Exploring the Item Order Effect in a Geoscience Concept Inventory

Molly A. Undersander,[1] Richard M. Kettler,[2,a] and Marilyne Stains[1,a]

ABSTRACT

Concept inventories have been determined to be useful assessment tools for evaluating students' knowledge, particularly in the sciences. However, these assessment tools must be validated to reflect as accurately as possible students' understanding of concepts. One possible threat to this validation is what previous literature calls the item order effect: that the ordering of items on an assessment may affect how students perform. This study attempts to test the item order effect using different orderings of pictorial and verbal items on a geoscience concept inventory. Two different orderings of the same concept inventory were created and distributed to students in introductory geoscience courses via an online survey software (n = 432). The students' performance on the pictorial and verbal items was statistically analyzed. Seven interviews were also conducted to probe students about their preferred item format and order. The results revealed no predictable item order effect, though suggestive trends are present that warrant further study. Implications for educators and researchers are discussed. © *2017 National Association of Geoscience Teachers*. [DOI: 10.5408/16-235.1]

Key words: pictorial, verbal, item order, concept inventory

BACKGROUND

Development of a Concept Inventory

Concept inventories have been recognized as an effective tool for assessing student knowledge in science fields (McConnell et al., 2006; Libarkin, 2008). Currently concept inventories have been developed in physics, chemistry, astronomy, biology, and geoscience (Libarkin, 2008). This study recognizes a concept inventory as "a multiple-choice instrument designed to evaluate whether a person has an accurate and working knowledge of a concept or concepts" (a survey is considered to have five to seven items where an inventory has at least 20 items) as defined by Lindell, Peak, and Foster (2007, 14) in their study on the methodologies of concept inventory development. Their study revealed that the process of developing concept inventories lacks a universal system when it comes to aspects such as defining the concept domain, establishing test specifications (how an item will be represented), and establishing reliability and validity. Reliability and validity can be tested using many sources of psychometric and statistical evidence, and it is important that researchers continue to collect this evidence in order to develop inventories, which accurately reflect students' level of conceptual knowledge (Arjoon et al., 2013).

Validity has three main facets: construct, i.e., importance of the item for understanding the discipline; content, i.e., from an expert perspective, does the item measure understanding in the discipline?; and communication, i.e., does the test taker interpret the item the way the test developer intended? (Libarkin, 2008). Certain factors, such as item order, which are not directly related to the content can inadvertently make certain items easier or more difficult and therefore can affect the validity of a test (Messick, 1995). This would have implications for inventory design and the current study aims to address this possible threat.

The Geoscience Concept Inventory

The Geoscience Concept Inventory (GCI) was developed by testing 69 items with 5,000 students enrolled in 60 different geoscience courses at over 40 institutions across the United States (Libarkin et al., 2005; Libarkin and Anderson, 2005; Libarkin et al., 2011). Approximately 75 interviews were conducted and nearly 1,000 open-ended questionnaires were collected (Libarkin et al., 2005; Libarkin, 2014). The development team first identified alternate conceptions held by geoscience students and created test items, which addressed these alternate conceptions (Libarkin and Anderson, 2006a; Ward et al., 2010). These items were externally reviewed by researchers and educators and then presented to students via pilot testing, which included think-aloud interviews. The results of the pilot testing were analyzed using exploratory factor analysis, item response theory (Rasch analysis) and scale development theory; interviews were analyzed using grounded theory (Libarkin and Anderson, 2006b, "The Geoscience Concept Inventory"; Ward et al., 2010; Libarkin, 2014). Several cycles of revising, re-piloting, and re-analyzing resulted in the first version of the GCI (Ward et al., 2010). A GCI WebCenter exists online where all the validated items are compiled (Ward et al., 2010). WebCenter visitors can, and are encouraged to, submit new geoscience items for discussion as potential GCI items until they are validated. Community members are invited to contribute to the development of the GCI via reviewing GCI items, proposing new areas for development, becoming co-authors of the GCI, and using the GCI to address student learning (Libarkin et al., 2011). Currently, there are almost 200 validated items on the WebCenter, which educators can pick and choose from to use in their own classrooms. This "pick and choose" process was used to create the geoscience concept inventory instrument used in the current study.

Received 13 December 2016; revised 12 May 2017 and 19 May 2017; accepted 23 May 2017; published 7 August 2017.

[1]Department of Chemistry, University of Nebraska-Lincoln, 649a Hamilton Hall, Lincoln, Nebraska 68588, USA
[2]Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, 126 Bessey Hall, Lincoln, Nebraska 68588, USA
[a]Authors to whom correspondence should be addressed. Electronic mail: rkettler1@unl.edu; mstains2@unl.edu.

The Item Order Effect

The item order effect is defined as the idea that the order of items on a test or concept inventory could affect how students perform on successive items (adapted from Oldendick, 2008). This effect was first studied by Mollenkopf in 1950. He developed a power test (created with multiple types/skill level of items with the idea that if given an infinite amount of time, the test taker could achieve a perfect score) and a speeded test (created with trivially easy items aimed at completing as many items as possible in a restricted amount of time) version of a mathematics and verbal skills test; for each of the four types of test, he created different orderings of the items based on difficulty where certain items appeared earlier in one ordering and later in the other ordering (Mollenkopf, 1950; Mead and Drasgow, 1993). The tests were given to 382 high school students. The study found trends in the students' item-level scores depending on the item order and whether the test condition was power or speed (Mollenkopf, 1950).

Since then, many studies have been conducted to either prove or disprove the item order effect (Bradburn and Mason, 1964; Monk and Stallings, 1970; Dean, 1973; Crano, 1977; Plake, 1980; Hodson, 1984; Leary and Dorans, 1985; Balch, 1989; Gohmann and Spector, 1989; Carlson and Ostrosky, 1992; Coniam, 1993; Neely et al., 1994; Gray et al., 2002; Pettijohn and Sacco, 2007; Tal et al., 2008; Weinstein and Roediger III, 2012). Some researchers, like Mollenkopf, studied item order based on difficulty (Monk and Stallings, 1970; Coniam, 1993), while others looked at ordering according to content/chapter order versus random ordering of items (Hodson, 1984; Gohmann and Spector, 1989; Coniam, 1993). Studies have suggested that factors such as cognitive load required to answer items preceding the item of interest could explain the item order effect (Schroeder et al., 2012). Leary and Dorans (1985) as well as Undersander et al. (2016) conducted literature reviews of the major studies conducted to date about the item order effect. Both literature reviews concluded that in aggregate, there is inconclusive evidence as to whether or not this item order effect actually plays a role in student performance. In fact, there seems to be as many significant findings as insignificant findings.

Although one of the main concerns regarding the item order effect is student performance on assessment tools, researchers are also concerned about the impact of item order effect on the determination of a test's item difficulty, reliability, and validity (e.g., Monk and Stallings, 1970; Hodson, 1984; Balch, 1989; Carlson and Ostrosky, 1992; Coniam, 1993; Pettijohn and Sacco, 2007). So far, no significant results have been found to suggest that reliability and validity are predictably affected by item order (Monk and Stallings, 1970; Balch, 1989; Carlson and Ostrosky, 1992; Pettijohn and Sacco, 2007).

Some of the main limitations of these early studies, however, is that the researchers could not always enforce randomization of the test version each student received since the tests were delivered on a paper medium (tests were assigned alphabetically, or handed out as students walked through the door), nor could they enforce sequential completion as the test developer intended, which would invalidate any item order effect if students could jump around to items out of order (Dean, 1973; Plake, 1980; Balch, 1989). This is not to say that item order effects may be the product of how students take tests, but simply that if researchers are trying to test a specific ordering and students do not answer the items in that exact order, then the ordering cannot really be tested. Another limitation of these prior studies is the varying student demographics, in terms of student age and discipline studied. Some studies focused on high school students (Mollenkopf, 1950), while others focused on undergraduate students (Gohmann and Spector, 1989) and even adults in the work force (Bradburn and Mason, 1964). A third limitation which hinders the comparison of all of these studies is the wide range of disciplines studied. Disciplines studied include business and economics (Dean, 1973; Gohmann and Spector, 1989; Carlson and Ostrosky, 1992), chemistry (Hodson, 1984; Undersander et al., 2016), general social science (Crano, 1977), geography (Monk and Stallings, 1970), job related interviews (Bradburn and Mason, 1964), math (Mollenkopf, 1950; Leary and Dorans, 1985), physics (Gray et al., 2002), psychiatric nursing (Plake, 1980), psychology (Balch, 1989; Neely et al., 1994), and verbal skills (Mollenkopf, 1950; Coniam, 1993). The sentiment of Bradburn and Mason (1964) which states that until any two studies are studying the exact same ordering effect in the same discipline and context, the effect cannot be confidently generalized for other external situations still stands.

To summarize, the literature regarding the item order effect is plentiful, yet conflicted. While some studies claim that an item order effect exists, others claim that it does not, and therefore it is unclear whether multiple orderings of test items would have any predictable, significant effect on students' test scores or the validity of a concept inventory. The current study aims to address some of the identified limitations of previous studies by investigating differences in performance at the item level on two versions of a geoscience concept inventory; undergraduate college students in an introductory geoscience course took this inventory online to ensure true randomization and forced sequential completion.

Item Format: Pictorial versus Verbal

Extensive research has been conducted on the use of pictorial (visual) items and verbal (textual) items in student learning and assessment. For the purpose of this study, a visual/pictorial item is an item that includes any form of visual aid or visual representation (e.g., real-life picture, diagram, chart, graph, table, time-scale, map, cross section), often to accompany other text (LaDue et al., 2015). In contrast, a verbal/textual item is an item with no accompanying visual aid; students must rely solely on the words of the item and answer choices. This research has led to the identification of best practices for both types of items. For example, some of the best practices for minimizing exam fatigue and confusion when using verbal items include monitoring the length of items and word choice (Haláková and Prokša, 2007). For pictorial items, researchers suggest that the most important factors to consider is the length and word choice of captions, the balance of supplemental text (captions) to visuals, that the visuals actually be supplemental and relevant to the text, and that the content of visuals be attended to with greater importance than aspects such as color or realism (Haláková and Prokša, 2007; Phillips et al., 2010).

Having balanced text to accompany visuals is especially important in assessment. Because not all students use

pictures the same way, supporting text can be useful in communicating what the item developer wants the student to get out of the picture, and in fact visuals may be essentially useless without some amount of supporting text (Holliday, 1975; Mayer, 1989; Mayer and Anderson, 1991; Mayer et al., 1996; Angeli and Valanides, 2004; Crisp and Sweiry, 2006; Phillips et al., 2010; Kapıcı and Savas-cı–Açıkalın, 2015). To this end, having too much or too little supporting text can also be detrimental because it may reinforce alternate conceptions and incorrect mental models if the visual is not explained well or is made to seem more complicated than it is, in which case students may ignore the picture altogether and a "good picture" can actually be ruined (Weidenmann, 1989, 163; Mayer and Anderson, 1991; Schnotz and Bannert, 2003; Phillips et al., 2010).

A separate factor to consider is that students' background knowledge can also affect the usefulness of textual and visual assessment items. McNamara (1996) found that students who had a better grasp of the targeted concepts required less explicit text and could employ more gap-filling, whereas students at a lower proficiency in a certain concept require more explicit text in order to learn the material. Similarly, Duran and Balta (2014) found in their study that students who already tended to excel at science were not affected by the presence or absence of visual items, whereas students who did not tend to excel at science performed better when they had pictorial items compared to students who did not excel at science and saw only verbal items. For this reason, Duran and Balta (2014) suggest that while pictures generally seem beneficial for STEM learning (LaDue et al., 2015), it is not always necessary to include visuals in each assessment item, and more importantly all items should be piloted to assess their value.

Visuals in Geoscience Education

The use of visuals has been found to be especially useful in geoscience education, and the development of spatial skills is crucial for students to succeed in geoscience courses, especially in classes such as structural geology and classes that require three-dimensional (3D) thinking (Orion et al., 1997; Libarkin and Brick, 2002; LaDue et al., 2015). As it relates to geoscience, spatial thinking can be described as "(a) recognizing, observing, recording, describing, classifying, remembering, and communicating the two- or three-dimensional shapes, structures, orientations, and positions of objects, properties, or processes; (b) mentally manipulating those shapes, structures, orientations, and positions by rotation, translation, deformation, or partial removal; (c) making interpretations about why the objects, properties, or processes have those particular shapes, structures, orientations, and positions; (d) making predictions about the consequences or implications of the observed shapes, structures, orientations, and positions; and (e) using spatial thinking as a short cut or metaphor to think about the distribution of processes or properties across some dimension other than length-space" (Ishikawa and Kastens, 2005,184–186). Students in introductory level courses often come in with under-developed and varying levels of spatial skills (Ishikawa and Kastens, 2005; Titus and Horsman, 2009). Since researchers have found a correlation between students' spatial skill level and their success in geoscience courses, there has been a call to action to try to reinforce these skills by practicing them during class time (Orion et al., 1997; Gobert and Clement, 1999; Titus and Horsman, 2009; Liben and Titus, 2012). Titus and Horsman (2009) suggest that teachers use a pre-assessment at the beginning of the semester to determine their students' spatial skill level so that they can help their students learn the material most effectively.

This study investigates student performance on isomorphic items one presented solely verbally the other containing maps. The past research previously described suggests that students may be more challenged by the item containing maps.

The Chemistry Concept Inventory

The current study builds on a prior study in chemistry in which the item order effect within a concept inventory was also explored (Undersander et al., 2016). The study was developed as result of observations made during the development of a concept inventory about acid/base/solubility chemistry. In particular, the researchers noticed during think-aloud interviews that students answered differently depending on the order that certain pictorial (visual) and verbal (textual) items were presented to them. This led to an investigation of the existence of an item order effect with the pictorial and verbal items.

The study was first conducted at a Western institution and was later replicated at a university in the Midwestern region of the United States in order to address limitations with data collection at the Western institution. Both studies produced statistically insignificant results, though large opposite trends were seen in the two studies (Undersander et al., 2016). At the Western university, we saw a trend in which students who saw the verbal item first did better on the pictorial item than the students who saw the pictorial item first. Conversely, we found at the Midwestern university that students who saw the verbal item first did better on the verbal item than the students who saw the verbal item second, with no major trend seen among the pictorial items.

The current study aims to directly replicate the implementation of the Midwestern university to test for an item order effect of pictorial and verbal items in the context of a geoscience concept inventory.

RESEARCH QUESTION

The primary questions being researched are as follows:

1. How does item order affect student performance on conceptually isomorphic items when presented with pictorial and verbal versions of the item?
2. Why might this item order effect be present or absent?
3. To what extent is the item order effect generalizable across science disciplines?

METHODS

The methods for this study were adapted from Undersander et al., 2016 in order to compare findings in in chemistry to those reported here in geoscience.

Concept Inventory Instrument Design

The concept inventory used in this study consisted of 20 validated geoscience items taken from the Geoscience

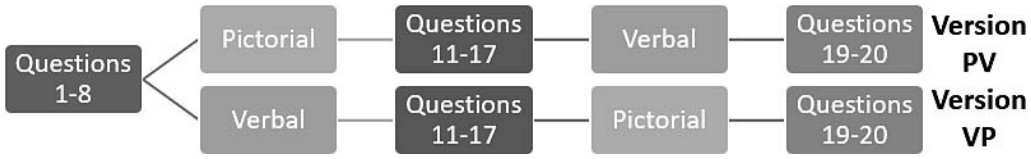

FIGURE 1: Schematic of item order for each version of the concept inventory. For each block of items, items remained in the same order for each version of the inventory. Each item was "named" after its order in version PV to make it easy to refer to the items as blocks. This is why even though Item 11 is the 10th item to appear on version VP, it is still referred to as Item 11 in Block 11–17.

Concept Inventory (Libarkin and Anderson, 2005). Of these 20 items, three addressed the same concept, namely the relationship between earthquakes, volcanoes, and plate tectonics; of these three items, two were presented in a form which used diagrams or maps (pictorial) and the other was presented purely in words (verbal). In order to test the item order effect, two versions of the inventory were created by switching the order of these pictorial and verbal items. In particular, in one version, the pictorial items appeared after eight control items followed by seven more control items and then the verbal item followed by two more control items. This is version PV. The other version had the position of the pictorial and verbal items switched so that the verbal item appeared after the first eight control items and the pictorial items appeared after the second seven control items. This is version VP. In both versions, the two pictorial items remained next to each other and in the same order, individually labeled as P1 and P2. The first eight control items always appeared in the same order, as did the second set of seven control items, and the last two control items. The two different versions are represented schematically in Fig. 1. Item P (P1 and P2) is shown in Figs. 2 and 3, respectively, and item V is shown in Fig. 4. Inventory version PV can be found in Supplementary Material A (available in the online journal and at http://dx.doi.org/10.5408/16-235s1).

Data Collection

Concept Inventory

This study was conducted using a mixed-methods exploratory design. The first part of the study involved quantitatively surveying students. Over the course of two semesters (fall and spring, successively), students who were enrolled in an introductory level geology course were asked to take a geoscience concept inventory in the form of an online survey delivered through Qualtrics online survey software. Students were asked to complete the inventory at the beginning and end of the semester for pre- and post-instruction data. The Qualtrics link randomized the version of the concept inventory that each student took and forced sequential completion; students could not return to previous items after they were completed. Students received the same inventory version at the end of the semester. Once the Qualtrics link was distributed to each class, the students had one week to complete the survey before it closed. Professors were requested to offer extra credit for completion of the inventory; however this was ultimately left to their discretion, and as a consequence, some offered it and some did not. When offered, the extra credit accounted for a small percentage of the total points available for the course.

Interviews

Qualitative data were collected through seven semi-structured think-aloud interviews conducted after the post-instruction concept inventory was distributed in both semesters. Students were given a paper medium of the same version of the concept inventory they took online. The use of multiple media was not a concern since Mead and Drasgow (1993) found no significant difference in performance when students take power tests on paper or online. Students were randomly chosen to be interviewed from a pool of students who were identified having answered their first item (P or V) correct on the online survey. For example,

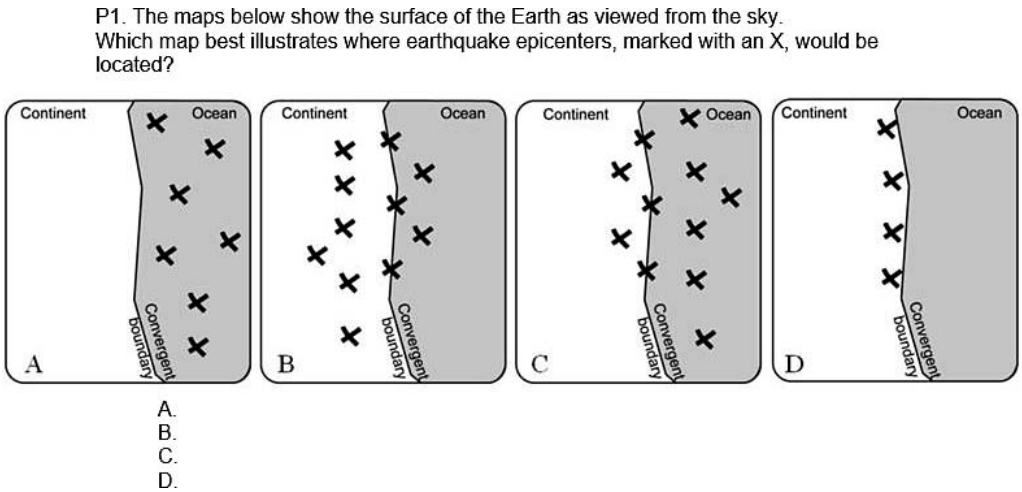

FIGURE 2: The pictorial item demonstrating the relationship between earthquakes and plate tectonics. The correct answer for item P1 is answer choice D.

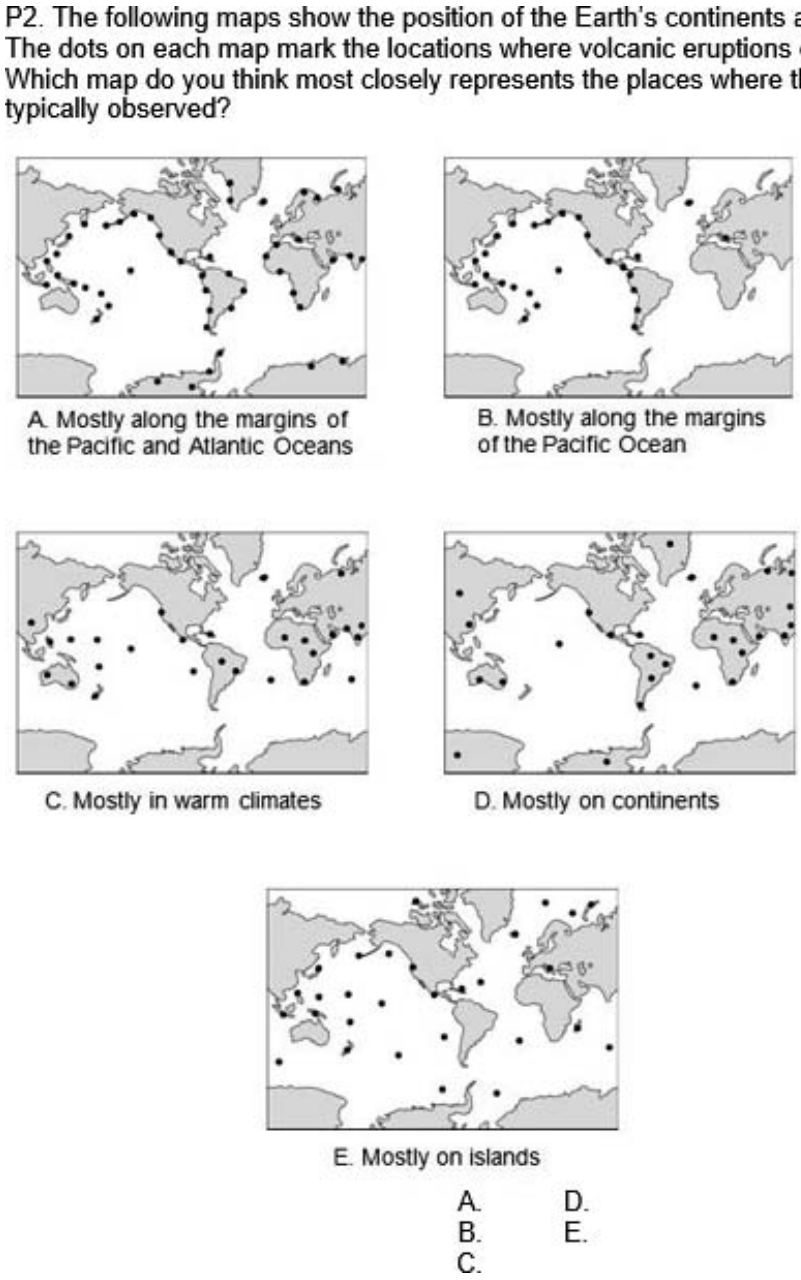

FIGURE 3: The pictorial item demonstrating the relationship between volcanoes and plate tectonics. The correct answer for item P2 is answer choice B.

someone who took version PV would have had to answer item P1 correctly to be included in the pool regardless of whether they answered item V correctly. Once this pool was formed, students were asked to volunteer to be interviewed with the incentive of a $15 gift card. Unfortunately, only seven students volunteered across the two semesters; we were thus not in a position to achieve saturation during the analysis of the interviews and the resulted emerging from them are thus preliminary. Students were asked to work out items P1, P2, V, and an arbitrary, unrelated item between P and V or vice versa to simulate the fact that the online medium had unrelated items separating the two types of conceptually isomorphic items. After working through the four items aloud, students were probed as to which version of the item (P or V) they preferred and which they thought

V. Which of the following responses best summarizes the relationship between volcanoes, large earthquakes, and tectonic plates?

A. Volcanoes typically occur on islands, earthquakes typically occur on continents, and both occur near tectonic plates
B. Volcanoes and large earthquakes both typically occur along the edges of tectonic plates
C. Volcanoes typically occur in the center of tectonic plates and large earthquakes typically occur along the edges of tectonic plates
D. Volcanoes and large earthquakes both typically occur in warm climates
E. Volcanoes, large earthquakes, and tectonic plates are not related, and each can occur in different places

FIGURE 4: The verbal item demonstrating the relationship between earthquakes, volcanoes, and plate tectonics. The correct answer for item V is answer choice B.

was more helpful to appear first on a concept inventory or other assessment tools to aid in answering successive items correctly. The interview protocol can be found in Supplementary Materials B (available in the online journal and at http://dx.doi.org/10.5408/16-235s2).

Student Population

A total of 487 pre- and post-surveys were collected. After cleaning the students' data, 432 data sets remained to be analyzed.

SPSS Statistics software was used to conduct *t*-tests in order to compare the students' total scores on the inventory between versions PV and VP as well as the students' total scores on the first eight control items between the two inventory versions. The first eight items that appeared on the concept inventory were universal in terms of order and content. Having these control items at the beginning of the concept inventory allowed us to compare and potentially control for students prior knowledge; this ensures that the students who took one version were not perhaps collectively more knowledgeable or otherwise dissimilar from the student population who took the other version of the concept inventory. After accounting for Type I and II error (Cunningham and McCrum-Gardner, 2007), the *t*-test results on the first eight items and on the concept inventory overall were statistically insignificant and thus the students who took the two versions were comparable in their content knowledge (see Supplementary Material C, Tables III–VI; available in the online journal and at http://dx.doi.org/10.5408/16-235s3).

Demographic data were collected at the end of the concept inventory as further means for determining that the two student groups were comparable between the PV and VP versions. Students were asked about their class standing, major, GPA, gender, whether the class was a repeat for them, and which lecture section they were in to determine whether an instructor effect existed. SPSS chi-square tests and Fisher 2×3 and 2×4 contingency tests were used to compare the various demographics. All tests came back statistically insignificant and all groups were determined to be demographically comparable (see Supplementary Material C, Tables VII–XVIII).

Data Cleaning and Analysis

Survey data were included in the analysis only if they met the following criteria: Students completed the inventory and self-reported not using outside resources as well as an effort level of 1, 2, or 3.

Before students began the online survey, they saw a screen that instructed them not to use any outside resources. Students who self-reported using outside resources at the end of the online survey were removed from the data pool.

Similarly, at the end of the survey students were asked to self-report their effort level on a scale of 1 to 4 with 1 being "I gave it my best effort" and 4 being "I didn't take it too seriously." Students who answered 4 were completely removed from the data pool in order to eliminate data that may be the result of pure guessing. The remaining students' data were then split into two populations. The first population includes students who answered with an effort level of 1, 2, or 3, 2 being "I gave my best effort for most of the questions, but not all" and 3 being "I tried on the ones I thought I knew, but didn't work too hard at the others." This will be referred to as the "moderate effort population." Our second population consists of students who self-reported an effort level of 1 or 2, which will be referred to as the "high effort population."

Finally, only data sets in which students completed all 20 inventory items were used. Unfinished submissions were disregarded, even if the target items had been answered. Had any cases of zero variance been found, they would have been removed; however, no such cases were found.

Because the concept inventory was administered online, only minimal data cleaning could be performed based on the amount of time it took students to complete the survey since Qualtrics only reports a start and stop time stamp as opposed to total time spent on the concept inventory. Some responses could be eliminated if the timestamps indicated an unreasonably short time spent from opening the survey to the submission of the survey. This was the case for very few submissions. On the other end of the spectrum, timestamps that showed the survey being open for multiple days could not be disregarded since we cannot ascertain the exact amount of time the student spent thinking about the items compared to the amount of time the survey was simply left open on the student's computer for them to return to. This emulates true power testing conditions.

The concept inventory data (i.e., individual item scores) were analyzed using SPSS statistics software. This software was used to perform chi-square tests based on whether the three items of interest were answered correctly or incorrectly. The Bonferroni correction was applied to all statistical data to account for Type I error, rejecting the null hypothesis when it is actually true, which resulted in a new level of significance (0.008; Cunningham and McCrum-Gardner, 2007). This decrease in threshold reduces the probability of Type I error, but increases the probability of Type II error which is that the null-hypothesis would not be rejected even if proven false (Cunningham and McCrum-Gardner, 2007). Statistical power accounts for Type II error by determining the probability of getting statistically significant results (O'Keefe, 2007). G*Power 3.1 software was used to calculate the statistical power of each test ($1-\beta$). The ideal statistical power to indicate a sufficient sample size is 0.8. When this statistical power is achieved, results can be confidently reported as an accurate reflection of the sample size. The lower the value for $1-\beta$ is relative to the desired value, the more likely it is that the sample size is not large enough to accurately reflect whether the reported two-tailed p values are actually significant regardless of the corrected threshold for determining significance.

Transcriptions of the seven interviews were coded for common themes. The main codes used included whether the students preferred to see the pictorial or verbal format first, whether they had no opinion, and why they held the opinion regarding item preferences. They were also coded based on whether the student tried to apply the first item they saw to the successive counterpart items.

RESULTS

Concept Inventory Results

Table I shows the statistical results for the percent of students who answered each of the three targeted items correctly.

TABLE I: Results for concept inventory. Percentages reflect number of students who correctly answered each item.

Pre-Instruction, Moderate Effort						Pre-Instruction, High Effort					
Item	*Version*		*Significance*			*Item*	*Version*		*Significance*		
	PV (n=121)	*VP (n=124)*	*p value*	***Φ***	***1-β***		*PV (n=101)*	*VP (n=107)*	*p value*	***Φ***	***1-β***
P1	35.5%	33.9%	0.784	-0.018	0.139	P1	40.6%	37.4%	0.635	-0.033	0.113
P2	28.9%	29.8%	0.875	0.010	0.139	P2	31.7%	29.9%	0.781	-0.019	0.113
V	43.8%	56.5%	0.048	0.126	0.139	V	47.5%	60.7%	0.056	0.133	0.113
Post-Instruction, Moderate Effort						**Post- Instruction, High Effort**					
Item	*Version*		*Significance*			*Item*	*Version*		*Significance*		
	PV (n=94)	*VP (n=93)*	*p value*	***Φ***	***1-β***		*PV (n=84)*	*VP (n=85)*	*p value*	***Φ***	***1-β***
P1	26.6%	33.3%	0.315	0.074	0.099	P1	26.2%	32.9%	0.336	0.074	0. 088
P2	31.9%	33.3%	0.836	0.015	0.099	P2	33.3%	35.3%	0.788	0.021	0. 088
V	56.4%	67.7%	0.110	0.117	0.099	V	58.3%	68.2%	0.182	0.103	0.088

In the pre-instruction concept inventory, accounting for the Bonferroni correction, no statistical significance was found in either the high effort or moderate effort populations for any of the three items [moderate effort P1: $\chi^2(1, N = 245, p = 0.784, \varphi = -0.018)$; moderate effort P2: $\chi^2(1, N = 245, p = 0.875, \varphi = 0.010)$; moderate effort V: $\chi^2(1, N = 245, p = 0.048, \varphi = 0.126)$; high effort P1: $\chi^2(1, N = 208, p = 0.635, \varphi = -0.033)$; high effort P2: $\chi^2(1, N = 208, p = 0.781, \varphi = -0.019)$; high effort V: $\chi^2(1, N = 208, p = 0.056, \varphi = 0.133)$]. In both effort populations, students performed approximately 3% better on item P1 when they saw this item first (version PV). Students performed almost identically on item P2 between the two effort populations and between the inventory versions. Students performed approximately 13% better on item V when they had version VP compared with their PV counterparts. Between the high and moderate effort populations, the students from the high effort population performed an average of 3.2% better collectively on all three items than the moderate effort population.

We hypothesized that the item order effect may be more present at the end of the semester, once students had learned the targeted content. However, in the post-instruction inventory, there was also no significance found in either population for any of the three items [moderate effort P1: $\chi^2(1, N = 187, p = 0.315, \varphi = 0.074)$; moderate effort P2: $\chi^2(1, N = 187, p = 0.836, \varphi = 0.015)$; moderate effort V: $\chi^2(1, N = 187, p = 0.110, \varphi = 0.117)$; high effort P1: $\chi^2(1, N = 169, p = 0.336, \varphi = 0.074)$; high effort P2: $\chi^2(1, N = 169, p = 0.788, \varphi = 0.021)$; high effort V: $\chi^2(1, N = 169, p = 0.182, \varphi = 0.103)$]. Both populations performed similarly on the concept inventory with a 6.7% difference between the two versions; the VP version was favored as opposed to the PV version, unlike the pre-instruction results. Both effort populations, students also performed almost identically on item P1 and P2. Similar to the pre-instruction inventory, item V had a relatively large difference between the two versions in both the high and moderate effort populations with an 11.3% difference in the moderate effort population and a 9.9% difference in the high effort population, both in favor of version VP. That is, students performed approximately 9%–13% better on item V when they had to complete V after seeing item P, although this is not statistically significant.

It is important to note that while trends are seen at the item variation level, t-tests performed on the total scores yielded clearly insignificant *p* values ($p = 0.037$ for all groups, see Supplementary Material C, Table VI). If this study were determining the presence of item-order effect based on total score alone, it would appear as if an item order effect does not exist. The complete absence of any statistical significance at the item variation level would also suggest that an item order effect does not exist. However, given that the statistical power (1-β) was medium (lower than ideal) where we saw possible trends emerging in all 4 tests on item V and in item P1 in the post-inventory, this could indicate that our sample sizes were simply not large enough to reflect any significance. Therefore these trends suggest that an item order effect may be present, though not visibly, based on our item variation statistical analysis. This shows that it is possible to see an item order effect in one level of analysis, but it could be hidden in the other.

Interview Results

Table II shows the results of the think-aloud interviews conducted to explore seven students' thoughts on item format and item order effect. Almost all of the students (n = 5/7) recognized the relationship between the pictorial and verbal items. One student felt that s/he did not connect the two items because of the different learning styles associated with the items (verbal versus visual). The other student who did not feel a strong connection between the items as the interview progressed identified that there was a relationship, but because the student knew the material so well they felt that previous items may have only affected successive items on a subconscious level.

A majority of the students (n = 5/7) indicated that they preferred seeing the verbal item first because they found it easier or found it helpful as a primer for the item's counterpart:

> *"[V] kind of like embeds the concept in your mind more. Also with the pictures you're kind of just like trying to relate them, you're not actually thinking like...volcanoes they're along plate boundaries."*

> *"I [could think] through all the answers more clearly because then I wouldn't have any picture in my head already to kind of like base my answers, so I would just be going off myself."*

TABLE II: Results from the think-aloud interviews based on which inventory version the students took with supporting quotes from the interviewees.

Interview Question	Choice Provided by Interviewee	Interviewee n = 7		Example of Quote Providing Justification for Choice
		PV	VP	
Did you prefer seeing the [verbal/pictorial] question first? Did it matter to you?	Pictorial	2	0	"I kind of like [P1 and P2] first because you're thinking about volcanoes and earthquakes separately and then [V] brings them both together and it's like oh I remember back on those diagrams they do both line up along the plate boundaries and they're similar so I do like this way better having [P1 and P2] first; it makes you think a little bit more rather than okay I answered [V] this way so these two should be this because of that."
	Verbal	2	3	"I like seeing the verbal one first just because it kind of like embeds the concept in your mind more. . .then you can apply it [to P] rather than [try] applying it [to P first] and not really thinking about the concept as much."
As you moved to each successive question, were you thinking about previous questions to help you, or were they just separated in your mind?	Yes	3	2	"I think this puts a more permanent idea in my head of what my answer would've been. . . so I don't think I would think too deeply into these [following] questions like kind of just going through blindly. . . [the first question] definitely affects my answers."
	No	0	1	". . .when I was reading through [V] I was focusing on which answered fit best and then we have diagrams it's like visual mode of thinking. . ."
	Subconscious	1	0	"I did make that correlation but I wouldn't say I explicitly used that for [V]."

The two students (n = 2/7) who said they preferred to see item P first had this to say about the item:

> *"I kind of like [P1 and P2] first because you're thinking about volcanoes and earthquakes separately and then [V] brings them both together and it's like oh I remember back on those diagrams they do both line up along the plate boundaries and they're similar so I do like this way better having [P1 and P2] first; it makes you think a little bit more rather than okay I answered [V] this way so these two should be this because of that."*

Contrary to the students who liked being able to form their own visualization first, one of the students who preferred seeing the pictorial items first liked having a visual to go off of because with the verbal item *"you have to visualize it in your head. . .I think that [P2] is personally easier than [V] because you can look and see where there's not a volcano in the Sahara so I know these three [answer choices] are out."*

Despite the students' preferences, the only two students who answered all three items correctly were the two students who preferred having the pictorial item first, who happened to both have version PV. All three of the students who had version VP who also preferred seeing V first, got one or both of the pictorial items incorrect after answering item V. Of the students who had version PV but said they would have preferred to have the verbal item first, one got all three items incorrect and one got P2 incorrect after answering item P1 first.

DISCUSSION AND CONCLUSION

Geoscience Concept Inventory

In this study, item order effect was examined using a geoscience concept inventory comprised of validated items from the GCI (Libarkin and Anderson, 2005), followed by think-aloud interviews with students from introductory geoscience courses. The analysis of the quantitative data yielded no statistically significant differences among any of the populations. The data do expose some trends, which may be proven significant if this study were to be replicated with a larger sample size to achieve higher statistical power. This supports the findings of several previous studies, which determined that item order effect, though sometimes appearing to be present, cannot be identified with any level of predictability or certainty (Bradburn and Mason, 1964; Monk and Stallings, 1970; Dean, 1973; Crano, 1977; Plake, 1980; Hodson, 1984; Leary and Dorans, 1985; Balch, 1989; Gohmann and Spector, 1989; Carlson and Ostrosky, 1992; Coniam, 1993; Neely, et al., 1994; Gray, et al., 2002;

Pettijohn and Sacco, 2007; Tal, et al., 2008; Weinstein and Roediger III, 2012).

In the semistructured interviews, students were probed on their opinion of which item format they thought was more helpful to appear first as a primer for successive items. Although a majority of the students stated that they believed the verbal item to be more useful as a primer, which corroborates the statistically insignificant trends in the data analyses, the overall results of the interviews, namely how students performed on the items in comparison to their preference reasoning, do not provide enough evidence to make a ruling on the presence of an item order effect.

Comparison to Chemistry Concept Inventory

Previously, Undersander et al. (2016) conducted a nearly identical experiment using a 20-item acid/base/solubility concept inventory in general chemistry and organic chemistry classes at the same institution as this study. Similar results as those presented here were found: the quantitative inventory results ($n = 365$) yielded no statistically significant data, although the power of the test was also a limitation. Interestingly, both studies produced similar trends: students who received version VP had a much higher performance on item V than did students who received version PV. The interview results ($n = 19$) were more varied; of the students who had a preference, three preferred the pictorial item first, four preferred the verbal item first, 10 did not have a preference, and two were unidentified.

One goal of this study was to test pictorial and verbal item order effect between different disciplines. While the disciplines of chemistry and geoscience appear to yield similar results, suggesting that there is an interdisciplinary lack of predictability of the item order effect, two studies are not enough to close the argument. Further studies should be conducted in other STEM (science, technology, engineering, and mathematics) disciplines as well as other non-STEM disciplines, which heavily use both pictorial and verbal item formats for learning and assessment.

Because these identical studies were conducted in two different science disciplines the similarly insignificant results suggest that the presence of item order effect can be rejected more definitively, at least in the realm of pictorial and verbal items. It will be difficult to develop a decisive conclusion on whether or not the item order effect actually exists and should be taken into account until studies start being replicated across various disciplines using the same ordering factor and likewise within the same discipline using different ordering factors. Despite the statistical insignificance, since both studies saw the same trends regarding item V, it is important to continue investigating this trend with larger sample sizes.

Implications

Since Mollenkopf's original study on item order effect in 1950, the literature has not been able to agree on the definitive predictability of the item order effect. The current study in congruence with Undersander et al. (2016) demonstrates that teachers should be able to generate multiple versions of tests and other assessments for purposes such as preventing cheating without the fear of negatively impacting any group of students, though educators are cautioned to monitor any trends that may be present in their students' scores.

Although teachers are not expected to observe any performance discrepancy among their students, researchers are cautioned to closely monitor any potential ordering effects when developing assessment tools. The observed trends in the data, while statistically insignificant and therefore unpredictable, indicate that an ordering effect could be present which would affect the validity, reliability, and item difficulty during psychometric analyses. Therefore, multiple orderings of assessment items should be implemented across a wide range of demographics when developing assessment tools for research purposes.

It may be possible that certain concepts are more susceptible to demonstrating an item order effect than others, as the low performance of students on both versions of the concept inventory on items P1 and P2 indicates. The concept tested in this study was centered on plate tectonics, more specifically the relationship between earthquakes, volcanoes, and their location on tectonic plates. Although students struggle with many geoscience concepts, plate tectonics seems to be one of the most prevalent where students have many alternate conceptions (Clark et al., 2011). Libarkin et al. (2005) found that there is a large disconnect between students' understanding of the location of earthquakes and volcanoes. They also found that students were using correct terminology to describe incorrect concepts, so the use of correct terminology doesn't necessarily imply student understanding. Even after receiving instruction, students tend to hold onto their alternate conceptions, and often visuals can reinforce these alternate conceptions if not used properly for instruction and assessment (Clark et al., 2011). Finally, items P1 and P2 may have had the lower performance level on both versions of the inventory due to students' difficulties in interpreting maps (Ishikawa and Kastens, 2005; Titus and Horsman, 2009). Instructors are encouraged to address students' alternate conceptions during instruction through the proper use of visuals in order to eliminate any confusion that may result when a student sees a visual on an assessment.

In terms of the statistical analysis, this study only took the statistical variation of each individual item (P1, P2, and V) into account and used the total score variation as a means to prove that the student populations were comparable via independent *t*-test. This in an important factor to distinguish in item order effect studies because an item order effect may be hidden in one type of statistic variation, but present in the other. In the current study, trends were seen which suggest an item order effect may exist at the item score level, but if the total score variation had been taken into account, or had been solely used instead of the item statistic variation, this would have strengthened the case against the presence of an item order effect. One type of future study that would add to the item order effect literature is one that would compare the statistical analysis of both the individual item score variation and total score variation, provided an acceptable statistical power is achieved.

LIMITATIONS

This study was conducted to address limitations identified in the study conducted by Undersander et al. (2016), yet this study still has limitations of its own.

As was seen in the chemistry study, the sample size for this study was smaller than ideal resulting in a small

statistical power which rendered the statistical analysis void in the sense of predictive ability. Despite this fact, trends are seen that suggest that an item order effect may still be present. We thus recommend that the study be reproduced using much larger sample sizes.

Because the focus of this study was to test students' performance on conceptually isomorphic items presented in visual and textual forms, the emphasis was on the content of the item rather than the actual type of visual used. For this reason, students were not probed during the interviews about their comfort level with the particular type of visual used, namely maps. There is a wide range of types of visuals utilized in geoscience instruction, and students could perform differently on these different types of visuals depending on the amount of instruction they receive for each. This could also affect students' preferences for either seeing the pictorial or verbal item first; if they don't feel comfortable with maps, they may say they prefer to see the verbal item first, but if they love working with cross-sections, they may say they prefer seeing the pictorial item first. Further studies should be conducted testing for orderings with various types of visuals (maps, 3D diagrams, cross-sections, timescales, real pictures, etc.).

One possible limitation was brought to light during an interview when the student commented on the fact that in the two maps which were used for P1 and P2, the colors of the land and ocean are reversed from each other (Phillips et al., 2010). This was originally not considered as a weight added to cognitive load since each item had been individually validated (Libarkin and Anderson, 2005). However, by putting these two items together, the validity and reliability of the items may be altered and should be tested if this instrument were to become standardized.

Students' comfort and skill level with certain visuals may also play into the amount of cognitive load and fatigue they experience, also affecting the performance scores on the individual items (Yen, 1993). For example, if students are uncomfortable working with a certain visual like maps, they may feel discouraged, which would impact their performance on following items. Similarly, the fact that items P and V were separated by several other items could also contribute to cognitive load and fatigue. The general location of items was not assessed for the role that location can play in priming and cognitive load. If this concept inventory were to be validated as a whole, it would be important not only to test the use of various types of visuals and their orderings with textual items, but to also test various versions of the concept inventory in which the spacing of related items is tested as well as the ordering of items related to difficulty (easy to hard/hard to easy) to see what effect these factors have on priming and cognitive load.

Some studies suggest that other ordering factors such as answer choice order could have a significant impact on student performance in terms of creating more cognitive load for students as they progress through an assessment. Schroeder et al. (2012) performed a study in which they investigated factors that affected student performance on the American Chemical Society. Exams. While they, along with Tellinghuisen and Sulikowski (2008), agree that answer choice order seems to have a significant impact on student performance, Schroeder et al. also noted that with the ACS exams it is impossible to study either answer choice order effect or item order effect independently. Therefore, there is a crucial need to further study which factors affect student performance the most when assessment tools are being developed. Because the item order factor could be isolated with the current instrument, answer choice order was not tested or considered in this study; however, it would be beneficial to recreate the study to test for answer choice order effects and hold the item order constant.

ACKNOWLEDGMENTS

We would like to acknowledge the University of Nebraska-Lincoln's Undergraduate Creative Activity and Research Experience (UCARE) Grant for partial funding of this project. We would also like to thank all the faculty and students whose cooperation and participation made this project possible.

REFERENCES

Angeli, C., and Valanides, N. 2004. Examining the effects of text-only and text-and-visual instructional materials on the achievement of field-dependent and field-independent learners during problem-solving with modeling software. *Educational Technology Research and Development*, 52:23–36.

Arjoon, J., Xu, X., and Lewis, J. 2013. Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education*, 90:536–545.

Balch, W. 1989. Item order affects performance on multiple-choice exams. *Teaching of Psychology*, 16:75–77.

Bradburn, N., and Mason, W. 1964. The effect of question order on responses. *Journal of Marketing Research*, 1:57–61.

Carlson, J., and Ostrosky, A. 1992. Item sequence and student performance on multiple-choice exams: Further evidence. *The Journal of Economic Education*, 23:232–235.

Clark, S., Libarkin, J., Kortz, K., and Jordan, S. 2011. Alternative conceptions of plate tectonics held by nonscience undergraduates. *Journal of Geoscience Education*, 59:251–262.

Coniam, D. 1993. Does the ordering of questions on a test affect student performance? *Education Research Journal*, 8:74–78.

Crano, W. 1977. Primacy versus recency in retention of information and opinion change. *The Journal of Social Psychology*, 101:87–96.

Crisp, V., and Sweiry, E. 2006. Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, 48:139–154.

Cunningham, J., and McCrum-Gardner, E. 2007. Power, effect and sample size using GPower: Practical issues for researchers and members of research ethics committees. *Evidence Based Midwifery*, 5:132–136.

Dean, M. 1973. The impact of exam question order effects on student evaluations. *The Journal of Psychology*, 85:245–248.

Duran, M., and Balta, N. 2014. The influence of figured and non-figured questions on secondary students' success at science exams. *Pakistan Journal of Statistics*, 30:1279–1288.

Gobert, J., and Clement, J. 1999. Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching*, 36:39–53.

Gohmann, S., and Spector, L. 1989. Test scrambling and student performance. *The Journal of Economic Education*, 20: 235–238.

Gray, K., Rebello, S., and Zollman, D. 2002. The effect of question order on responses to multiple-choice questions, *in* Physics Education Research Conference, American Institute of Physics, p. 1–4.

Haláková, Z., and Prokša, M. 2007. Two kinds of conceptual

problems in chemistry teaching. *Journal of Chemical Education*, 84:172–174.
Hodson, D. 1984. The effect of changes in item sequence on student performance in a multiple-choice chemistry test. *Journal of Research in Science Teaching*, 21:489–495.
Holliday, W. 1975. The effects of verbal and adjunct pictorial-verbal information in science instruction. *Journal of Research in Science Teaching*, 12:77–83.
Ishikawa, T. and Kastens, K. 2005. Why some students have trouble with maps and other spatial representations. *Journal of Geoscience Education*, 53:184–197.
Kapıcı, H., and Savaşcı-Açıkalın, F. 2015. Examination of visuals about the particulate nature of matter in Turkish middle school science textbooks. *Chemical Education Research and Practice*, 16:518–536.
LaDue, N., Libarkin, J., and Thomas, S. 2015. Visual representations on high school biology, chemistry, earth science, and physics assessments. *Journal of Science Education and Technology*, 24:818–834.
Leary, L., and Dorans, N. 1985. Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55:387–413.
Libarkin, J. 2008. Concept inventories in higher education science. National Research Council. Available at http://www7.nationalacademies.org/bose/Libarkin_CommissionedPaper.pdf (accessed 1 September 2016).
Libarkin, J. 2014. Research related to understanding. Available at https://geocognitionresearchlaboratory.wordpress.com/research-in-the-grl/research-related-to-understanding/ (accessed 1 September 2016).
Libarkin, J., and Anderson, S. 2005. Assessment of learning in entry-level geoscience courses: Results from the geoscience concept inventory. *Journal of Geoscience Education*, 53:394–401.
Libarkin, J., and Anderson, S. 2006a, Development of the geoscience concept inventory: Proceedings of the National STEM Assessment Conference, Washington DC, p. 148–158.
Libarkin, J., and Anderson, S., 2006b, The Geoscience Concept Inventory: Application of Rasch Analysis to Concept Inventory Development in Higher Education, *in* Liu, X. and Boone, W., Applications of Rasch Measurement in Science Education, Maple Grove, MN, JAM Press, p. 45–73.
Libarkin, J., Anderson, S., Science, J., Beilfuss, M., and Boone, W. 2005. Qualitative analysis of college students' ideas about the Earth: Interviews and open-ended questionnaires. *Journal of Geoscience Education*, 53:17–26.
Libarkin, J., and Brick, C., 2002. Research methodologies in science education: Visualization and the geoscience. *Journal of Geoscience Education*, 50:449–455, doi: 10.5408/1089-9995-50.4.449.
Libarkin, J., Ward, E., Anderson, S., Kortemeyer, G., and Raeburn, S. 2011. Revisiting the Geoscience Concept Inventory: A call to the community. *GSA Today*, 21:26–28.
Liben, L., and Titus, S. 2012. The importance of spatial thinking for geoscience education: Insights from the crossroads of geoscience and cognitive science. *The Geological Society of America*, Special Paper 486:51–70.
Lindell, R., Peak, E., and Foster, T. 2007. Are they all created equal? A comparison of different concept inventory development methodologies, *in* Physics Education Research Conference, American Institute of Physics, p. 14–17.
Mayer, R., and Anderson, R. 1991. Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology*, 83:484–490.
Mayer, R. 1989. Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology*, 81:240–246.
Mayer, R., Bove, W., Bryman, A., Mars, R., and Tapangco, L. 1996. When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology*, 88:64–73.
McConnell, D., Steer, D., Owens, K., Knott, J., Van Horn, S., Borowski, W., Dick, J., Foos, A., Malone, M., McGrew, H., Greer, L., and Heaney, P. 2006. Using conceptests to assess and improve student conceptual understanding in introductory geoscience courses. *Journal of Geoscience Education*, 54:61–68.
McNamara, D., Kintsch, E., Songer, N., and Kintsch, W. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14:1–43.
Mead, A. and Drasgow, F. 1993. Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:449–458.
Messick, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50:741–749.
Mollenkopf, W. 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., and Stallings, W. 1970. Effects of item order on test scores. *Journal of Educational Research*, 63:463–465.
Neely, D., Springston, F., and McCann, S. 1994. Does item order affect performance on multiple-choice exams? *Teaching of Psychology*, 21:44–45.
O'Keefe, D. 2007. Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, 1:291–299.
Oldendick R. 2008. Question order effect. Available at http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n428.xml (accessed 14 December 2015).
Orion, N., Ben-Chaim, D., and Kali, Y. 1997. Relationship between earth-science education and spatial visualization. *Journal of Geoscience Education*, 45:129–132.
Pettijohn II, T., and Sacco, M. 2007. Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology*, 34:142–149.
Phillips, L., Norris, S., and Macnab, J. 2010. Visualization in mathematics, reading and science education. Dordrecht, the Netherlands: Springer Science and Business Media.
Plake, B. 1980. Item arrangement and knowledge of arrangement on test scores. *The Journal of Experimental Education*, 49:56–58.
Schnotz, W., and Bannert, M. 2003. Construction and interference in learning from multiple representation. *Learning and Instruction*, 13:141–156.
Schroeder J., Murphy K,. and Holme T., 2012, Investigating factors that influence item performance on ACS Exams. *Journal of Chemical Education*, 89:346–350.
Tal, I., Akers, K., and Hodge, G. 2008, Effect of paper color and question order on exam performance. *Teaching of Psychology*, 35:26–28.
Tellinghuisen J.. and Sulikowski M. M. 2008. Does the answer order matter on multiple-choice exams? *Journal of Chemical Education*, 85:572.
Titus, S., and Horsman, E. 2009. Characterizing and improving spatial visualization skills. *Journal of Geoscience Education*, 57:242–254.
Undersander, M., Lund, T., Langdon, L., and Stains, M. 2017. Probing question order effect while developing a chemistry concept inventory. *Chemical Education Research and Practice,* 18:45–54.
Ward, E., Libarkin, J., Raeburn, S., and Kortemeyer, G. 2010. The Geoscience Concept Inventory WebCenter provides

new means for student assessment. *eLearningPapers,* 20:1–14

Weidenmann, B. 1989. When good pictures fail: An information-processing approach to the effect of illustrations. *Advances in Psychology*, 58:157–170.

Weinstein, Y., and Roediger, H. 2012. The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition*, 40:727–735.

Yen, W. 1993. Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30:187–213.