Res Sci Educ
DOI 10.1007/s11165-014-9435-9

# Using a Constructed-Response Instrument to Explore the Effects of Item Position and Item Features on the Assessment of Students' Written Scientific Explanations

**Meghan Rector Federer · Ross H. Nehm · John E. Opfer · Dennis Pearl**



**Abstract** A large body of work has been devoted to reducing assessment biases that distort inferences about students' science understanding, particularly in multiple-choice instruments (MCI). Constructed-response instruments (CRI), however, have invited much less scrutiny, perhaps because of their reputation for avoiding many of the documented biases of MCIs. In this study we explored whether known biases of MCIs—specifically item sequencing and surface feature effects—were also apparent in a CRI designed to assess students' understanding of evolutionary change using written explanation (Assessment of COntextual Reasoning about Natural Selection [ACORNS]). We used three versions of the ACORNS CRI to investigate different aspects of assessment structure and their corresponding effect on inferences about student understanding. Our results identified several sources of (and solutions to) assessment bias in this practice-focused CRI. First, along the instrument item sequence, items with similar surface features produced greater sequencing effects than sequences of items with dissimilar surface features. Second, a counterbalanced design (i.e., Latin Square) mitigated this bias at the population level of analysis. Third, ACORNS response scores were highly correlated with student verbosity, despite verbosity being an intrinsically trivial aspect of explanation quality. Our results suggest that as assessments in science education shift toward the measurement of scientific practices (e.g., explanation), it is critical that biases inherent in these types of assessments be investigated empirically.

M. R. Federer (✉)
Department of Teaching and Learning, Ohio State University, 333 Arps Hall, 1945 North High Street, Columbus, OH 43210, USA
e-mail: federer.21@osu.edu

R. H. Nehm
Center for Science and Mathematics Education, Stony Brook University, Stony Brook, NY 11794, USA

J. E. Opfer
Department of Psychology, Ohio State University, Columbus, OH 43210, USA

D. Pearl
Department of Statistics, Ohio State University, Columbus, OH 43210, USA

Published online: 29 October 2014

**Keywords** Constructed response instrument · Item order effects · Item surface features · Scientific explanation · Evolution

## Introduction

Assessment of students' scientific knowledge and reasoning plays a central role in research on science teaching (NRC 2001, 2007). Well-developed assessment instruments can provide valid and reliable inferences about students' understanding and guide evidence-based teaching and learning (NRC 2001). Many assessment instruments, however, have not been shown to produce valid and reliable inferences. Consequently, these assessment instruments may fail to guide or predict learning outcomes. For assessment instruments in content-rich domains, such as biology (Nehm et al. 2012), these risks are particularly acute. This has spurred efforts to understand how different assessment formats and features differentially inform our inferences about student understanding, as well as efforts to develop new tools to measure more authentic practices and performances (Nehm et al. 2012; NRC 2001, 2007).

To improve the quality of science assessment instruments, the NRC (2001) report and the new *Framework* (NRC 2012) recommend that research-based models of cognition and learning guide assessment designs. Such recommendations are largely based on a wide array of cognitive differences documented between experts and novices (e.g., Chi et al. 1981). A cognitive model that addresses the novice-to-expert progression in student thinking is crucial as it helps to explain differences in levels of performance, thus guiding the development and interpretation of assessment instruments. While such approaches have proven useful in a variety of scientific domains (Opfer et al. 2012; White and Frederickson 1998), the potential biases intrinsic to these types of assessment tasks have not been explored in many domains. This limitation raises questions about the inferences that may be drawn about students' performances as measured by various types of practice-based assessments.

As the importance of scientific explanation becomes increasingly emphasized in science education research and policy documents (AAAS 1994, 2011; Duschl et al. 2007; NRC 1996, 2012), the inclusion and evaluation of assessments that provide students with opportunities to participate in authentic scientific practices is critical (NRC 2012). Such experiences are a necessary part of developing a deeper understanding of science, as learners are required to elicit and connect their ideas about relevant scientific concepts (Lee et al. 2011). Additionally, as the process of making connections between ideas and forming arguments may result in more coherent knowledge frameworks, evaluation of the linkages between students' scientific and naïve knowledge is important for understanding the processes by which students are constructing scientific understanding (Nehm and Ha 2011; Nehm et al. 2011). In the sections that follow, we (1) highlight some perspectives on explanatory practice in science education, (2) discuss the advantages and limitations of multiple-choice (MC) and constructed-response (CR) assessment tasks for evaluation of scientific practices, and (3) review prior assessment research that has explored item sequencing and feature effects on measures of student performance.

## Explanatory Practice in Science Education

Engaging students in the construction of scientific explanations is a central aspect of science education, research, and policy (AAAS 1994, 2011; Duschl et al. 2007; NRC 1996, 2012). The inclusion of such practices represents a shift in focus from learning as understanding of specific content knowledge to learning as an integration of disciplinary core ideas with

opportunities to participate in authentic scientific practices (NRC 2012). If students are to meet performance expectations for explanatory practice, it is necessary to have assessment instruments that are capable of measuring student performance on such scientific practices. Such assessments must be capable of evaluating both the structure of the explanation as well as the conceptual components (i.e., content knowledge). As science education and research moves in the direction suggested by the *Framework* (NRC 2012), we must ask whether current assessment practices provide valid and reliable inferences about student performance on scientific practices in conjunction with disciplinary knowledge.

Research examining students' explanatory practices has resulted in numerous perspectives about the scientific practice of explanation, what constitutes a scientific explanation, and how such explanations should be evaluated (e.g., Berland and McNeill 2012; Osborne and Patterson 2011; Russ et al. 2008). The question of what defines a scientific explanation is an important step in developing assessment tools aligned with performance expectations for the scientific practice. While space precludes a more detailed discussion of the nuances of explanatory practices in science education, we have attempted to briefly identify some of the major perspectives. Hempel and Oppenheim (1948) provide a general framework for scientific explanations as a method for answering the question "why" (or "how") rather than only the question "what." More generally, an explanation consists of an explanandum (what is to be explained) and an explanans (what is doing the explaining). Perspectives from the philosophy of science provide additional clarification of the function of scientific explanations, including providing information about a cause (causal account; e.g., Lewis 1986; Salmon 1984; Scriven 1959), to connect a diverse set of facts under a unifying principle (unification account; e.g., Friedman 1974; Kitcher 1981), or to do both (kairetic account; e.g., Strevens 2004). However, there is general consensus among philosophers of science, and among science educators and researchers, that the concept of "cause" is central to the process of scientific explanation (see Kampourakis and Zygzos 2008 for a review).

In addition to a general understanding of explanatory practice in science, it is important for science educators and researchers to consider how explanations are conceptualized within specific disciplines or for specific core ideas. For example, in describing the particular nature of *evolutionary explanations* (as a type of scientific explanation), Kampourakis and Zygzos (2008, p. 29) stated: "To give causes in an evolutionary explanation is not to give complete accounts but useful and enlightening partial accounts." While conceptualizing explanations as "partial accounts" may not be applicable for other areas of science, here it is an appropriate and acceptable explanatory account. It is clear that an understanding of what constitutes a scientific explanation within the context of specific disciplinary knowledge is an important precursor to developing performance expectations and assessments.

The diversity of perspectives presented above make it clear that in order to adequately evaluate students' explanatory practices, the science education and research community needs to establish clear guidelines and performance expectations that distinguish between the various practices by which scientific knowledge is constructed and communicated. The explanatory tasks presented in this paper align best with the perspective identified by Osborne and Patterson (2011) and Kampourakis and Zygzos (2008). That is, students are asked to provide an explanation for a natural phenomenon but are *not* asked to provide an argument or justification for why that explanation might be truthful. In addition, in evaluating student explanations, mechanistic (causal) accounts are recognized as important forms of scientific explanation that are distinct from those that are teleological or otherwise representative of non-normative reasoning patterns. Thus, students' explanatory practices in response to the Assessment of Contextual Reasoning about Natural Selection (ACORNS) instrument are evaluated for evidence of both causal (scientifically normative) and non-causal (non-normative) reasoning.

## Assessing Students' Explanatory Practice

As identified in the previous section, a more unified understanding of scientific explanation is a necessary step in the development of assessment instruments designed to evaluate student understanding through performance on scientific practices. Two widely used formats of assessment in science education are multiple-choice (MC) and constructed-response (CR) tasks. The utility of either format for classroom or research practice is dependent on the purpose of the assessment, and multiple studies have suggested that assessments based on MC versus CR items may generate different measures of competency (e.g., Bridgeman 1992; Nehm and Schonfeld 2008; Nehm et al. 2012; Opfer et al. 2012). Prior research has documented several limitations of multiple-choice instruments (MCI) and constructed-response instruments (CRI) on the measurement of student knowledge (e.g., Bennett and Ward 1993; Cronbach 1988; Liu et al. 2011; Messick 1995; Popham 2010). This section identifies some of the major advantages and limitations of each item format for assessing students' conceptual understanding through performance on an explanation task.

### Multiple-Choice Tasks

MC tests have been shown to be limited in their ability to assess the depth of students' knowledge organization, synthesis, and communication (Liu et al. 2011; Martinez 1999; Popham 2010), and may often be poor predictors of real-world performance (Nehm and Ha 2011; Nehm et al. 2011; NRC 2001). In some instances, MC tests have also been shown to produce unintended, negative consequences, such as the learning of incorrect ideas as a part of the testing process (Kang et al. 2007; Mandler and Rabinowitz 1981; Roediger 2005). While there is little doubt that MCIs are cost-effective and capable of providing reliable and valid inferences about some kinds of conceptual knowledge, they do not adequately measure all types of learning outcomes, such as the formulation of scientific explanations or other practices of communicating scientific understanding (AAAS 2011; NRC 2012). However, despite the documented limitations of MCIs, these disadvantages may be more reflective of their typical design (measuring recall using either–or choices) than their intrinsic capacity to measure complex thinking (Martinez 1999). For example, while the process of knowledge recognition and construction may not be the same, studies investigating construct equivalence between MC and CR items have documented high correlations (Rodriguez 2003).

### Constructed-Response Tasks

In contrast to MC tasks, an advantage of using CRIs is that they permit students to construct heterogeneous responses comprised of non-normative and scientific elements, providing greater insight into student thinking than assessments evaluating for "right or wrong" responses. Moreover, CRIs are generally supported as having a broader capacity for measuring higher-order cognitive processes, such as explanation and justification, eliciting different levels of cognitive activity during problem solving than MC tests (Liu et al. 2011; Martinez 1999; Ward et al. 1987). Likewise, performance on CRIs has been found to have greater correspondence to clinical interviews than some MCIs, suggesting that CRIs may provide a more valid measure of complex student reasoning (Nehm et al. 2012). Therefore, use of CRIs may elucidate student thinking and provide a more comprehensive analysis of student knowledge about complex processes than MCIs.

The above limitations of MC formats and advantages of CR formats motivated the studies presented in this paper. Specifically, this paper explores whether factors known to affect

measures from MCIs similarly affect measures from CRI and focuses on the effects of two factors on measures of student performance: item position (variation in the position of an item within a sequence) and item features (superficial characteristics of the item that are independent of conceptual understanding). The following section provides a brief overview of these two issues in the MC assessment literature and discusses the potential implications for CR assessments.

## Research on Assessment Bias

### Item Sequencing Effects

Over half a century of investigations have examined item-sequencing effects within MCIs, with research in this area concentrated around a simple but important question raised by Leary and Dorans (1985):

> If the items that compose a test are presented in one arrangement to one individual and the same items are then rearranged into a different sequence and administered to another individual, can one assume that the two individuals have taken the same test? (p. 389)

A variety of research findings have lead to some confusion regarding the role that sequencing effects may play in MCIs, but the lack of consensus notwithstanding, there are a few trends that can be identified from the item sequencing literature. Sequencing items according to their difficulty (e.g., easy to hard vs. hard to easy) has been shown to affect student performance, with sequences arranged from easy-to-hard associated with higher student performance (e.g., MacNicol 1956; Mollenkopf 1950; Monk and Stallings 1970; Sax and Cromack 1966). Mollenkopf's (1950) study on the effect of section rearrangement in verbal and mathematics aptitude tests represents one of the earliest investigations of such order effects. In this study, Mollenkopf documented significant effects of item rearrangement for verbal tests and mathematics tests. MacNicol's (1956) comparison of easy-to-hard (E–H), hard-to-easy (H–E), and randomized item sequences for a verbal test revealed that random orders produced comparable performances on E–H sequences, both of which were significantly higher than performances on H–E sequences.

Comparisons of different testing environments have generally documented decreased performance on item sequences administered under speeded (i.e., timed) conditions, as compared to power (i.e., untimed) conditions (Mollenkopf 1950). This appears to be particularly important for qualitative items (e.g., verbal/written) as opposed to quantitative items (Kingston and Dorans 1984), emphasizing the need to assess the relative impact of item location within a test on performance. Additional differences have been found relating to the type of test, such as aptitude versus achievement, with performances on tests related to aptitude skills being more susceptible to sequencing effects (e.g., Gray 2004).

The diversity of item sequencing studies using MCIs, and the general lack of research on question order effects using CRIs—despite their prevalence in recent research—motivated the studies presented in this paper. Results from prior research demonstrated significant effects of item sequencing for particular item types and arrangements. Of particular interest to the work presented in this paper are the documented differences in item sequencing effects between MC tests comprised of quantitative (e.g., computational problems) and qualitative (e.g., problems with a reading passage) items. CR assessments are inherently qualitative and therefore measures of student performance might be subject to item sequencing effects or measurement biases, such as errors of omission (i.e., leaving out previously stated information), similar to those of qualitative MC tests.

### Item Feature Effects

Early research on surface features in science education focused on problem representation and categorization in an attempt to identify the variety of ways in which novices represented and solved problems using their prior knowledge and experiences (e.g., Chi et al. 1981). In these initial expert–novice studies, researchers demonstrated that novice problem solvers tend to categorize problems according to the item surface features rather than recognizing common conceptual themes or groupings. Surface features can be defined as the superficial characteristics of the item that can be changed without altering the underlying concept. Focusing on these can be problematic for novices if the surface features of otherwise isomorphic items are perceived as different, thereby influencing the elicitation and measurement of different ideas.

Since the initial studies, significant effects of item surface features have been documented in a variety of scientific domains, using both MCIs and CRIs. For example, familiarity with the construct to be tested has been associated with higher performance (relative to performance on unfamiliar constructs) on MC assessments and higher confidence in response accuracy in physics education (e.g., Caleon and Subramaniam 2010) and chemistry education (e.g., Rodrigues et al. 2010). In addition, the use of isomorphic problems (with variable surface features) to study student expertise and problem solving processes in chemistry and physics suggests that the transfer of knowledge across problem contexts is inhibited by misconceptions (naïve ideas [NIs]) about scientific concepts and that the process of transfer can be particularly difficult if the items do not share common surface features (e.g., Singh 2008).

In the domain of biology, Clough and Driver (1986) were among the first to acknowledge and explore item context effects in evolutionary explanations, documenting substantial consistency "…in the use of the accepted scientific framework but little consistency in the use of identifiable alternative frameworks" during comparison of responses to evolutionary prompts (p. 490). Similarly, Settlage and Jensen (1996) found that parallel items elicited considerably different response patterns, suggesting that the item context influenced participant responses. More recently, Nehm and Reilly (2007) explored how CR item sets about evolutionary change produced significantly different elicitation patterns for key concepts (KCs) and NIs about natural selection and documented that student knowledge and NIs about evolutionary change vary greatly according to the specific contexts in which they are assessed (Nehm and Ha 2011). For example, isomorphic CR items that differ only in the subject feature (e.g., plant vs. animal) or the familiarity of the subject feature (e.g., penguin vs. prosimian) have been shown to produce markedly different measures of both students' evolutionary knowledge and their NIs or misconceptions (Nehm and Ha 2011; Nehm et al. 2012; Opfer et al. 2012). In addition, some students will correctly explain the evolutionary gain of traits using a variety of KCs, while seldom mentioning these same concepts when explaining the evolutionary loss or decline of traits. These studies indicate that assessment of students' understanding of one type of evolutionary change (i.e., one item context) is often a poor predictor of their understanding of another type.

## Research Questions

Although the documentation of various item feature effects on reasoning using CRIs has resulted in important advances in the measurement of students' evolutionary knowledge frameworks, many unanswered questions remain about biases intrinsic to constructed-response assessment. This paper presents the results of three independent studies that explored whether and how the *sequencing* of items and their corresponding *features* impacted measures of explanation performance on a CRI. While the research presented in this paper is framed

within the context of students' *evolutionary explanations*, the primary focus is that of item sequencing, an aspect of assessment that has been largely unaddressed in the science education literature.

Specifically, the research was motivated by the following questions:

(1) To what extent does item position impact measures of student understanding (i.e., response accuracy) on CR explanation tasks?
(2) How do different item features differentially affect measures of students' explanatory practice?
(3) How does the magnitude of sequencing effects relate to the specific item features of the assessment?

## The ACORNS Constructed Response Instrument

In order to explore whether item sequencing and/or item features affect measures of student performance, we needed an instrument that had previously been evaluated for validity and reliability. While there are many published instruments available, few contain CR items that have been rigorously evaluated. Therefore, we chose the ACORNS CRI (Nehm et al. 2012), a previously published and evaluated instrument, to determine whether CR assessments may be susceptible to similar biases that are noted for MC assessments.

The ACORNS is a short answer, diagnostic instrument built on the work of Bishop and Anderson (1990) to assess student reasoning about the construct of natural selection. The instrument consists of an open-ended, isomorphic framework that prompts: "How would biologists explain how a living X species *with/without* Y evolved from an ancestral X species *without/with* Y?" The isomorphic nature of this instrument is of central importance for its use in the research presented here, as it allows for the construction of multiple items that are conceptually identical while differing in the specific scenarios presented (i.e., *X* and *Y* represent variable surface features). Thus, it was possible to construct item prompts and sequences suited to the three research questions mentioned above. In addition, the ACORNS is also one of the few CRIs in science education that has multiple published sources of evidence regarding the validity and reliability of inferences derived from measures of student explanations (e.g., Nehm and Schonfeld 2008; Nehm 2010; Nehm et al. 2011, 2012; Opfer et al. 2012).

## Research Methods

This paper used a mixed-methods approach to identify sources of and solutions to assessment bias for constructed-response items. Prior research on aspects of assessment structure has noted that the type of assessment (Gray 2004), sequencing of items (MacNicol 1956; Mollenkopf 1950; Monk and Stallings 1970; Sax and Cromack 1966), and item features (Nehm et al. 2012; Opfer et al. 2012) influence measures of student performance on a particular task. However, as noted in the introduction, the majority of research on assessment structure and its corresponding impact on student evaluation has been largely based in multiple-choice assessment formats (Leary and Dorans 1985). This study explored the relationship between two components of CR assessment structure—item sequencing and item features—and how they influenced measures of student performance on explanation tasks.

## Study Participants

To address our research questions, explanations of evolutionary change were gathered from a large sample of undergraduate students enrolled in introductory biology courses for majors and for non-majors at a large, public, Midwestern research university. Table 1 provides a breakdown of the participant samples used for each assessment. It is important to note that the exposure to evolutionary knowledge varied between the levels of introductory biology. Evolution was posited as an underlying theme in the teaching of the non-majors biology courses, whereas evolution was identified as one of four major themes (i.e., learning outcomes) for the majors-level biology courses. This variation in the teaching and exposure to evolutionary knowledge may have been an additional influence on measures of explanation performance and is discussed further in "Study Limitations" section.

Population demographics for each sample were generally representative of the larger student body of the university. Of the study samples, 58 % (non-majors level biology courses) and 55 % (majors level biology courses) were female (university enrollment: 48 %), and the average reported age was 19.7 and 20.5 years, respectively. Ethnicity was not reported for all samples, however the majority (79 %) of non-majors biology participants identified as non-Hispanic whites (university enrollment: 70 %).

## Study Design

To fully address our research questions, three independent assessments were administered in order to examine individual aspects of item sequencing and item features. Each assessment was presented electronically to participants (independent samples),[1] with items presented one at a time. Participants were not provided with information on how their responses would be scored and the instructions were open-ended ("please respond to the best of your ability"). Response time was not limited, and the median time for completion of an item was 1.9 min, with 90 % of the response times between 0.6 and 9.4 min, while the median number of words per explanation was 30 with 90 % of the responses containing between 10 and 74 words. Prior research has supported that the ACORNS scoring modules are limited in their capacity to identify student reasoning when response length is short (e.g., fewer than five words; Authors, unpublished data). For this reason, our analyses only included explanations from individuals whose responses to each item contained a minimum of five words.

Each of the three assessments employed a sequentially counterbalanced design (4×4 Latin Square) in order to evaluate potential effects of item location on sample level performance (e.g., Holland and Dorans 2006). The defining principle of a Latin Square task is that each item or task can appear only once in each row and once in each column of the ordered sequence (see Table 2). Participants in each sample were randomly assigned to one of four possible item sequences, each of which consisted of four isomorphic CR items that varied with respect to surface features. Item features were varied across the three assessments to evaluate effects on individual-level performance (Table 3). Three levels of item features are examined in this paper: item familiarity (i.e., familiar vs. unfamiliar), taxa type (i.e., animal vs. plant), and trait polarity (i.e., direction of change, gain vs. loss). While altering the item features within the assessment allowed for the presentation of different evolutionary change scenarios, the

[1] While the administration of three independent assessment tasks to three different student cohorts is a limitation of this study (see Study Limitations for a detailed discussion), we argue that the isomorphic nature of the assessment items allows for comparison across participant samples.

**Table 1** Participant demographics for each version of the ACORNS assessment

| Assessment | Course enrollment | Items/task | Sample (×Items) |
|---|---|---|---|
| 1 | Introductory biology (2 courses, non-majors) | 4-item | $N$=309 ($n$=1,236) |
| 2 | Introductory biology (1 course; majors) | 4-item | $N$=262 ($n$=1048) |
| 3 | Introductory biology (1 course; majors) | 4-item | $N$=157 ($n$=628) |

isomorphic nature of the ACORNS instrument provided conceptual continuity within and across item sequences.

The similarity of item features was established using PageRank (Page et al. 1998), a central component of the Google search engine highly useful for estimating the frequency with which individuals might normally encounter specific text (Griffiths et al. 2007). PageRank values therefore serve as a proxy for participants' familiarity *between* plant and animal taxa (e.g., snails and elms are more similar and PageRank than fish and elms) and *within* taxa (e.g., elms has a lower PageRanks value than labiatae) (Nehm et al. 2012; Opfer et al. 2012).

## Response Scoring

Prior research on students' evolutionary reasoning has documented the variety of cognitive elements that are used when constructing evolutionary explanations (Nehm 2010). Given the centrality of causality to scientific explanation (see above), each written explanation was quantified by tabulating the frequency of normative and non-normative causal elements as outlined by the ACORNS scoring rubric (Nehm et al. 2010) (Table 4). Normative scientific elements included seven KCs of natural selection and six NIs about natural selection widely discussed in the evolution education literature (Table 5). In addition to KC scores, we tallied the number of *different* KCs used by an individual *across* the item sequence, which refers to our composite measure of KC diversity (KCD) (for more details, see Nehm and Reilly 2007). All explanations were independently scored by a minimum of two expert raters who demonstrated high inter-rater reliability (kappa coefficients >0.8). In cases of disagreement between raters, consensus was established prior to data analysis.

## Data Analysis

Statistical analyses of the results from the three studies were performed in PASW (SPSS, Inc.) and JMP (SAS, Inc.). Univariate comparisons were made using the Wilcoxon signed rank test (pairwise comparisons); relationships with ordinal variables were made using Kendall's tau, or in the case of quantitative variables using Spearman's rho. Quantitative variables, such as

**Table 2** Simple Latin square design

| | Position 1 | Position 2 | Position 3 | Position 4 |
|---|---|---|---|---|
| Sequence 1 | 1 | 2 | 3 | 4 |
| Sequence 2 | 2 | 3 | 4 | 1 |
| Sequence 3 | 3 | 4 | 1 | 2 |
| Sequence 4 | 4 | 1 | 2 | 3 |

Each item can only appear once in each row and once in each column, therefore the number of item sequences is dependent on the number of items

**Table 3** Item feature categories for each of the ACORNS assessments examined in this study

| ACORNS version | Taxon type | Trait polarity | Familiarity |
|---|---|---|---|
| Assessment 1<br>*Item Position* | Animal | Gain, Gaining, Losing, OR Loss[a] | Familiar |
| Assessment 2<br>*Familiarity (feature)* | Animal | Gain | Familiar |
| | | | Unfamiliar |
| | Plant | | Familiar |
| | | | Unfamiliar |
| Assessment 3<br>*Polarity (feature)* | Animal | Gain | Familiar |
| | | Loss | |
| | Plant | Gain | |
| | | Loss | |

Items within each assessment were ordered using a counterbalanced design (Table 2)

[a] Participants responding to Assessment 1 were presented with only one of the four trait polarities, therefore the features of item sequence in this assessment remained constant across all four explanation tasks

verbosity measures, that more closely followed a normal distribution were analyzed using parametric methods. Thus, analyses that controlled for potential confounders, or that examined specific interactions with item sequencing, were based on a mixed effects repeated measures model. For these analyses, model assumptions were examined using residual plots and the sensitivity of conclusions to parametric assumptions were examined with permutation tests.

## Results

The results below are organized into three sections corresponding with the three different assessments evaluated in this paper. In each section, we review the item sequences and item features used to explore each source of potential bias for CR assessment, followed by the results for each assessment.

### Assessment 1: Item Position Effects in Constructed Response Assessment

The first assessment focused on documenting whether or not item-sequencing effects might occur for constructed response explanation tasks. In order to investigate the effect of item position on response scores, participants responded to a four-item sequence prompting them to explain evolutionary change. Importantly, as this assessment was simply to identify the effects of item sequencing, item features within a sequence were held constant across the four versions of the assessment task. Specifically, all items in the presented sequence were *familiar* (for details on familiarity ranking, see Nehm et al. 2012; Opfer et al. 2012) and asked about evolutionary change in *animal* taxa. In addition, students responded to only one type of trait *polarity* (e.g., gain, gaining, losing, or loss) in all four items. Therefore, we had 16 different treatments, with random assignment to an item sequence (i.e., order of item presentation) and trait polarity (see Table 3), resulting in a total of 1,236 evolutionary explanations for analysis. Importantly, this design allowed us to compare whether effects of item position varied in accordance with the features of the individual item sequences.

Analysis of student response scores, as measured by the total frequency of KCs and NIs revealed a sizeable effect of item position on overall measures of student performance on an

**Table 4** Sample scoring of student explanations

| Concept type | Concept scored | Explanatory element | Concept definition | Example of explanation illustrating each concept |
|---|---|---|---|---|
| Key concepts | Variation | Normative causal idea | The presence and causes of variation (i.e., mutation, recombination, sex) | *An ancestral labiatae plant's DNA mutated, giving it pulegone. This mutation did not harm the species, rather it made it equally or more successful from the labiatae without pulegones*. This allowed the mutation to be passed on via reproduction and eventually formed a new species. (Student #24207) |
| | Differential survival/ reproduction | Normative causal idea | The differential reproduction and/or survival of individuals | An ancestral labiatae species had a mutation that caused a pulegone. *This pulegone was more beneficial to survival and reproduction in their environment than no pulegone. The labiatae with a pulegone produced more offspring and lived longer*, which kept their genes in the species. (Student # 23834) |
| Naïve ideas | Teleology | Non-normative causal idea | Goal or "need" of trait causes it to exist; no mention of variation in trait existing prior to "need" | The plant may have evolved from a species which did not have pulegone because *it needed to attract certain insects*. The plant could have developed some way to produce *pulegone in order to attract certain animals or insects to spread its seeds and reproduce*. (Student # 23969) |
| | Intentionality | Non-normative causal idea | Events are directed by a mental agent | The winged seeds with the help of wind create dispersal. The ancestral elm species adapted to their environment and *realized that* in order to produce more and more elm plants through dispersal, their seeds must be winged in order to disperse better. (Student # 23856) |

*Italicized text* corresponds with the identified key concept or naïve idea (for more detailed scoring descriptions and examples, see Nehm and Schonfeld 2008; Nehm 2010; Nehm et al. 2012; Opfer et al. 2012)

Adapted from (Nehm et al. 2012; Opfer et al. 2012), Table 3

**Table 5** Normative and non-normative causal elements scored for in students' evolutionary explanations

| Key concepts (normative causal) | | Naïve ideas (non-normative causal) | |
|---|---|---|---|
| KC1 | Causes of variation | NI1 | Need as a goal (teleology) |
| KC2 | Heritability of variation | NI2 | Use and disuse |
| KC3 | Competition | NI3 | Intentionality |
| KC4 | Biotic potential | NI4 | Adapt (acquired traits) |
| KC5 | Limited resources | NI5 | Energy reallocation |
| KC6 | Differential survival | NI6 | Pressure |
| KC7 | Change over time | | |

explanation task (Fig. 1a). While student explanations largely included more scientifically normative elements (i.e., KCs) than non-normative elements (i.e., NIs), pairwise comparisons of responses scores indicated that KC use decreased significantly across the item sequence (Item 1→Item 4; Wilcoxon signed rank test, $p<0.001$). KC use tended to be highest for items in Position 1 of the sequence, relative to other item positions, with more than 36 % ($n=112$) of students having more KCs. In addition, response scores for items located at the start of the sequence (i.e., 1 and 2) were consistently higher relative to those at the end of the sequence (i.e., 3 and 4). However, no differences were found between measures of performance on items at the end of the sequence (Item 3→Item 4; Wilcoxon signed rank test, $p>0.86$). Despite overall declines in KC frequencies across item sequences, the diversity of KCs (KCDs) remained consistent, with the majority of students using only two (27.1 %) or three (28.8 %) different KCs across their four explanations. However, KCD and total KC use were strongly correlated, with higher diversity corresponding with higher total KC use (Spearman's rank correlation, $r=0.81$).

In contrast to the observed patterns for KCs, use of NIs did not appear to be impacted by item position. Pairwise comparisons of responses scores revealed no significant differences in the use of NIs (Item 1→Item 4; Wilcoxon signed rank test, $p=0.36$) (Fig. 1a). While students who incorporated more NIs across their four explanations were more likely to incorporate a higher diversity of NIs (Spearman's rank correlation, $r=0.80$), the majority of students (81.8 %) incorporated one or no NIs in their evolutionary explanations.

Analysis of item sequencing effects for each of the four item feature groups provided insight into the overall trends discussed above (Fig. 1b–e). Student responses to item sequences about the "gaining" or "losing" of a trait appear to be driving the item position effects for the overall sample, with KC use decreasing significantly with item position (Item 1→Item 4; Wilcoxon signed rank test, Gaining: $p<0.001$; Losing: $p<0.05$). KC use in response to "loss" items also decreased significantly with item position (Item 1→Item 4; Wilcoxon signed rank test, $p<0.001$), despite the observed variability in responses to the "fish" item. In addition, the magnitude of sequencing effects appeared to be greatest for student performance on items about trait loss, with use of scientifically normative elements decreasing across the four explanations for more than 37 % of students.

In addition to significant effects of item position on the elicitation of scientifically normative and non-normative elements, response verbosity (number of words per explanation) significantly declined across item sequences (Item 1→Item 4; Wilcoxon signed rank test, $p<0.0001$). Explanations for the first item in a sequence tended to be more verbose relative to explanations for the fourth item (Table 6). On average, explanations for the first item were 40.4 words long and decreased to 28.2 words for the fourth item. Explanations about the evolutionary gaining of a trait demonstrated the greatest change across the item sequence, from an

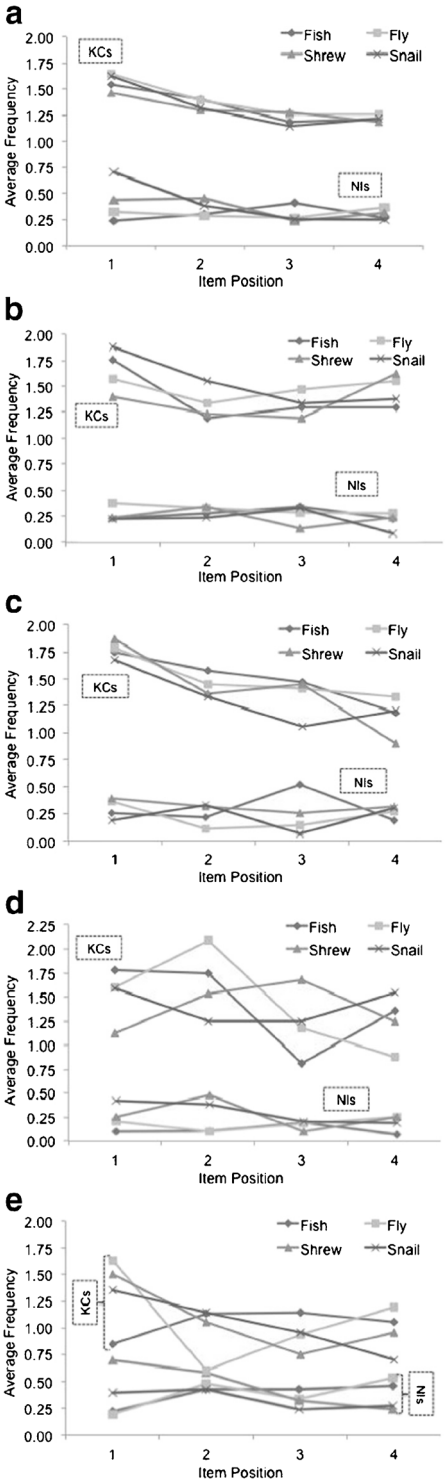


**Fig. 1** Item position effects on measures of scientifically normative (KCs) and non-normative (NIs) elements in students' evolutionary explanations. Each *graph* represents the average use of KCs and NIs across: **a** all feature (*trait polarity*) groups; **b** Gain; **c** Gaining; **d** Losing; **e** Loss item sequences

**Table 6** Average (±SEM) response verbosity across ACORNS item sequences

| | | Item 1 | Item 2 | Item 3 | Item 4 |
|---|---|---|---|---|---|
| Assessment 1 | Overall Verbosity | 40.43±1.22 | 34.14±1.05 | 30.03±1.64 | 28.27±1.59 |
| | Gain | 38.83±2.07 | 35.00±2.26 | 33.34±5.50 | 31.85±5.52 |
| | Gaining | 42.95±2.62 | 33.00±2.19 | 28.74±1.99 | 25.16±1.57 |
| | Losing | 41.04±2.48 | 36.39±1.98 | 31.24±2.26 | 30.13±2.18 |
| | Loss | 38.34±2.62 | 32.87±2.01 | 26.81±1.70 | 26.06±1.46 |
| Assessment 2 | Overall Verbosity | 35.79±4.18 | 32.22±4.00 | 30.65±3.81 | 32.88±4.11 |
| | F(A) | 34.58±4.08 | 32.80±3.81 | 31.36±4.08 | 42.25±5.60 |
| | F(P) | 39.44±5.13 | 39.68±5.26 | 35.71±4.21 | 32.08±3.73 |
| | U(A) | 32.92±2.72 | 32.67±3.85 | 29.72±3.94 | 30.66±3.99 |
| | U(P) | 36.23±4.80 | 23.71±3.09 | 25.80±3.00 | 26.51±3.12 |
| Assessment 3 | Overall Verbosity | 41.30±6.61 | 36.26±5.81 | 36.75±5.88 | 32.33±5.18 |
| | G(A) | 47.64±7.63 | 35.44±5.67 | 30.36±4.86 | 33.92±5.43 |
| | G(P) | 35.72±5.72 | 31.51±5.05 | 35.59±5.70 | 32.87±5.26 |
| | L(A) | 46.03±7.37 | 45.05±7.21 | 41.10±6.58 | 33.56±5.37 |
| | L(P) | 35.82±5.74 | 33.05±5.29 | 36.38±5.83 | 32.51±5.21 |

Data represents the response verbosity for items in each position in an item sequence. Assessment 2: *F(P)* familiar plant (elm/winged seeds); *U/(P)* unfamiliar plant (labiatae/pulegone); *F(A)* familiar animal (snail/poison); *U(A)* unfamiliar animal (suricata/pollex). Assessment 3: *G(A)* trait gain, animal (snail/teeth); *G(P)* trait gain, plant (grape/tendrils); *L(A)* trait loss, animal (mouse/claws); *L(P)* trait loss, plant (lily/petals)

average of 42.9 words for the first item to 25.1 words for the fourth item. Response verbosity was also significantly related to the use of KCs (Spearman's rank correlation, $r$=0.63) and KCD scores ($r$=0.505), with corresponding increases in verbosity with the addition of KCs. In contrast, student use of NIs in their explanations was not significantly associated with response verbosity ($r$=0.10).

## Assessment 2: Differential Effect of Item Familiarity on Measures of Explanatory Practice

The second assessment explored the role of item familiarity, and its interaction with item sequencing, for constructed-response explanation tasks. In order to investigate the potential effects of variable familiarity on response scores, participants responded to a four-item sequence prompting them to explain *familiar* and *unfamiliar* examples of evolutionary change. Importantly, all of the items in the presented sequences were of the same *polarity* (gain). Therefore, for analysis we had four different treatments, with random assignment to an item sequence (see Table 3), resulting in a total of 1,048 evolutionary explanations for analysis.

Analysis of response scores found no significant effect of item position on explanations for three out of the four item feature combinations presented in the sequence (Fig. 2a). While explanations for all item prompts tended to include more scientifically normative elements than non-normative elements, responses to the *familiar*/*animal* item contained significantly more scientifically normative elements, on average, relative to responses to the *familiar*/*plant* item and all *unfamiliar* items. In addition, pairwise comparisons for the *familiar*/*animal* item indicated that KC use decreased significantly with item position (Item 1→Item 4; Wilcoxon signed rank test, $p$<0.015). In contrast, KC use for the other items tended to increase, although not significantly (Item 1→Item 4; Wilcoxon signed rank test, $p$=0.069).

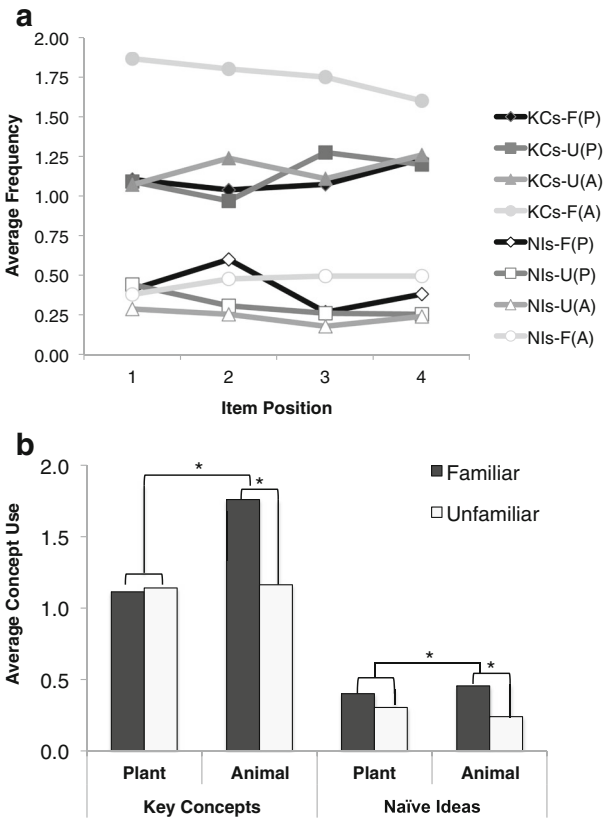


**Fig. 2** Explaining *familiar* and *unfamiliar* evolutionary change: **a** Item position effects and **b** item feature effects. Response scores consisted of the average frequency of key concepts (*KCs*) and naïve ideas (NIs) for each of four taxa/trait combinations: *F(P)* familiar plant (elm/winged seeds), *U(P)* unfamiliar plant (labiatae/pulegone), *F(A)* familiar animal (snail/poison), *U(A)* unfamiliar animal (suricata/pollex); *$p<0.05$

Along with the relative stability of KC use across the item sequence, the diversity of KCs was relatively consistent between the different item sequences. However, KCD was much lower in responses to this item set compared to the items used in Assessment 1, suggesting the familiarity of items plays a role in the elicitation of scientifically normative elements in students' evolutionary explanations. The largest group of students used one or fewer different KCs (34.5 %) in their responses, and only 9.6 % of students utilized more than four different KCs across their explanation sequence. However, KCD scores and total KC use were significantly correlated, with higher diversity corresponding with higher total KC use (Spearman's rank correlation, $r=0.64$). Corresponding with the observed patterns for KCs, student use of NIs in their explanations did not appear to be impacted by item position (Item 1→Item 4; Wilcoxon signed rank test, $p=0.48$). In addition, while the majority of students used one or fewer NIs per explanation (72.1 %), the diversity of NIs was not significantly related to the frequency of NIs in student responses.

Analysis of item familiarity, independent of item position, identified significant effects of the item feature on measures of student performance (Fig. 2b), with explanations about evolutionary change in *familiar* taxa containing more KCs, as well as almost 40 % more NIs ($F[1, 1046]=17.717$, $p<0.0001$). Similarly, the taxa of the item appeared to influence response scores, with

items asking students to explain evolutionary change in animals containing, on average, more KCs than those about plants. Still, the majority of student explanations (58.4 %) only contained 1–2 KCs, regardless of item familiarity. In addition, responses to *familiar* item contained more NIs relative to *unfamiliar* items, with 37.1 % of explanations about *familiar* evolutionary change scenarios containing 1–2 NIs compared to only 23.9 % for *unfamiliar* scenarios.

As indicated above, KCD of responses to this version of the assessment was lower than that of responses to the first version of the assessment (discussed above). However, examination of diversity scores for feature groups found that KCD scores for explanations about evolutionary change in *familiar* items was, on average, 0.5 KCs greater than KCD scores for explanations of *unfamiliar* items ($t[533]=4.412$, $p<0.001$). Similar results were not found when comparing diversity scores for explanations of evolutionary change in *plants* versus *animal*s, controlling for familiarity.

Analysis of the interaction between item feature effects and item position effects for this assessment revealed no significant effects on measures of scientifically normative elements in student explanations. However, significant effects were found for measures of non-normative elements ($F[1, 1046],=7.505$, $p<0.0006$) in student responses. While inclusion of non-normative ideas was most prevalent in Positions 1 and 2 of the item sequence, NIs tended to increase across the item sequence for *familiar* items (i.e., a *familiar* item in Position 3 tended to have more NIs compared to a *familiar* item in Position 2).

Finally, while results of Assessment 1 indicated that response length was significantly associated with item position, the results of the second assessment highlight the role of item features for response verbosity (Table 6). Specifically, explanations about *familiar* taxa were significantly more verbose than responses about *unfamiliar* taxa, however differences in response lengths were mitigated by the interaction between familiarity and taxon type. In addition, responses that were more verbose were significantly more likely to contain scientifically normative elements (Spearman's rank correlation, $r=0.44$, $p<0.01$) and have higher KCD ($r=0.43$, $p<0.01$), thereby increasing the explanation response score. However, response verbosity did not appear to be related to the use of non-normative elements ($r=0.06$), indicating that as students write longer explanations they are not also more likely to incorporate additional NIs about evolutionary change.

### Assessment 3: Differential Effect of Item Polarity on Measures of Explanatory Practice

The third assessment explored the role if item polarity, and its interaction with item position, for constructed-response explanation tasks. Similar to the previous assessment, participants responded to a four-item sequence prompting them to explain the evolutionary *gain* or *loss* of a trait in *familiar* taxa (no *unfamiliar* items were presented in this assessment). Therefore, for analysis we had four different treatments, with random assignment to an item sequence, resulting in a total of 628 evolutionary explanations for analysis.

Analysis of student response scores revealed no significant differences in the use of scientifically normative or non-normative elements in relation to item position within the sequence (Fig. 3a). While the use of KCs was relatively consistent across an item sequences, with the majority of explanations (59.1 %) containing 1–2 KCs on average, more than 20 % of students never used a KC in their explanations and more than 33 % of students incorporated only 1–2 NIs in their responses. However, explanations about the evolutionary *gain* of a trait contained significantly more KCs ($F[1,622=7.906$, $p<0.005$) and fewer NIs ($F[1,622]=12.293$, $p<0.001$) than explanations about the *loss* of a trait, regardless of item taxon (Fig. 3b).

Along with the differences in KC use across item features, the diversity of KCs was significantly higher for *gain* items relative to *loss* items within a sequence ($t[310]=2.307$, $p<0.04$). Student responses averaged 1.56 KCs in response to *gain* item prompts, with 49 % of

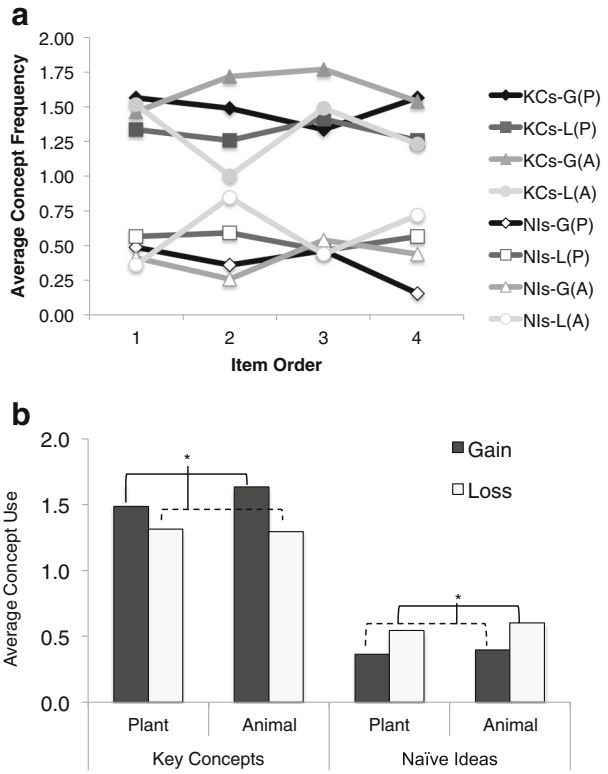


**Fig. 3** Explaining evolutionary *gain* and *loss* of traits: **a** item position effects and **b** item feature effects. Response scores consisted of the average frequency of key concepts (*KCs*) and naïve ideas (*NIs*) for each of four taxa/trait combinations: *G(A)* trait gain, animal (snail/teeth); *G(P)* trait gain, plant (grape/tendrils); *L(A)* trait loss, animal (mouse/claws); *L(P)* trait loss, plant (lily/petals), $*p<0.05$

these responses containing 2–4 KCs, whereas 62.5 % of responses to *loss* item prompts contained 0–1 KCs and 43 % of student explanations about trait *loss* contained 1–2 NIs. In addition, explanations of evolutionary change in *animals* tended to have higher KCD and to incorporate more NIs relative to *plant* item prompts. Similarly, while analysis of the interaction between item position and item features revealed no significant effects on measures of scientifically normative ideas, non-normative elements tended to increase across an item sequence, especially for explanations of trait *loss*.

Lastly, as with Assessment 2, the results of this assessment highlight the effects of item features on response verbosity. Despite no differences in response accuracy across the item sequence (as measured by the frequency of KCs and NIs), response verbosity significantly decreased with item position across the sequence (Item 1 → Item 4; Wilcoxon signed rank test, $p<0.004$) by an average of 8.7 words (Table 6). In addition, while explanations about *animal* taxa contained significantly more words than *plant* taxa ($t[622]=-2.449$, $p<0.002$), differences in response lengths were mitigated by the interaction between the two feature categories (i.e., taxon type and trait polarity), as verbosity did not significantly differ between items about the *gain* or *loss* of a trait. Responses that were more verbose, regardless of feature categories, were significantly more likely to contain scientifically normative elements (Spearman's rank correlation, $r=0.50$), but not the use of non-normative elements (Spearman's rank correlation, $r=0.07$).

## Summary

Together, the results from this study provide significant insight into the effects of item surface features and item sequencing for constructed response assessment and the evaluation of student understanding through performance on scientific practices like explaining. Analyses revealed that while there are a variety of factors that influence undergraduate biology students' performances on explanation tasks, measures of performance do differ in accordance with item position within an assessment sequence (Table 7). This is in concordance with previous research on measures of student performance on multiple-choice tasks in a variety of disciplines, and highlights the need for research on the assessment of scientific practices. Importantly, factors such as the specific item (i.e., features like familiarity or the polarity of evolutionary change) appear to mitigate effects of item sequencing for CR assessment, suggesting that educators and researchers need to consider both the structure of individual items and the structure of the overall instrument when assessing student performance on explanation tasks.

**Table 7** Summary of results for each ACORNS assessment

| ACORNS Study foci | Assessment 1 Item position | Assessment 1 Familiarity (feature) | Assessment 3 Polarity (feature) |
|---|---|---|---|
| Results: Item Position | | | |
| KCs | 1st Item>4th Item** | 1st Item>4th Item for Snail** | No difference across items |
| KCD | No difference between item sequences | No difference between item sequences | No difference between item sequences |
| NIs | No difference across items | No difference across items | No difference across items |
| Results: Item Features | | | |
| KCs | Gain=Gaining=Losing>Loss** | Familiar>Unfamiliar*; Animal>Plant** | Gain>Loss** |
| KCD | Gain=Gaining=Losing>Loss** | Familiar>Unfamiliar** | Gain>Loss* |
| NIs | Loss>Gain=Gaining=Losing** | Familiar>Unfamiliar** | Loss>Gain** |
| Feature Interaction Effects | | | |
| KCs | N/A | Familiar Animal>Familiar Plant=Unfamiliar Animal/Plant** | No difference across items |
| KCD | N/A | N/A | N/A |
| NIs | N/A | Familiar Animal>Familiar Plant=Unfamiliar Animal/Plant** | Animal Gain/Loss>Plant Gain/Loss** |
| Feature/Sequencing Interaction Effects | | | |
| KCs | N/A | No difference across items | No difference across items |
| KCD | N/A | N/A | N/A |
| NIs | N/A | 1st Item<4th Item for Familiar Animals/Plants** | 1st item<4th item for Loss in Animals/Plants** |
| Response Verbosity | | | |
| KCs | Increases with verbosity** | Increases with verbosity** | Increases with verbosity** |
| KCD | Increases with verbosity** | Increases with verbosity** | Increases with verbosity** |
| NIs | No difference across items | No difference across items | No difference across items |

*KCs* key concepts, *KCD* key concept diversity, *NIs* naïve ideas

*$p<0.05$; **$p<0.01$

## Discussion

As the use of assessment tasks that integrate scientific practices with core disciplinary concepts becomes more common in the classroom and in science education research, it becomes critical that we investigate the advantages and limitations of the assessment practices being utilized. While many issues of assessment structure have been widely discussed for MCIs (Leary and Dorans 1985), the comparative lack of discussion for CRIs is indicative of a large gap in assessment design research. With increasing emphasis on assessing student *performance* on tasks (NRC 2012; Pellegrino 2013), assessments asking for students to actively engage in a scientific practice (i.e., *construct* responses as opposed to *selecting* responses) need to be evaluated for the comparative risks and benefits of assessing student performance using different instruments. However, a brief review of recent research reveals that assessments utilizing constructed-response explanation items contain little to no empirical rationale for the structure of the instrument in terms of the number or arrangement of items and their features (e.g., Gotwals and Songer 2010; Lee et al. 2011; McNeill et al. 2006; Nehm and Reilly 2007; Peker and Wallace 2011; Songer et al. 2009). Despite the fact that such research covers a wide variety of item types, numbers of items, and overall instrument structure (i.e., arrangement of items), there is no indication that the intrinsic biases of overall instrument structure were considered. In the following sections, we situate our results within the larger context of assessment research and discuss the potential implications for CR instruments in biology education.

### Item Sequencing: Eliminating Assessment Bias with Counterbalanced Designs

Prior research on instrument biases for MC assessments identified several potential effects of item sequencing on measures of student performance, including differential performance in accordance with subject matter (Mollenkopf 1950), item difficulty (MacNicol 1956; Monk and Stallings 1970) and quantitative versus qualitative item types (Kingston and Dorans 1984). In spite of item sequencing being a well-documented assessment bias for MCIs (Leary and Dorans 1985), comparable research on instrument biases for CR assessments—which are inherently qualitative—has to date been, unaddressed.

As demonstrated by our first assessment, measures of student performance (i.e., frequency of scientifically normative and non-normative ideas, response verbosity) on a CRI can be susceptible to effects of item sequencing, with item-level analyses being particularly affected. This is particularly problematic for the development of new assessments for evaluating student understanding through performance on scientific practices, including explanation and argumentation (NRC 2012). However, the results from our set of assessments identified two potential solutions to CR item sequencing effects: counterbalanced designs and variable surface features (discussed in the next section). The use of counterbalanced designs can mitigate the effects of item location across treatments and provide more valid measures of item-level performance across a sample (Holland and Dorans 2006); however, this approach must be weighed relative to the other goals of instrument design.

### Item Features: Using Variable Features to Mitigate Declining Responses

Item surface features are often cited when referring to differences in expert and novice problem representation (e.g., Chi et al. 1981). While this can be particularly problematic if the surface features are perceived as different, our current study demonstrates that variable item features can also be particularly beneficial. Features such as subject and subject familiarity have been shown to produce markedly different measures of student performance (and confidence in)

response accuracy—for both MCIs and CRIs (e.g., Caleon and Subramaniam 2010; Nehm et al. 2012; Rodrigues et al. 2010). In addition, the assessments presented in this paper suggest that while students appeared to respond differently to isomorphic CR items that contained variable surface features, the response differences resulted in a *better* measure of individual student performance (as opposed to sample-level measures). Student responses to our variable CR item sequences (i.e., Assessments 2 and 3) contained more diverse sets of scientifically normative concepts compared to those responding to CR item sequences with similar features (i.e., Assessment 1).

While prior assessment research has documented the effects of a variety of factors at the item level, issues at the instrument level, such as item sequencing, have not been investigated for CR assessments. The results of this study emphasize the importance of considering the potential biases of *both* individual CR items and overall CR instrument structure during classroom and large-scale assessment. Sequences consisting of more variable item features significantly influenced students' response accuracy across CR items, suggesting that variable features serve to mitigate the observed effects of item sequencing observed with Assessment 1. Overall, our study indicates that measures of student explanation quality across multiple items may be less susceptible to item sequencing biases than measures that examine responses to individual items in isolation.

Of additional interest for the design and evaluation of CR assessments is that students' use of scientifically normative concepts was directly related to the particular item features of the sequence, independent of item order. Items with "easier" features tended to elicit more KCs than those with more "difficult" features. For example, the *familiarity* of the item features contributed significantly to the use of both normative and non-normative concepts in student responses. This raises significant questions about the use of *familiar* (and *unfamiliar*) item contexts in classroom assessments and high-stakes testing. Measures of student performance on items with novel or unfamiliar features may under-represent students' understanding of the construct of interest (e.g., the core idea of evolution; NRC 2012).

### Verbosity Impacts all Explanation Performance Measures for CRIs

The relationship between response verbosity and student performance is perhaps not surprising when measuring student performance on CR items. However, higher student performance was driven solely by higher frequencies of scientifically normative ideas, *not* lower frequencies of NIs, which remained relatively stable across item sequences. This suggests that students simply tend to be more verbose at the beginning of a CR assessment, perhaps due to the nature of responding to isomorphic items; however, our results are somewhat inconsistent across studies (see also Rector et al. 2012). In particular, changes in verbosity were susceptible to differences in item features, such as *familiarity*, suggesting that the surface features of an item are potentially more influential on response processes than the sequencing of the items in the assessment task.

The purpose of this study was to investigate the relationship between factors known to influence measures of student concept recognition using MC tests on measures of student recall using CRIs (specifically, a widely used written explanation assessment). The three assessments presented in this paper serve to highlight the interactions among these assessment features and their effects on measures of student performance on CR tasks. Overall, our results have identified several potential limitations of CR assessments that should be considered when evaluating student performance and recall of core ideas (cf. NRC 2012). Educators and researchers need to carefully consider the effects of item sequencing and item features on student responses, in particular on response verbosity, when using a CR assessment. In the sections that follow, we discuss some limitations and the relevance of our studies to the general

science education research community, and offer suggestions for the direction of future research on assessing scientific practices.

## Study Limitations

One overarching limitation of our work is that the students in our samples had different exposures to biology instruction, and in particular to evolutionary content, and therefore the results for each assessment are not directly comparable. Assessment 1 measured the explanatory practice and recall of core concepts in evolution by biology non-majors enrolled in courses where evolution was posited as a main theme. In addition, these students wrote a paper on evolution as a course assignment, which may have enhanced their understanding of evolutionary processes and their ability to write explanations about evolutionary change. In contrast, Assessments 2 and 3 measured the explanatory practice and knowledge recall of introductory biology majors enrolled in their first course series containing evolution instruction as a major theme (one of four main learning outcomes for the course), which is generally more advanced than the non-major course. Surprisingly, our results indicated that students responding to Assessment 1 (non-majors) incorporated slightly more KCs into their responses than students responding to Assessments 2 or 3 (majors).

We suggest that the similarity of item features in Assessment 1 is the most parsimonious explanation for the differences in KC use between the three assessments. However, there are other factors that could explain the differences in measured performance, such as the amount or type of biological knowledge held by the participants, instructional goals of the course, or type of biological content. Research exploring the effects of variable surface features on item sequencing (i.e., Assessments 2 and 3) using samples drawn from populations of non-majors or advanced biology majors would provide greater insight into the role that prior knowledge and course instruction plays in explaining differences in performance measures. Future work is needed to expand the populations of study to include a broader spectrum of educational levels and experience with biology to determine whether the findings are generalizable.

A second limitation of our study is that differences in scientific reading and writing ability (e.g., English language learners [ELLs] vs. native speakers) were not explicitly taken into consideration. Measures of scientific understanding of ELLs that are dependent on the inclusion of key terms and phrases may not be representative of actual knowledge or recall of concepts. However, there are similar constraints on recall when using MCIs that incorporate large amounts of text (Martiniello 2008). Even so, despite our lack of consideration for differences in reading and writing ability, the items used in this study have been previously validated among diverse populations using clinical interviews and multiple-choice assessments, and shown to produce valid and reliable inferences (Nehm and Schonfeld 2008; Nehm et al. 2012). Future research might investigate tasks that allow participants to respond via other modes of communication (i.e., visual representations, graphs) important for assessment of students' scientific literacies.

A third limitation of our study is related to our scoring methodologies, which were atomistic rather than holistic (e.g., Songer et al. 2009). Correspondingly, our item sequencing results may be different if student responses were evaluated for overall quality rather than individual components. For example, while the use of total KC scores (per item) provides a measure of students' evolutionary knowledge, the use of KCD (across-item) scores prevents the comparison of sequencing patterns. While not within the scope of the current study, clinical interviews that examine item sequencing effects would provide additional insight into the accuracy of measures of students' evolutionary knowledge and explanatory practice. Despite the potential differences in scoring approaches, it remains clear that item surface features are an important factor when evaluating students' evolutionary explanations across an item sequence.

Finally, while the results from our three assessments offer clear implications for constructed response assessment in biology education (see Implications section), it is important to consider the potential application for other scientific domains. The nature of our study was to investigate how factors known to influence measures of students' biological understanding using MC tests influence measures of student biological understanding using a CR task. While item-sequencing effects using CR items have been relatively unexamined in the research literature, several recent studies in chemistry and physics education have examined how variation in surface features of isomorphic problems relates to student problem solving success (e.g., Gulacar and Fynewevr 2010; McClary and Talanquer 2011; Papadouris et al. 2008; Singh 2008). While it does not appear that surface features have been directly manipulated in these studies as they were in ours, item features did vary across problems in an assessment. Furthermore, research suggests that the transfer of knowledge across problem contexts is often inhibited by misconceptions (NIs) about scientific concepts, particularly if problems do not share surface features. Likewise, researchers have recognized the importance of assessment research that considers "the order in which questions were asked and the proximity of the paired questions" when examining whether or not students are able to appropriately transfer knowledge from one problem to the next (Singh 2008 p. 8.C).

The results of these other studies, and ours, support the need for further examination of student reasoning and knowledge construction across different item features when making inferences about student performance on constructed response assessments in domains other than evolution.

## Implications for Previous and Future Research

### Assessing Evolutionary Understanding

Recent advances in evolution education research have documented a variety of difficulties that students have when reasoning about evolutionary change. Much research has focused on identifying the types of concepts that students have difficulty with that hinder the development of evolutionary reasoning. For example, students have been shown to have difficulty with concepts that are fundamental to an understanding of evolution, such as common ancestry (Catley et al. 2013; White and Yamamoto 2011). Similarly, when students hold misconceptions about the role of random processes or adaptation, they tend to produce non-normative explanations of biological change (Nehm et al. 2012; Opfer et al. 2012; Garvin-Doxas and Klmkowsky 2008). Other areas of research have addressed the types and patterns of intuitive reasoning used when explaining evolutionary change. For example, the use of teleological, or agency-driven reasoning, is among the most prevalent in student explanations, even from a young age (Kelemen 2012). Although our study documented many of the same types of evolutionary ideas (both normative and non-normative or naïve) that have been uncovered in prior science education research, it adds a new perspective on why these ideas might occur in assessment responses.

Our results suggest that both item sequencing and the diversity of item features in an assessment influence the frequency with which normative ideas are elicited in students' constructed responses. While prior work on evolution assessment has investigated the types and magnitudes of scientifically normative (e.g., Bishop and Anderson 1990; Nehm and Reilly 2007) and non-normative ideas (e.g., Bishop and Anderson 1990; Clough and Wood-Robinson 1985; Nehm and Reilly 2007) in undergraduates, and the effects of item features on measures of student performance (e.g., Nehm and Schonfeld 2008; Nehm and Ha 2011), these studies

did not use a counterbalanced design to control for or consider the effects of item sequencing on their measures of student performance. However, declines in student performance (as measured by KCs) across item sequences were noted in some studies (e.g., Nehm and Reilly 2007). Importantly, the majority of prior research on student understanding of evolution, including by Nehm and Reilly), has used items with some degree of variation in surface features. Our results provide empirical support for the continued use of differential item features when measuring student performance across item sequences, as the impact of item order was greater in our studies when item features were very similar.

Another consideration for measurement of student performance using constructed response assessment is the manner in which student knowledge is quantified. In this study, we used both KC frequency and KCD as measures of student performance. Previous work on evolution assessment has argued that KCD provides a more accurate measure of student knowledge and understanding as it represents the number of different, scientifically normative ideas a student uses in response to a set of isomorphic CR items (e.g., Nehm and Reilly 2007). Our results suggest that *both* total KC frequency and KCD can be used to obtain accurate measures of student performance, but the method that provides the best measure of student knowledge depends upon the level of focus. When assessment practices are focused on quantifying overall performance, the *diversity* of concepts present (KCD) provides a more holistic measure of students' understanding of the construct. However, the consistent use or *frequency* of particularly KCs provides a better measure of individual student performance.

## Assessing Student Reasoning: Universal Implications

A potential concern for this and other item sequencing research is the manner in which the items are presented to students. For example, presenting students with items of the same format in sequential order may provide different measures of student knowledge than a mixed format assessment. The inclusion or interspersion of items of an alternate format (e.g., Item 1: CR, Item 2: MC, Item 3: CR) might also mitigate sequencing effects in constructed response assessments. Likewise, our results support the use of a counterbalanced design to moderate item sequencing effects. This was particularly important for CR items that are located at the end of the assessment, as student performance on these items may be subject to other measurement errors, such as errors of omission (i.e., leaving out previously stated information).

Use of a counterbalanced design approach can also facilitate the evaluation of population performance on specific items or tasks, independent of item location. Such designs are ubiquitous in the cognitive sciences (e.g., Birney et al. 2006; Jensen et al. 1999; Perret et al. 2011; Pollatsek and Well 1995); however, Latin Square designs such as those employed in our three studies are rare in science assessment research, particularly in the area of explanatory practice (Table 1). The results of our study and previous work on item sequencing effects suggest that order does matter for both multiple choice and some types of constructed response assessments. Importantly, the order of items with similar surface features can significantly affect student performance, highlighting the need to increase our understanding of item sequencing effects in different science domains.

In addition to addressing the issue of item sequencing effects in CR assessments, our results further clarify the role of item surface features in science assessment. There is much research discussing the ways in which novice problem solvers use surface features to identify and categorize problems (e.g., Chi et al. 1981). However, many commonly used assessment instruments were not developed with such features as familiarity and subject feature in mind. For example, the Conceptual Inventory of Natural Selection (CINS; Anderson et al. 2002), a

commonly used MC instrument in biology education research, *does* incorporate different taxon/trait combinations but *does not* vary with respect to trait polarity or item familiarity (it only includes relatively familiar animals in trait gain scenarios). Given that our results clearly indicate that familiarity and trait polarity play an important role in both the mitigation of item sequencing effects and overall measures of student performance, this raises concerns about the inferences that can be drawn from this and other similarly developed instruments.

Overall, the results of our study raise important questions about the design and implementation of assessments measuring student understanding through performance on scientific practices (Nehm et al. 2012; NRC 2012; Opfer et al. 2012). While the order of item presentation has not been widely examined, our results suggest that it is an important feature that should be considered in constructed-response assessments. Likewise, little research has been done to investigate how evaluations of performance change based on the number of explanatory tasks given. Assessments that seek to measure student performance on a particular construct may be biasing student responses based on the structure of the items. Future work on constructed response assessments will need to consider the effects of item sequencing and features on measures of student understanding through performance on scientific practices, such as explanation.

Given our current results, we recommend that classroom assessments include a diversity of problem features to ensure equivalent response verbosities. Likewise, the items should be only moderately isomorphic so as to minimize errors of omission or repetitive responses. Performance on items that are too similar, such as those in Assessment 1, may significantly bias estimates of student knowledge recall because of declining response verbosity. Increasing the diversity and familiarity of item features within an assessment, as in Assessments 2 and 3, provided a solution for decreasing responses, allowing for more a more accurate measure of individual student performance.

**Acknowledgments** This research was supported by the National Science Foundation REESE program (R.H. Nehm) and the Marilyn Ruth Hathaway Education Scholarship (M.R. Federer). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the NSF or The Ohio State University.

## References

American Association for the Advancement of Science [AAAS]. (1994). *Benchmarks for science literacy*. New York: Oxford University.

American Association for the Advancement of Science [AAAS]. (2011). *Vision and change in undergraduate biology education*. Washington, DC. http://visionandchange.org/.

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, 39*, 952–978.

Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berland, L. K., & McNeill, K. L. (2012). For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson. *Science Education, 96*(5), 808–813.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: the development of the Latin Square Task. *Educational & Psychological Measurement, 66*(1), 146–171.

Bishop, B., & Anderson, C. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27*, 415–427.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–271.

Caleon, I. S., & Subramaniam, R. (2010). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education, 40*, 313–337.

Catley, K. M., Phillips, B. C., & Novick, L. R. (2013). Snakes, eels, and dogs! Oh my! Evaluating high-school students' tree-thinking skills: an entry point to understanding evolution. *Research in Science Education, 43*(6), 2327–2348.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Clough, E. E., & Wood-Robinson, C. (1985). How secondary students interpret instances of biological adaptation. *Journal of Biological Education, 19*, 125–130.

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education, 70*(4), 473–496.

Cronbach, L. J. (1988). *Five perspectives on validity argument* (In H. Wainer and H.I. Braun (Eds)). Hillsdale, NJ: Lawrence Erlbaum.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: learning and teaching science in grades K-8*. Washington DC: National Academies.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*(1), 5–19.

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). *CBE Life Sciences Education, 7*(2), 227–233.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Gray, K. E. (2004). The effect of question order on student responses to multiple choice physics questions. Master thesis, Kansas State University. Retrieved from http://web.phys.ksu.edu/dissertations/

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. *Psychological Science, 18*, 1069–1067.

Gulacar, O., & Fynewevr, H. (2010). A research methodology for studying what makes some problems difficult to solve. *International Journal of Science Education, 32*(16), 2167–2184.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed, pp. 187–220). Westport: American Council on Higher Education and Praeger.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 135–175.

Jensen, P., Watanabe, H. K., & Richters, J. E. (1999). Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology, 27*(6), 439–445.

Kampourakis, K., & Zygzos, V. (2008). Students' intuitive explanations of the causes of homologies and adaptations. *Science & Education, 17*, 27–47.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of cross-validation on long-term retention. *European Journal of Cognitive Psychology, 19*, 528–558.

Kelemen, D. (2012). Teleological minds: how natural intuitions about agency and purpose influence learning about evolution. In K. S. Rosengren, S. K. Brem, E. M. Evans, & G. M. Sinatra (Eds.), *Evolution challenges: integrating research and practice in teaching and learning about evolution* (pp. 66–92). Oxford: Oxford University.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147–154.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*(4), 507–531.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387–413.

Lee, H.-S., Liu, L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115–136.

Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical papers* (Vol. 2, pp. 214–240). Oxford: Oxford University Press.

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment, 16*, 164–184.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test*. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory, 7*(2), 79–90.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218.

Martiniello, M. (2008). Language and the performance of English Language Learners in math word problems. *Harvard Educational Review, 78*, 333–368.
McClary, L., & Talanquer, V. (2011). College chemistry students' mental models of acids and acid strength. *Journal of Research in Science Teaching, 48*(4), 396–413.
McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*(2), 153–191.
Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741–749.
Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika, 15*, 291–315.
Monk, J. J., & Stallings, W. M. (1970). Effect of item order on test scores. *Journal of Educational Research, 63*, 463–465.
National Research Council. (1996). *National science education standards*. Washington, DC: National Academies.
National Research Council. (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: National Academies.
National Research Council. (2007). *Taking science to school: learning and teaching science in grades K-8*. Washington, DC: The National Academies.
National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies.
Nehm, R. H. (2010). Understanding undergraduates' problem-solving processes. *Journal of Biology and Microbiology Education, 11*(2), 119–121.
Nehm, R. H., & Reilly, L. (2007). Biology majors' knowledge and misconceptions of natural selection. *Bioscience, 57*(3), 263–272.
Nehm, R. H., & Schonfeld, I. (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and oral interview. *Journal of Research in Science Teaching, 45*(10), 1131–1160.
Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256.
Nehm, R. H., Ha, M., Rector, M., Opfer, J. F., Perrin, L., Ridgway, J., & Mollohan, K. (2010). Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (ACORNS). *Technical Report of National Science Foundation REESE Project, 0909999*. www.evolutionassessment.org.
Nehm, R. H., Ha, M., & Mayfield, E. (2011). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196.
Nehm, R. H., Beggrow, E., Opfer, J., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher, 74*(2), 92–98.
Opfer, J., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. *The Journal of Research in Science Teaching, 49*(6), 744–777.
Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: a necessary distinction? *Science Education, 95*, 627–638.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: bringing order to the web (Tech. Rep.)*. Stanford, CA: Stanford Digital Library Technologies Project.
Papadouris, N., Constantinou, C. P., & Kyratsi, T. (2008). Students' use of the energy model to account for changes in physical systems. *Journal of Research in Science Teaching, 45*, 444–469.
Peker, D., & Wallace, C. S. (2011). Characterizing high school students' written explanations in biology laboratories. *Research in Science Education, 41*, 169–191.
Pellegrino, J. W. (2013). Proficiency in science: assessment challenges and opportunities. *Science, 340*, 320–323.
Perret, P., Bailleux, C., & Dauvier, B. (2011). The influence of relational complexity and strategy selection on children's reasoning in the Latin Square Task. *Cognitive Development, 26*, 127–141.
Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: a suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(3), 785–794.
Popham, W. J. (2010). *Classroom assessment: what teachers need to know*. Pearson: Pearson Allyn & Bacon.
Rector, M., Nehm, R. H., & Pearl, D. (2012). Learning the language of evolution: lexical ambiguity and word meaning in student explanations. *Research in Science Education, 43*(3), 1107–1133.
Rodrigues, S., Taylor, N., Cameron, M., Syme-Smith, L., & Fortuna, C. (2010). Questioning chemistry: the role of level, familiarity, language, and taxonomy. *Science Education International, 21*(1), 31–46.
Rodriguez, M. C. (2003). Construct equivalence of multiple choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40*, 163–184.
Roediger, H. L., III. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology, 31*(5), 1155–1159.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: a framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world* (pp. 79–118). Princeton: University Press.

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement, 3*, 309–311.

Settlage, J., & Jensen, M. (1996). Investigating the inconsistencies in college student responses to natural selection test questions. *Electronic Journal of Science Education, 1*, 1.

Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science, 130*, 477–482.

Singh, C. (2008). Assessing student expertise in introductory physics with isomorphic problems: II. Effects of some potential factors on problem solving and transfer. *Physics Education Research*, 4(1), 010105-1–010105-10.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*(6), 610–631.

Strevens, M. (2004). 'Scientific explanation', in macmillan encyclopaedia of philosophy, (2nd ed.).

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). *A comparison of free-response and multiple-choice questions in the assessment of reading comprehension (RR 87–20)*. Princeton, N.J.: Educational Testing Service.

White, B. Y., & Frederickson, J. R. (1998). Inquiry, modeling, and metacognition: making science accessible to all students. *Cognition & Instruction, 16*, 3–118.

White, B. T., & Yamamoto, S. (2011). Freshman undergraduate biology students' difficulties with the concept of common ancestry. *Evolution: Education & Outreach, 4*(4), 680–687.