explain why either the average beak size in the finch population increased or the average leg-length ratio in the anole population decreased by (a) stating each general principle and (b) applying it to the specific context. The closest students came to stating the general principle was by indicating, or at least implying, that three criteria must be met for a claim that natural selection caused the change to be valid.

> "… It was established that this trait is heritable, and the graph shows variation among the traits."

> "… Short leg-ratios are a reproductive advantage, so all three ingredients of natural selection are present. If all three are present, that means that natural selection is occurring."

Fewer than 10% of students stated these general principles. In addition, some students that mentioned the three criteria in their reasoning were not specific about how these criteria were met, suggesting they were using the terms without understanding.

> "The process of natural selection caused the changes in the leg length ratio of anoles. In order for them to survive there had to be variation, heritability, and a reproductive advantage. From the graph you can see that there was variation among the species. since the length of their legs is heritable, there is heritability. Also, there was a reproductive advantage so that they would be able to survive and also explains why the majority of the anoles have a certain leg length."

Students were more likely to directly apply the principles in their reasoning; however, this was still uncommon (fewer than 20% of students).

Many students' reasoning was based on organisms as actors who "*need*" to change their traits. Alternatively, some students focused on a trait being "*useful*," "*better,"* or *"fit."* Beyond the organism and trait focus, some students' reasoning was vague about the underlying mechanism describing it as *"adaption*" and/or *"survival."*

> [After citing evidence from the graphs] "This evidence suggests that the anoles in the experimental habitat have come to adapt to their environment to survive the longest."

This level of reasoning may stem from misconceptions, an incomplete understanding of the mechanism, and/or lack of clarity in applying and/or communicating their thinking.