Contextual Constraints in Science Assessment: Effects of Item Sequencing and Item Features on Measures of Student Performance using a Constructed Response Instrument

Meghan A. Rector

School of Teaching and Learning
The Ohio State University

Introduction

Assessment of students' scientific knowledge and reasoning processes is a complex task, yet essential to effective teaching and learning (NRC 2001, 2007). Measurement of students' knowledge is known to be influenced by a variety of factors, which has spurred efforts to understand how different assessment formats and item features differentially inform our inferences about student understanding, as well as the development of new tools and practices to measure authentic problem solving performances (NRC 2001, 2007; Gitomer & Duschl, 2007). Recent work has documented limitations of multiple-choice instruments [MCI] and constructed-response instruments [CRI] on the measurement of student knowledge. Research has documented that MC tests have limited ability to assess students' knowledge selection, organization, and communication (Martinez, 1999), and are often poor predictors of real-world performance (NRC 2001; Nehm, Ha, & Mayfield, 2011). Furthermore, MCIs and CRIs are known to elicit different levels of cognitive activity during problem solving, with CRIs generally acknowledged as having a broader capacity for measuring higher order cognitive processes (Guildford, 1967; Ward, Dupree, & Carlson, 1987; Martinez, 1999). In addition, performance on CRIs has been found to have greater correspondence to clinical interviews than MCIs, suggesting that CRIs are a more valid measure of student reasoning than MCIs (Nehm & Schonfeld, 2008). However, the documented limitations of MCIs may be more reflected of their typical use (measuring recall) than their capacity to measure complex thinking (Aiken, 1982).

Several recent studies in science education using MCIs and CRIs have documented item feature effects that significantly control the elicitation and measurement of student knowledge. For example, familiarity with the construct to be tested has been associated with higher performance on MC assessments and higher confidence in response accuracy in physics education (e.g., Caleon & Subramaniam, 2010) and chemistry education (e.g., Rodrigues, Taylor, Cameron, Syme-Smith, & Fortuna, 2010). In biology education, recent research using isomorphic CR items differing in only the subject feature (e.g. plant vs. animal) or the familiarity of the subject feature (e.g., penguin vs. prosimian) will produce markedly different measures of both students' evolutionary understanding and their naïve ideas or misconceptions (Nehm & Ha, 2011; Opfer et al., 2011). Likewise, items using the same organisms, but prompting explanations for different polarities of trait evolution (e.g., gain vs. loss) will also produce significantly different measures of student understanding and naïve ideas (Nehm & Ha, 2011; Opfer et al., 2011). Careful control of such item features in assessment design is necessary to produce valid inferences about the scientific and intuitive ideas that student's harbor.

Although the documentation of various item features and familiarity effects on reasoning using CRIs have resulted in important advances in the measurement of students' evolutionary understanding, many unanswered questions remain about CRI biases. One area that remains to be investigated is whether the sequencing (order) of CRI items in an assessment has a significant or meaningful impact on measures of student performance and whether the features of the items relate to the magnitude of sequencing effects across an assessment. Such concerns are not new to educational research, but have been largely restricted to MCIs (e.g., Mollenkopf 1950; Leary & Dorans 1982, 1985). Indeed, a large body of work has focused on research examining item sequencing effects (e.g., Mollenkopf, 1950; Monk & Stallings, 1970), later expanding to include

investigations of interactive effects among item features and external factors (e.g., test anxiety, gender, and levels of student achievement) on student performance (e.g., Munz & Smouse, 1968; Plake, 1980). Research in this area has been concentrated around a simple but important question: "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? " (Leary & Dorans, 1985, p.389). Our study investigates this question using three CRIs about evolutionary change.

Research on Item Sequencing

Despite more than half a century of investigation, research on item sequencing using MCIs has not resulted in any clear consensus (Leary & Dorans, 1985). A few trends, however, have been found across studies. First, sequencing of item according to difficulty (e.g., E-H, H-E) has been shown to affect student performance, with E-H sequences associated with higher student performance. It is widely assumed that E-H sequence might therefore reduce the effects of external factors, such as test anxiety, on student performance, but investigations of these interactions have produced inconsistent results (Cronbach, 1984; Leary & Dorans, 1985). Second, studies that compared different testing environments, such as speeded (timed) and power (untimed) conditions, have generally documented decreased performance on item sequences administered under speeded conditions. This appears to be particularly important for qualitative items (e.g. verbal/written) as opposed to quantitative items. And third, great differences with respect to student performance have been found for aptitude tests relative to achievement tests, with performances on tests related to aptitude skills being more susceptible to different item sequences (e.g. H-E, E-H, vs. random) (e.g., Gray, 2004).

The relative breadth of item sequencing research using MCIs, and the general lack of research using CRIs, motivates the studies presented in this paper. Results from prior research demonstrate 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 quantitative and qualitative MC items. CR assessments are inherently qualitative and therefore might demonstrate similar effects of item sequencing on measures of student performance. If measurement of student performance using CRIs is subject to effects of item sequencing, this represents an unexplored bias in CR assessment.

Research Questions

The overarching question addressed in this paper is, how does the diversity or similarity of item features in an assessment relate to the magnitude of sequencing effects and other factors associated with student performance? Specifically, we investigate whether item sequencing differentially impacts the measurement of student performance and whether this measurement is correlated with the type of item features in the assessment. Such feature effects have been documented using CR items (e.g., Nehm & Ha, 2011), however prior research has not examined item sequencing effects or putative interactions between sequencing and types of item features on measures of student performance. In order to determine the impacts of item sequencing and item features on measurement of student performance, we conducted three studies.

Our first study investigated how the *similarity* of item features was related to the magnitude of item sequencing effects on student performance measures. This study was guided by two questions: First, we asked to what extent does item sequencing impact measurement of student performance on an isomorphic, constructed-response assessment (as measured by response accuracy)? Given that prior research on MC tests suggests that sequencing may be an

important factor in MC assessment, we expected that similar trends might occur using CRIs. Based on prior research, we predicted that as student's progress through short-answer, CR item sequences, their performance would decline. Second, we asked to what extent is the magnitude of sequencing effects mediated by the item features in the sequence? Prior research has indicated that student performance is significantly higher when reasoning about familiar items compared to unfamiliar items (e.g., Caleon & Subramaniam, 2010; Rodrigues, Taylor, Cameron, Syme-Smith, & Fortuna, 2010; Nehm & Ha, 2011; Opfer et al., 2011). Additionally, in the context of evolutionary assessment, previous research has documented significantly higher student performance when reasoning about evolutionary gain of a trait compared to the evolutionary loss of a trait. However, this research has not investigated the possibility of interactions between item features and sequencing. We predicted that performance on items characterized by features associated with higher reasoning would be less difficult and therefore less influenced by item sequencing than items characterized by more difficult features (e.g. unfamiliar/ loss). Therefore, we also predicted that interactive effects between item features and sequencing would be least significant for items about evolutionary gain of a trait, followed sequentially by the gaining, losing, and loss of a trait.

Our second and third studies explored how the *diversity* of item features was related to the magnitude of item sequencing effects on measures of student performance. Prior research, including Study 1, has investigated the influences of item sequencing and/or item features in MCIs and CRIs, but has not examined which component of an assessment has a greater impact on measurement of student performance. Specifically, we asked whether certain item features (e.g., familiarity, polarity, taxon/trait, etc.) or item sequencing had differential effects on measures of student performance. We predicted that item sequencing effects would be moderated

by items representing a diverse array of item features and that the category of item features would be significantly associated with measures of student performance across an item sequence.

To address our research questions, we gathered explanations of evolutionary change from a sample of undergraduate students enrolled in non-majors and majors level introductory biology courses at a large Midwestern university. Student responses to a previously published CRI were gathered using an online survey system. This CRI has been shown to generate reliable and valid inferences about student's evolutionary reasoning across levels of biology coursework. Previous research examined more than 130 clinical interviews across an array of item formats to ensure that the items are valid and reliable measures of student knowledge (Nehm et al., in press, Nehm & Schonfeld, 2008).

Scoring of Constructed-Response Items

Students are known to recruit a variety of cognitive elements when constructing evolutionary explanations (Nehm, 2010). These elements may be arranged in a variety of ways, for example, as well-structured networks of scientific knowledge (schemas), mixed models of scientific and naïve components, or models consisting entirely of naïve knowledge elements. For our study of the effects of item sequencing on student performance, we quantified student knowledge by tabulating the frequency of scientific ideas (key concepts) and naïve ideas (misconceptions) in their evolutionary explanations. CR items were presented one at a time under power conditions. The average time for completion of an item was 2.86 minutes, and the average number of words per response was 29.9. We only included responses from individuals who completed all survey items with greater than five words on each item for our analysis.

We scored each student response for seven key concepts of natural selection and six naïve ideas about natural selection (for scoring details, see Nehm & Schonfeld, 2008, and Nehm

et al., 2010). In addition to key concept scores, we tallied the number of *different* key concepts used by a student across the item sequence, which refers to the composite measure of key concept diversity (KCD) (for more details see Nehm & Reilly, 2007). Key concepts included elements such as the presence and causes of variation, the heritability of variation, and differential survival and reproductive success. Naïve ideas examined in this study included teleological reasoning, use and disuse, and intentionality. Two experts who have previously demonstrated high inter-rater reliability (kappa coefficients > 0.8) independently scored all explanations. In cases of disagreement between raters, consensus was established prior to data analysis.

Study 1. Sequencing & Feature Effects in Constructed-Response Assessment: Similar Item Features

Participants & Methods

Student responses were collected from two introductory biology courses for non-majors taught by the same instructor in the same academic quarter in 2010 (n=705, n=611). The number of participants consenting to use their data was relatively low at 35% (n=461). Of the original participants, 67% (n=309) wrote explanations that met the minimum response length requirement (5 words per response), resulting in a total of 1,236 student explanations of evolutionary change for analysis. The average age of the participants was 19.7 years, with 58% of the sample female. Non-Hispanic Whites comprised the majority of the sample (79%) and the average course grade was 3.49 (on a 4.0 scale).

Four items prompted students to explain evolutionary change in four taxa and trait combinations: fish fins, fly wings, shrew incisors, and snail feet. The items used in this study were isomorphic in structure ("How would biologists explain how a living X species

with/without OR small/large Y evolved from an ancestral X species *with/without OR small/large* Y?") but differed in the specific taxa and traits (i.e. X and Y varied). Importantly, the taxa and trait combinations used in this study were all familiar, minimizing potential item feature effects. Students were randomly assigned to one of four groups representing four types of trait change (gain, loss, increase, or decrease). Students were additionally assigned at random to one of four item sequences, resulting in sixteen different treatments (Table 1).). Statistical analyses were performed in PASW (SPSS, Inc.) and JMP (SAS, Inc.).

Table 1. Sequences and features of the constructed response items used in each study. All items were structurally isomorphic and asked students to explain: "How would biologists explain how a living "X" species *with/without OR small/large* "Y" evolved from an ancestral "X" species *with/without OR small/large* "Y"?"

Taxon/Trait	Study 1	Study 2		Study 3		**Familiarity**
Animal	Fish / Fin Fly / Wing Shrew / Incisors Snail / Feet	Snail / poison	N/A	Snail / teeth	Mouse / claws	*Familiar*
		Prosimian / tarsi		N/A		*Unfamiliar*
Plant	N/A	Elm / winged seed		Grape / tendrils	Lily / petals	*Familiar*
		Labiatae / pulegone		N/A		*Unfamiliar*
Trait Polarity	*Gain, Gaining, Losing, or Loss*	*Gain*	*Loss*	*Gain*	*Loss*	
Item Sequence (1/2/3/4)	Fish / Fly / Shrew / Snail Fly / Fish / Snail / Shrew Shrew / Snail / Fish / Fly Snail / Shrew / Fly / Fish	Elm / Lab. / Snail / Pros. Lab. / Elm / Pros. / Snail Snail / Pros. / Elm / Lab. Pros. / Snail / Lab. / Elm		Grape / Lily / Snail / Mouse Lily / Grape / Mouse / Snail Snail / Mouse / Grape / Lily Mouse / Snail / Lily / Grape		

Results

Analysis of student response scores (measured by frequency of KCs and NIs) revealed that item sequencing had sizeable effects on measures of student performance. Overall, students incorporated more KCs than NIs into their evolutionary explanations, the use of KCs decreased significantly with item location, from the first item to the fourth item in the sequence (Wilcoxon signed rank test, $p. < 0.001$) with 36.2% (n=112) of students having more KCs on items

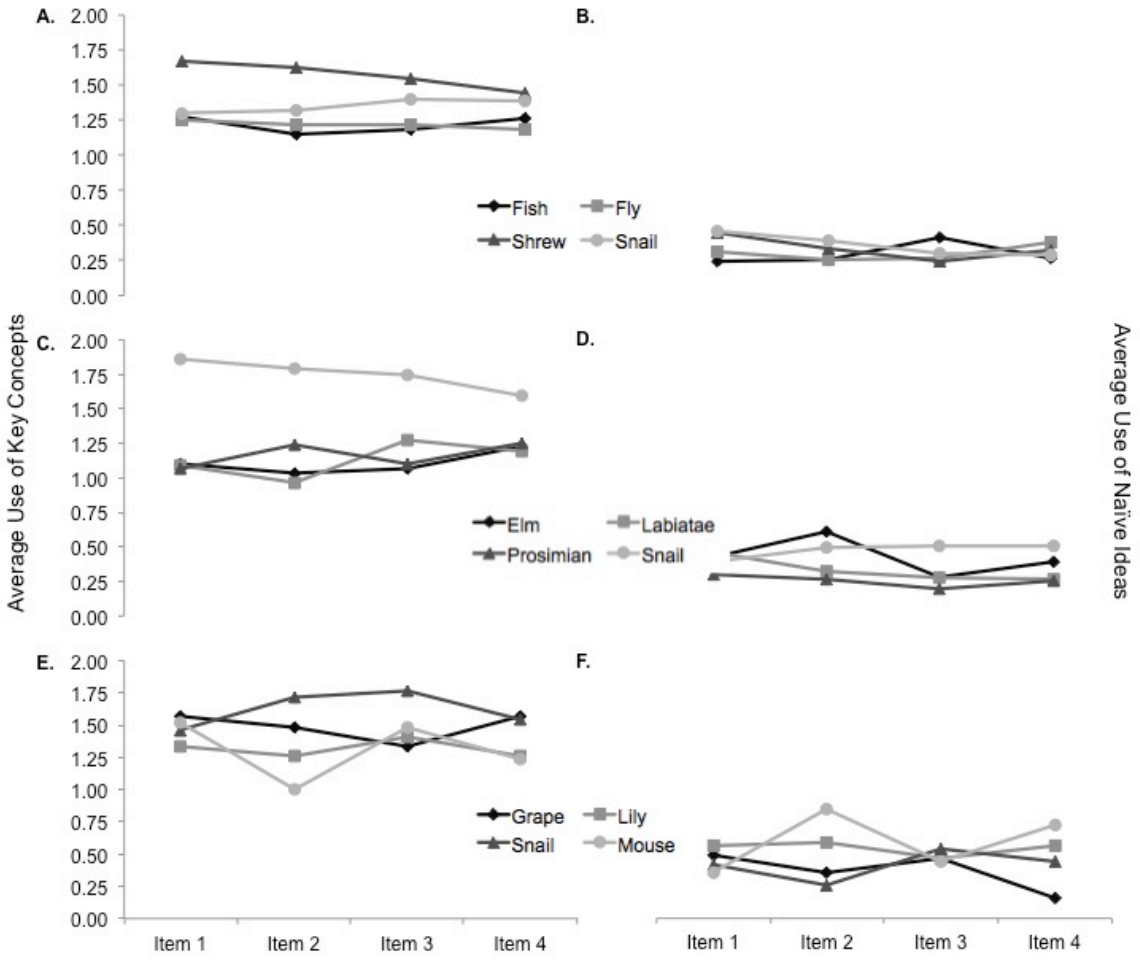
presented first (Figure 1A). Significant decreases in KC use were also found between items presented first and items presented second ($p. < 0.003$) and between second and third ($p. < 0.03$). However, student use of KCs between the third and fourth items in a sequence was not significantly different ($p. < 0.86$) and the use of NIs in student explanations did not differ across items ($p. < 0.36$) (Figure 1B). Despite decline in overall frequencies of KCs with item order, key concept diversity (KCD) scores revealed no significant differences between different item sequences, with the majority of students only using two (29%) or three (37%) KCs across a four-item sequence.

Item features also had a sizeable effect on measures of student performance. Responses to trait gain, gaining, and losing scenarios were significantly higher (i.e., more KCs) relative to response from the trait loss group (Figure 2A). While performance on gain and gaining item sequences were relatively similar, the same was not true for losing and loss item sequences, which were significantly different for all items. Additionally, the magnitude of sequencing effects appeared to be greatest for student performance on items about trait loss (relative to trait gain), with KCs decreasing across the four items for 37% of students. In contrast, item features (e.g., gaining or losing) appeared to mediate item sequencing effects, as there was little change in the frequencies of KCs and NIs across the four items. In general, student use of KCs was more likely to change with respect to item sequencing and features than use of NIs, however this could be due to the relatively infrequent use of NIs by our student sample.

In contrast to item sequencing, KCD scores were significantly different across item features. Items asking about trait loss had significantly lower scores compared to the other three groups (Wilcoxon signed rank test, $p. < 0.0001$). More than one-third of students responding to trait loss item sequences had a KCD score of 0 or 1, compared to only 18.7-21.5% of students in

the other groups. Additionally, only 8.9% of students responding to loss item sequences used 4 or more different KCs, compared to 20-28.9% of students in the other three groups. KCD scores were also significantly related to academic success in the course, regardless of treatment (i.e., item features), $X^2(9) = 34.36$, $p. < 0.000$, with the majority of students with a KCD score of 4 or greater having received an 'A' (26.4%) or 'A-' (25.7%) in the class. Likewise, 30.8% of students that received a 'B or below' had a KCD of 0 or 1.

Figure 1. Effects of item sequencing on measurement of key concepts in student responses to constructed-response items. A/B. Study 1 – Similar Item Features; C/D. Study 2 – Varied Familiarity; E/F. Study3 – Varied Trait Polarity.



Response verbosity (number of words) was related to item sequencing, with responses to the first item in a sequence greater relative to the fourth item (Figure 3A.). Declines in response

verbosity were significant for responses to all groups (Wilcoxon signed rank test, p. <0.0001). Responses to the first item were on average 40.4 words long and decreased to an average of 28.2 words in the fourth item. Response to items about the evolutionary Gaining of a trait demonstrated the greatest change across the item sequence, from an average of 42.9 words in the first item to an average of 25.1 words in the fourth item.

Verbosity was also significantly related to KC use (Spearman's rank correlation, r = 0.63) and KCD scores (r = 0.505), with corresponding increases in verbosity with the addition of KCs. In addition, academic success in the courses we studied was positively related to the frequency of KCs and response verbosity (r = 0.245). Students who earned an 'A' in the course had longer responses and more KCs than students who earned a 'B or below'. Item sequencing effects were more pronounced in the 'A' group, as they used more words and KCs in Item 1. Other demographic variables, such as gender and academic major, were not related to student performance or item sequencing effects. Student use of NIs was more weakly associated with response verbosity (r = 0.10).

Overall, students' evolutionary explanations tended to incorporate more KCs than naïve ideas (NIs), with 42% of students using a total of 3-4 KCs and 30% using a total of 1-2 NIs across an item sequence. Additionally, use of more KCs corresponded with decreased likelihood of using a NI. For example, of students who used 4-5 KCs in their first response (n=20), only10% were likely to also use a NI. In contrast, students who did not use any KCs in their first response (n=81) were 59.4% more likely to use a NI in their evolutionary explanation.

Discussion

Prior research has documented multiple effects of item sequencing for multiple-choice assessments (e.g., Mollenkopf, 1950; Munz & Smouse, 1968; Kingston & Dorans, 1984). Factors such as item sequencing, difficulty, and features have been clearly linked to student performance on various assessments (e.g., Cronbach, 1984; Leary & Dorans, 1985). To our knowledge, however, these factors had not been empirically examined for CRIs. The results of Study 1 clearly demonstrate effects of item sequencing and item features on measures of student performance using a CRI. Our first research question addressed the extent to which item sequencing impacts measurement of student performance. Despite the use of isomorphic items, student performance on CR items significantly declined across the item sequence. Interestingly, while KC frequencies declined across an item sequence NIs did not, suggesting that elicitation of accurate scientific knowledge may be more affected by item sequencing than naïve ideas.

Our second research question addressed the extent to which the magnitude of sequences of effects was mediated by item features within a sequence. Examination of the item features revealed that certain item features (e.g. gain, gaining, losing) significantly mediated the magnitude of sequencing effects across an isomorphic item sequence. Our results were consistent with prior research (Nehm & Ha, 2011); across all item sequences, student responses to gain items were higher than student responses to loss items. Importantly, KC and NI frequencies and KCD scores were significantly related to the item features, with items characterized by more difficult features (i.e., loss) having the least impact on item sequenceing effects. . However, contrary to our initial prediction, performance on item sequences about the gaining *and* losing of a trait were more similar to responses about trait gain, indicating that student reasoning about trait loss is different than the losing of a trait. Additionally, student performance did not

significantly differ between item sequences about gaining or losing, suggesting that reasoning about transitional trait change (regardless of direction) is different that reasoning about gain or loss.

An interesting finding from Study 1 was that while student performance declined significantly after the first and second items in the sequence, the decline in KCs was directly related to the response verbosity. Across an item sequence, as response length decreased, response accuracy also decreased, resulting in responses to items towards the end of a sequence more likely to end up with scores that appeared to underrepresent students' knowledge. It is possible that students recognized the isomorphism of the items in the instrument, resulting in errors of omission. If sequential isomorphic questions are causing errors of omission or resulting in test fatigue, CRIs using such formats could be underestimating students' understanding and performance. For example, in our study 21% of students, on average, dropped a KC between their first response and their second response. If this decrease in KC use is due to student recognition of the isomorphism of items, then isomorphic item sequences with similar features may be limiting inferences made from student explanations. This suggests that the similarity and difficulty of item features is an important consideration when interpreting student performance on individual items within a sequence.

Study 1 examined the effects of item sequencing on student performance and the use of item features to mediated such sequencing effects. The items used were carefully chosen to consistently administer the item features of familiarity (i.e., all familiar), taxon (i.e., all animals) and polarity of evolutionary change (e.g., gain, gaining, losing, or loss). In order to further clarify the impact of item features and sequencing on student performance using CRIs, we conducted

two follow-up studies to examine in more detail the impact of specific item features on the interaction of item sequencing effects.

Study 2. Sequencing & Feature Effects in Constructed-Response Assessment: Diverse Item Features - Familiarity

Participants & Methods

Student explanations were collected from two introductory biology courses for biology majors during the 2010 academic year. The average age of the participants was 20.5 years, with 55% of the sample female (we are unable to report information regarding the ethnicity of participants in this sample). The majority of the participants (89.5) had only completed 1-2 college biology courses prior data collection. We collected explanations using CR items that allowed us to focus on the effects of different item features on student performance across item sequences. As in Study 1, four items prompted students to explain evolutionary change in four taxa and trait combinations. However, for this study students were asked items that were consistent with respect to polarity of trait change (i.e., gain of trait) but varied in taxon type (i.e. animal vs. plant) and familiarity of taxon (i.e., familiar vs. unfamiliar). The items used in this study were isomorphic in structure to the items used in Study 1, but differed in taxon/trait combinations in accordance to the criteria above (Table 1). In addition, students were randomly assigned to one of four item sequences, resulting in eight total treatments. Analysis of student explanations (n=262) for familiarity effects resulted in a total of 1,048 student explanations about evolutionary change across familiar and unfamiliar items. Statistical analyses were performed in PASW (SPSS, Inc.) and JMP (SAS, Inc.).

Results

Analysis of student response scores (measured by frequency of KCs and NIs) revealed that item sequencing was not a significant factor for student responses to three out of the four items (Figure 1C). Interestingly, student responses to the Snail poison item, which represented a familiar, animal item type, contained on average significantly more KCs than the other items and KC decreased significantly across the item sequence ($F[1,1046] = 5.951$, $p. < 0.015$). On average students incorporated more KCs than NIs into their evolutionary explanations, the use of KCs and NIs did not significantly decrease with item location (Figure 1D). Although overall the number of KCs was greater in student explanations, the diversity of key concepts (KCD) did not vary significantly across different item sequences ($F[3, 359] = 0.004$, $p. = 1.00$). On average, 34.5% of students incorporated 0 – 1 different KCs into their explanations and only 9.6% used more than 4 KCs across the four items.

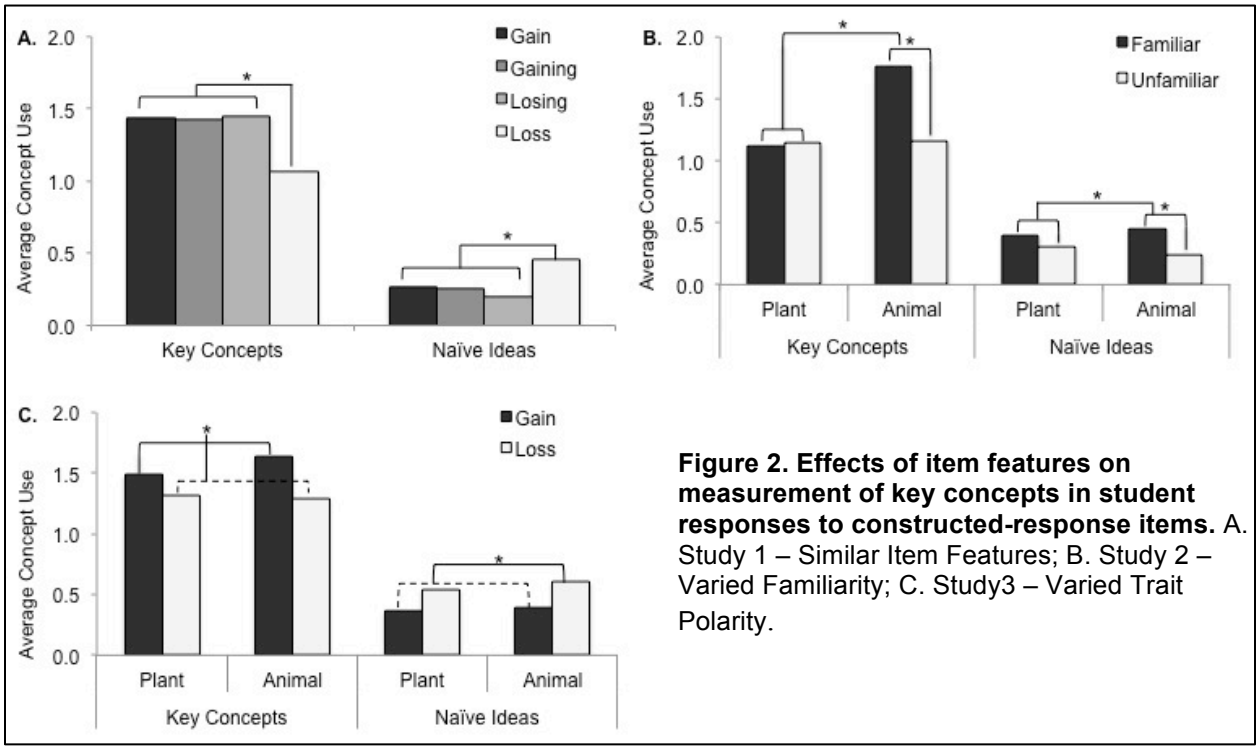


Figure 2. Effects of item features on measurement of key concepts in student responses to constructed-response items. A. Study 1 – Similar Item Features; B. Study 2 – Varied Familiarity; C. Study3 – Varied Trait Polarity.

Analysis of student responses with respect to item features (i.e., familiarity and taxon) revealed significant effects of features on measures of student performance (Figure 2B). Student explanations of evolutionary change about familiar items contained significantly more key concepts (F[1, 1046] = 17.717, p. < 0.0001) as well as almost 40% more naïve ideas (F[1, 1046] = 16.868, p. < 0.0001). However, the average use of KCs (1-2 KCs, 58.4%) was consistent across both familiar and unfamiliar items. In contrast, 31.7% of student responses to familiar items used one to two NIs compared to only 23.9% in response to unfamiliar items. The taxon (i.e., animal vs. plant) of the items was also significantly related to the use of KCs (F[1, 1046] =24.597, p. < 0.0001), but not NIs. Items asking students to explain evolutionary change in animals contained, on average, more KCs than those about plants. The KCD used in student responses was significantly different between familiar and unfamiliar items(t[533] = 4.412, p. < 0.0001), with responses to familiar items averaging 2.13 KCs compared to 1.67 KCs for unfamiliar items. No differences in KCD were found between animal and plant items.

Examination of the interaction between item familiarity and taxon revealed significant effects for the use of both KCs (F[1, 1046] = 4.933, p. < 0.027) and NIs (F[1, 1046) = 11.274, p. < 0.001) in student responses. Student performance on the familiar animal item (e.g, snail poison) was significantly higher than on the unfamiliar animal (e.g., prosimian tarsi) and plant items. Interestingly, use of naïve ideas was also higher for familiar items of both taxa. Furthermore, analysis of the interaction between item sequencing and item features (i.e., familiarity and taxon) revealed no effects for the measurement of KCs in student responses, however measures of NIs were significantly affected (F[1, 1046], = 7.505, p. < 0.0006). Student use of NIs was most prevalent in the first and second items of the sequence, however, NIs tended

to increase across the item sequence for the familiar items. A similar pattern was not found for items representing different taxa.

Students' response verbosity was significantly related to the use of KCs ($F(90, 957) = 4.306$, p. < 0.0001), but not NIs ($F(90, 957) = 1.149$, p. = 0.170) (Figure 3B). In addition, response verbosity was significantly related to item familiarity ($t[1046] = 3.710$, p. =0.018), but not the taxon (i.e., animal or plant) ($t[1046] = -0.764$, p. < 0.772). Student responses to more familiar items were significantly longer than responses to unfamiliar items, however differences in response lengths are mediated by the interaction between item familiarity and taxon. Verbosity was also significantly related to KCD, with lower KCD scores (e.g. 0 or 1) significantly associated with less verbose responses ($F(144, 903) = 2.994$, p. < 0.0001).

Discussion

Our second study explored the relationship between the familiarity of the item and item sequencing. Prior research on familiarity has documented effects of problem familiarity for multiple-choice assessments (e.g., Caleon & Subramaniam, 2010). Factors such as accuracy, confidence, and context have all be associated with problem familiarity and student performance on various assessments (e.g. Caleon & Subramaniam, 2010; Rodrigues, et al. 2010). However, as in our first study, the effect of item familiarity on student performance had yet to be empirically examined for CRIs.

Our overarching research question for this study was whether or not the diversity of item features was related to the magnitude of item sequencing effects on measures of student performance. The results from Study 2 indicate that item features do mediate the effects of item sequencing for KCs, which were relatively consistent across items sequences that varied with

respect to familiarity and taxa. In contrast, the use of NIs varied across the item sequence, but generally increased from the first item to the fourth item. Of the three factors examined in this study, familiarity appeared to be the most related to measures of student performance, with more accurate responses to familiar items. However, student performance (as measured by KCs) on familiar and unfamiliar items was relatively consistent from the first to the fourth item. Therefore, interactions among item sequencing and familiarity of item features may provide a more accurate measure of student performance across a constructed-response item sequence.

Corroborating the results of our first study, response verbosity was significantly related to KC use. However, the results of this study suggest that verbosity may also be related to other features of the assessment. Item familiarity was significantly related to students' response verbosity suggesting that students are writing longer, more accurate explanations when responding to items that are familiar than items that are unfamiliar. This raises a serious concern for assessment of student understanding of evolutionary change. If student performance is linked to familiarity, then it suggests that students have difficulty transferring their knowledge and understanding of evolutionary change to novel situations. Likewise, if familiarity and verbosity are linked, then students are likely to write less in response to novel situations and more likely to be evaluated as having an inaccurate or naïve model of evolutionary understanding.

Study 3. Sequencing & Feature Effects in Constructed-Response Assessment: Diverse Item Features – Trait Polarity

Participants & Methods

Student explanations were collected from two introductory biology courses for biology majors during the 2010 academic year. The average age of the participants was 20.1 years, with 55.4% of the sample female (we are unable to report information regarding the ethnicity of

participants in this sample). The majority of the participants (87.1) had only completed 1-2 college biology courses prior data collection. We collected explanations using CR items that allowed us to focus on the effects of different item features on student performance across item sequences. As in the previous studies, four items prompted students to explain evolutionary change in four taxa and trait combinations. However, for this study students were asked items that were consistent with respect to familiarity (i.e., familiar) but varied in the polarity of trait change (i.e. gain vs. loss) and taxon (i.e., animal vs. plant). The items used in this study were isomorphic in structure to those used previously, but differed in taxon/trait combinations in accordance with the above criteria (Table 1). Students were randomly assigned to one of four item sequences, resulting in eight total treatments. Analysis of student explanations (n=156) for effects due to polarity of trait change resulted in a total of 624 student explanations about the evolutionary gain and loss of traits.

A potential bias in the prior two studies is that they did not consider R/W's as a factor that might differentially impact student performance in constructed-response assessment. In an attempt to assess the relationship between scientific R/W ability and measured performance on CR items, we asked students participating in Study 3 to self-report their R/W in science prior to completing the item sequence. Students rated their R/W ability in science on a Likert scale (1 – 5), with higher numbers corresponding to a perceived higher R/W ability.

Results

Prior to discussing item sequencing and feature results for this study, we report on the relationship between student performance and scientific reading and writing (R/W) scores. More than 85% (n=134) of students self-reported a R/W score of 4 or 5, indicating that they perceive themselves to have a relatively high ability to read and write in science. While students who self-

rated higher R/W's had, on average, more scientifically accurate responses (as measured by frequency of KCs and NIs) and more verbose responses, R/W scores were only weakly correlated with NI use (Spearman's rank correlation, r = -0.112) and response verbosity (r = 0.174). The relationship between R/W scores and KC use (r = 0.232) and KCD (r = 0.236) was slightly stronger, with more diverse responses corresponding with R/W scores of 5, followed sequentially by 4, and 3 and below.

Analysis of student response scores across the four item sequences revealed no significant differences between students' use of KCs or NIs across an item sequence (from the first item to the fourth item). Incorporation of KCs into student responses was relatively consistent across item sequences (Figure 1E). On average, the majority of student responses (59.1%) contained 1-2 KCs per item. However 23.6% of students never used a KC in their responses and more than one-third of students (36.4%) used 1-2 NIs in their responses. In contrast, the use of NIs was highly inconsistent between items and did not relate to the sequencing of items within the assessment (F[1, 622] = 0.000, p. = 1.00) (Figure 1F). Corroborating our results from the previous two studies, while the measured use of KCs in student explanations was greater than NIs, the diversity of key concepts was not significant across item sequences (F[3, 152] = 0.600, p. =0.616). On average, 17.3% of students incorporated 0 – 1 different KCs into their explanations and 23.7% used more than 4 KCs across the four items.

In contrast to item sequencing, item features (i.e. trait polarity and taxon) had significant effects on measures of student performance. Student responses to items that differed in trait polarity (i.e., gain vs. loss) varied significantly with respect to KCs and NIs (F[1, 622] = 7.906, p. < 0.005; F[1, 622] = 12.293, p. < 0.001). Responses to items about the gain of a trait contained

on average more KCs relative to loss items (Figure 2C). Students averaged 1.56 KCs in response to gain items, with 49% of these responses contained 2 – 4 KCs. Accuracy of student responses to loss item sequences was significantly lower, averaging 1.3 KCs across items and 62.5% of these responses contained 0 – 1 KCs. Responses to items about the loss of a trait contained also more NIs, with 43% containing 1 – 2 NIs in responses to trait loss compared to 30.8% in responses to trait gain. Examination of student response scores with respect to item taxon revealed no differences in measured student performance across items about different animal and plant taxa (See Figure 2C). Student use of KCs were not significantly different between animals and plants (F[1,622] = 0.488, p. = 0.485), averaging 1.4 KCs for both animal and plant items, with the majority of responses (59.1%) containing 1 – 2 KCs across the item sequence. The use of NIs in student responses were also not different between animal and plant items (F[1, 622] = 0.657, p. = 0.418), with more than two-thirds of student responses containing 0 NIs.

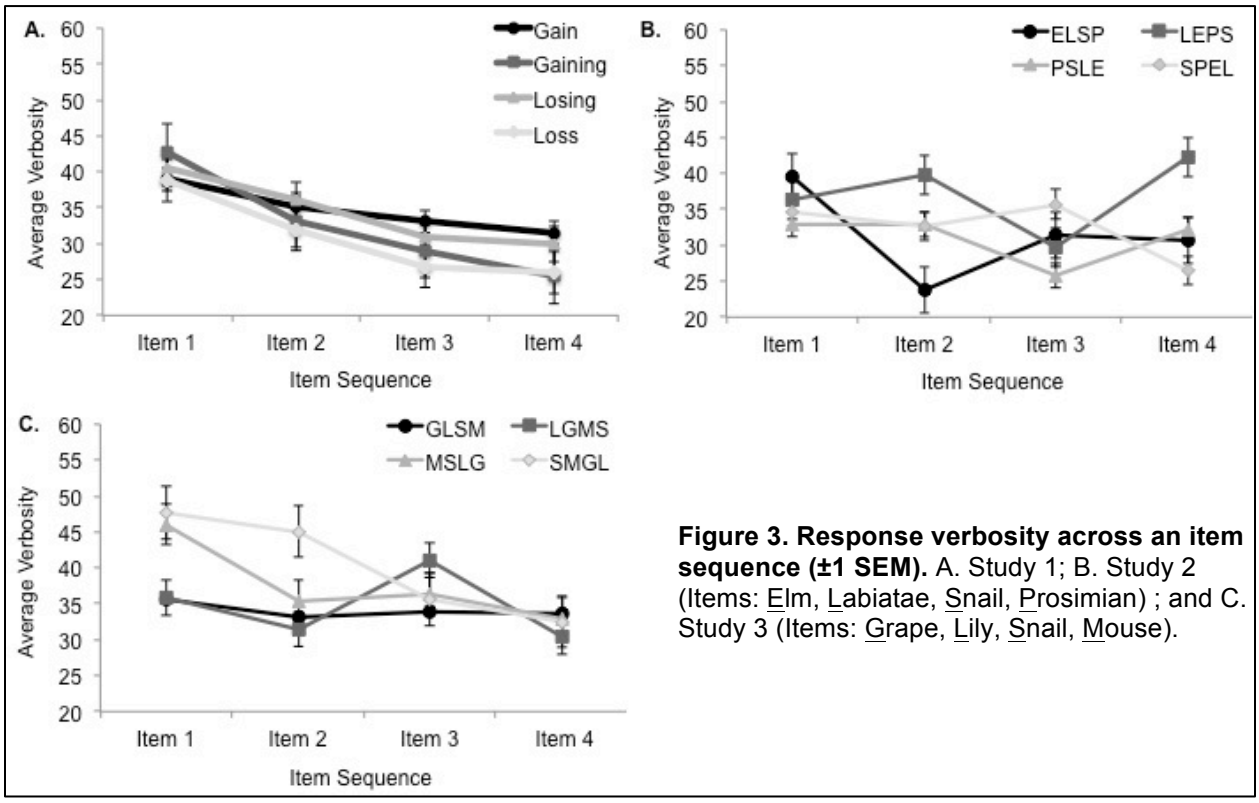


Figure 3. Response verbosity across an item sequence (±1 SEM). A. Study 1; B. Study 2 (Items: Elm, Labiatae, Snail, Prosimian) ; and C. Study 3 (Items: Grape, Lily, Snail, Mouse).

Supporting the differences in the distribution of KCs across gain and loss items, KCD scores were significantly different between gain items and loss items ($t[310] = 2.037$, $p. < 0.04$). More than one-third of responses to loss items had a KCD score of 0 or 1, compared to 27% of responses to gain items. In addition, KCD scores of 3 – 6 were more frequently found in response to gain items. There were no significant differences in KCD scores between plant and animal items, although KCD scores on plant items tended to be lower than those on animal items.

All together, interactions between the item features significantly affected the measured use of NIs in student responses ($F[1, 622] = 6.866$. $p. < 0.009$), but not KCs ($F[1, 622] = 3.022$, $p. = 0.083$). Responses to trait loss and gain in animals contained more NIs than equivalent items about plants. Similarly, analysis of the interaction between item sequencing and item features (i.e., trait polarity and taxon) revealed a significant effect on measures of NIs ($F1, 622] = 7.885$, $p. = 0.005$) but not KCs. While student use of NIs tended to increase across an item sequence, responses to trait loss in animals in plants were higher than trait gain.

The verbosity of student responses decreased significantly declined across an item sequence ($F[1, 622] = 8.571$, $p. < 0.004$) by an average of 8.7 words from the first item to the fourth item (Figure 3C). Verbosity was also significantly related to the subject of the item (taxon), ($t[622] = -2.449$, $p. <0.002$) with more verbose responses to animal items. However, the polarity of the item had no significant effect on verbosity ($t[622] = -1.260$, $p. = 0.304$). Students' response verbosity was also significantly related to the use of KCs ($F[98, 525) = 4.235$, $p. < 0.0001$) but not NIs.

Discussion

Our third study examined the relationship between trait polarity and item sequencing. Prior research, including Study 1, has indicated that trait polarity has significant effects on measures of student performance to CR items (Nehm & Ha, 2011). Student performance on items about trait gain tends to be higher than student performance on trait loss. As in Study 2, our overarching research question for this study was whether or not the diversity of item features was related to the magnitude of item sequencing effects on measures of student performance. Our results clarify the role of trait polarity and taxon/trait in mediating item sequencing effects.

Similar to Study 2, trait polarity appeared to be more related to student performance than the item taxon/trait. Student performance on items that differed according to polarity was significantly higher when responding to items about trait gain, regardless of taxon/trait. In addition, our results indicate that performance (as measured by KCs) on items that vary with respect to trait polarity was relatively consistent across an item sequence. In contrast, the use of NIs varied across the item sequence, but generally increased from the first item to the fourth item. This provides further support for the role of item features such as trait polarity in mediating effects of item sequencing in CR assessment.

Consistent with our previous studies, response verbosity was significantly related to both KC use and item features. However, unlike the results of Study 2, the taxon/trait of the item was significantly related to students/ response verbosity. Students wrote longer, more accurate explanations when responding to items about evolutionary change in animals relative to change in plants. Differences in performance due to the subject of the item is another concern for valid assessment of student understanding. As with familiarity, it suggests that students have difficulty transferring or applying understanding of evolutionary change across different taxa. Therefore,

assessments using CR items across various taxa/traits may result in responses that vary in accuracy and verbosity relative to items that are similar in taxa/traits.

Overall Conclusions

Prior studies using constructed-response assessment have clearly documented significant effects of item features on student performance, and our work corroborates this research (e.g., Nehm & Ha, 2011). However, our results also indicate that item sequencing is an important consideration for measuring students' scientific understanding using constructed-response assessments, particularly when they are similar in surface features. The studies presented in this paper serve to clarify the interaction of these assessment features and their effects on the measurement of student performance using CR assessment (Table 2). We present a conceptual model that attempts to elucidate the effects of the assessment features examined in this paper on student performance (Figure 4). In addition, we present verbosity as a feature of responses to CR items that is related to student performance.

Item Sequencing

As demonstrated by our first study, student responses to CR items with similar item features tended to decline in accuracy (as measured by KCs) across an item sequence, suggesting that item level evaluation of student performance is influenced by sequential presentation of items. However, effects of item sequences in an assessment on measures of student performance appeared to be mediated by the diversity of item features in our second and third studies. In particular, the familiarity of items appeared to significantly control for effects of item sequencing. In addition, analysis of the diversity of key concepts, as opposed to the frequency of individual concepts, was not impacted by different item sequences. This suggests that KCD

provides a holistic evaluation of student performance on CR item sequences, independent of item order. In addition to potential influence on concept use, our results support a relationship between item sequencing and response verbosity. Responses at the start of an item sequence were consistently more verbose than those at the end of an item sequence.

Item Features

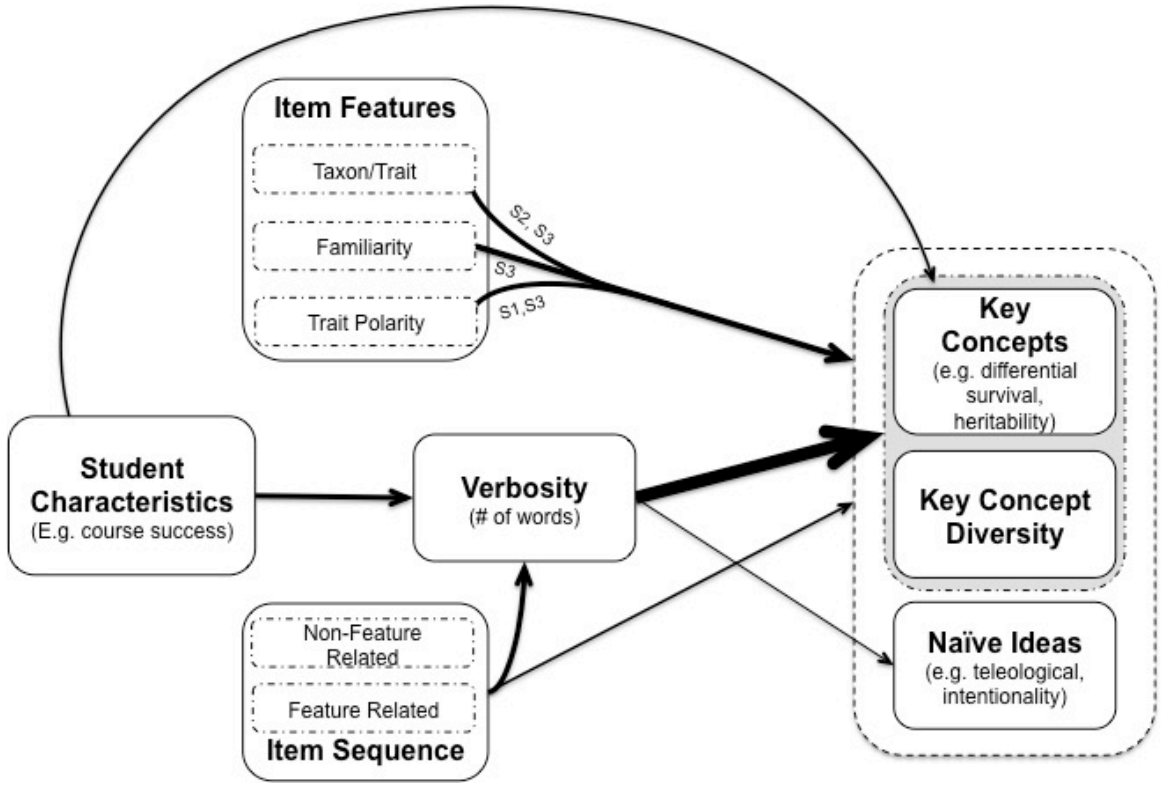
Item features were directly related to KC use, with "easier" features eliciting more KCs than "more difficult" features, independent of the item order. Familiarity of item context appears to be the most related to concept use, and was significantly related to both use of KCs and NIs in student responses. However, concept use was also influenced by the other item features examined in this paper. In addition, item features were significantly related to the verbosity of student responses, which in turn influences concept use. Based on our results, we argue that item features are more influential on student performance than item sequencing. While both assessment features were directly and indirectly (through response verbosity) related to concept use, item features were consistently related to student performance, regardless of item sequence. In contrast, effects of item sequencing on student performance were somewhat mediated by the presence of diverse item features.

Verbosity

Our results also suggest that scientifically accurate concept use in CR item sequences may be largely product of response verbosity. Key concepts were significantly more related to verbosity than were naïve ideas; students who wrote more tended to incorporate more key concepts per response, but not more naïve ideas. While verbosity was significantly related to

item sequences with similar features, the relationship was inconsistent for item sequences with diverse features. This suggests that item features are potentially more important for response verbosity. Features such as item familiarity were particular influential on response verbosity and accuracy. Furthermore, the use of diverse item features in a CRI mediated differences in response verbosity across item sequences.

Figure 4. Conceptual model of student explanations to constructed response items about evolutionary change.



Participant Characteristics

In addition to features of the assessment, characteristics of participants appear to be directly related to student performance. Characteristics such as cumulative GPA and course grade appear to be related to concept use in student responses. Students with higher academic success tended to use more KCs and fewer NIs than students with lower academic success.

However, we were not able to fully evaluate these relationships in our current paper, as our sample populations were relative homogenous and high performing. An increased sample of participants across different levels of academic success may clarify this association.

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. Study 1 measured the knowledge of biology non-majors enrolled in courses where evolution was posited as a main theme. In contrast, Studies 2 and 3 measured the knowledge of introductory biology majors, enrolled in their first course series containing evolution instruction, which is generally more advanced than the non-major course. Our results indicated that students in Study 1 (non-majors) incorporated slightly more KCs into their responses, on average, than students in Study 2 or 3 (majors). We suggest that the similarity of item features in Study 1 is the simplest explanation for differences in KC use between the three studies. However, there are other factors that could explain the differences in measured performance. It could be that sequencing effects are related to the amount or type of biological knowledge held by the participants. Biology majors are enrolled in a course that is designed to provide them with an introductory overview of an array of biological topics for future courses. The course for biology non-majors is much more focused on core themes and theories that can be used to organize their understanding of biology topics. Differences in sequencing effects between the two groups may be influenced by the instructional goals or the type of biological content.

Table 2. Summary of Results.

Study foci	Study 1	Study 2	Study 3
Item Sequencing			
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
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 Interactions			
KCs	N/A	Familiar Animal > Familiar Plant = Unfamiliar Animal/Plant**	No difference across items
KCD		N/A	N/A
NIs		Familiar Animal > Familiar Plant = Unfamiliar Animal/Plant**	Animal Gain/Loss > Plant Gain/Loss**
Feature & Sequencing Interactions			
KCs	N/A	No difference across items	No difference across items
KCD		N/A	N/A
NIs		1st Item < 4th Item for Familiar Animals/Plants**	1st item < 4th item for Loss in Animals/Plants**
Verbosity	*1st Item > 4th Item** for Gaining, Losing, and Loss*	*No difference across items*	*1st Item > 4th Item**
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

***p. < 0.01; *p. < 0.05*

A second limitation of our studies is that differences in reading and writing ability (e.g., English language learners vs. native speakers) were not explicitly taken into consideration for two of the three studies. While the results from Study 3 suggest that reading and writing ability may not be a strong predictor of student performance, students may have varying degrees of success interpreting the item and constructing a scientifically accurate written response, relative to performance on multiple choice assessments or clinical interviews. However, the items used in this study have been previously validated with clinical interviews and multiple-choice assessments, and shown to produce reliable inferences (Nehm & Schonfeld, 2008; Nehm, et al. in press).

Finally, while the results of our three students offer clear implications for constructed response assessment in biology education, suggestions for other scientific domains may be limited. The nature of our study was to investigate the relationship between item features and item sequencing effects on student performance on isomorphic, constructed response items. Are assessments of this type limited to biology? While item-sequencing effects in constructed response items have been relatively unexamined in the research literature, several recent studies in chemistry education have examined how variation in surface features relates to student problem solving success (Eg. Gulacar & Fynewever, 2010.; McClary & Talanquer, 2011) While it does not appear that surface features have been directly manipulated in these studies, features did vary across problems in an assessment.

Implications for Previous & Future Research

Although our study design and item sequences were different from prior studies, our findings suggest that prior work using sequential, isomorphic, constructed-response items in some cases may have underestimated the accuracy of students' scientific knowledge.

Specifically, our results suggest that item sequencing and the diversity of item features in an assessment influence the frequency of accurate, scientific knowledge. While our results do support the effects of differential item features on measurement of student performance, they also support the impact of item order on assessment of students' scientific knowledge. Importantly, impacts of item order were greater in our studies when the item features were very similar, a format that, to our knowledge, has not been used in other studies. The majority of research on evolution assessment, including Nehm & Reilly (2007), has used items that are isomorphic in structure with some variation in surface features.

Another consideration for measurement of student performance using constructed-response assessment is the manner in which student knowledge is quantified. Our studies presented here used both key concept frequency and key concept diversity as measures of student performance. Previous work on evolution assessment has argued that KCD provides a more accurate measure of student understanding as it represents the number of different, scientifically accurate ideas a student uses in response to a set of isomorphic CR items (e.g., Nehm & Reilly, 2007). The results of our three studies support that both total key concept frequency and KCD can be used to obtain accurate measures of student performance. However, our study suggests that the method that provides the best measure of student knowledge depends upon the level of focus. Across our three studies, total key concept frequency varied to some degree with respect to item sequencing and features. In contrast KCD was equivalent across item sequences within an assessment, suggesting that it provides a measure of student understanding somewhat buffered from item sequencing effects. Therefore, if researchers are interested in overall performance on an item sequence, KCD is a holistic measure of students' understanding that is

not significantly influenced by the order of items within a sequence. In contrast, key concept frequency provides a measure of student performance on individual items within a sequence.

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-responses assessments. A potential concern for this and other item sequencing research findings 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 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. This also highlights the important of response verbosity with respect to student performance. Students should be encouraged to construct more verbose responses in order to more accurately measure their knowledge and understanding. Computerized assessments could easily facilitate increased minimum verbosity of responses by requiring a minimum length before the participant can submit their response.

Future work on constructed response assessments should consider the effects of item sequencing and item features on measures of student performance. Given our results, we recommend that assessments include a diversity of surface features to ensure equivalent response verbosities. Likewise, the items should be only moderately isomorphic. Performance on items that are too similar, such as those in Study 1, may significantly bias estimates of student understanding because of declining response verbosity. High similarity of item features may result in item recognition or repetitive responses. However, as documented in Study 2, increasing

the diversity and familiarity of item features within a sequence resulted in relatively consistent response lengths. Nonetheless, student performance on items containing diverse item features did significantly differ across levels of familiarity and taxon/trait.

Acknowledgments

We thank Judy Ridgway and Minsu Ha for help with data collection and analysis and the National Science Foundation REESE program (DRL 0909999) and the Marilyn Ruth Hathaway Education Scholarship (Rector) for financial support. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the lab and do not necessarily reflect the view of the NSF.

References

Aiken, L.R. (1982). Writing multiple-choice items to measure higher-order educational objectives. Educational and Psychological Measurement, 42, 803.806

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.

Cronbach, L.J. (1984). Essentials of psychological testing. New York: Harper and Row.

Gitomer, D.H. & Duschl, R.A. (2007). Establishing multilevel coherence in assessment. In: Moss, P.A. (Ed.). Evidence and decision making. The 106th yearbook of the National Society for the Study of Education, Chicago, 288-320.

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

Guilford, J.P. (1967). The nature of human intelligence. New York: McGraw-Hill.

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

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.

Leary, L.F. & Dorans, N.J. (1982). The effects of item rearrangement on test performance: A review of the literature (RR 82-30). Educational Testing Service, Princeton, N.J.

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.

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

McClary, L. and Talanquer, V. (2011). College chemistry students' mental models of acids and acid strength. Journal of Research in Science Teaching, 48(4): 396-413.

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). IEffect of item order on test scores. Journal of Educational Research, 63, 463-465.

Munz, D.C. & Smouse, A.D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. Journal of Education Psychology, 59, 370-374

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. National Academy Press, Washington, D.C.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. National Academy Press, Washington, D.C.

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., Rector, M.A., and Ha, M. (2010). "Force Talk" in evolutionary explanation: metaphors and misconceptions. Evolution Education and Outreach, 3, 506-613.

Nehm, R.H. (2010). Understanding undergraduate's problem solving processes. Journal of Biology and Microbiology Education, 1(2): 119-121.

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. & 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.E., and Ha, M. (in press). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. The American Biology Teacher

Opfer, J.E., Nehm, R.H., Ridgway, J.S., Mollohan, K., Perrin, E. (2011). Applying cognitive science to assessment of evolution education. Paper presented at the National Association for Research in Science Teaching, Orlando, FL. April 3-6.

Plake, B.S. (1980). Item arrangement and knowledge of arrangement on test scores. Journal of Experimental Education, 49, 56-58.

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.

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). Educational Testing Service, Princeton, N.J.