Chemistry Education Research and Practice

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c5rp00228a

Probing the question order effect while developing a chemistry concept inventory

Molly A. Undersander,[a] Travis J. Lund,*[b] Laurie S. Langdon*[c] and Marilyne Stains*[a]

The design of assessment tools is critical to accurately evaluate students' understanding of chemistry. Although extensive research has been conducted on various aspects of assessment tool design, few studies in chemistry have focused on the impact of the order in which questions are presented to students on the measurement of students' understanding and students' performance. This potential impact has been labeled the question order effect in other literature and may be considered as a threat to the construct validity of the assessment tool. The set of studies described in this article tested whether question order effects were present within a concept inventory on acid-based chemistry. In particular, we tested whether the order of two conceptually isomorphic questions, one pictorial and one verbal, affected students' performance on the concept inventory. Two different versions of the inventory were developed and collected from students enrolled in the second semester of first-year university chemistry courses (general chemistry; N = 774) at two different institutions and to students enrolled in the first semester of organic chemistry (N = 163) at one of the two institutions. Students were further divided in two groups based on their self-reported level of effort in answering the concept inventory. Interviews were also conducted with a total of 19 students at various stages of the studies. Analyses of differences in students' responses to the two versions of the inventory revealed no question order effect in all settings. Implications for instructors and researchers are provided.

Received 21st December 2015,
Accepted 27th September 2016

DOI: 10.1039/c5rp00228a

www.rsc.org/cerp

Introduction

Concept inventories are increasingly used as one measure of student learning by instructors, departments, and institutions as they seek evidence for effectiveness of course reform efforts. The first major concept inventory to be developed was the Force Concept Inventory (FCI) in physics (Hestenes *et al.*, 1992), and it has been widely used to demonstrate that students learn more in courses that employ interactive learning strategies (Hake, 1998). The current study emerged from a context in which faculty at a large research university in the Rocky Mountain region of the United States were engaged in establishing learning goals, implementing research-based instructional strategies, and attempting to measure the effects of these course transformations within the first-year university chemistry (General Chemistry, GC). The Chemistry Concept Inventory (Mulford and Robinson, 2002) was adapted for use in the first semester of the GC course, but at the time, no concept inventories had been published that faculty found suitable in measuring concepts of interest in the second semester of the GC course. Thus, a 20-item concept inventory targeting solubility and acid–base concepts was developed and tested for this purpose. Faculty involved in this initiative valued student conceptual understanding as well as their abilities to think about chemistry at the particulate level. As part of the larger course reform efforts, new homework and recitation activities were developed that required students to interpret and produce their own particulate-level representations. Thus, development of this new concept inventory included questions that incorporated these pictures.

In developing the concept inventory, efforts were made to ask several questions related to each sub-concept of interest. One pair of questions, and the focus of this study, asked students to consider relationships among strength, concentration, and pH of acidic solutions. As shown in Fig. 1, one question (labelled P, pictorial) required students to interpret molecular-level representations of acidic solutions, and one question was based on text (verbal, V).

Mid-semester cognitive interviews, in which students thought aloud as they worked through the entire 20-item instrument, revealed that some students seemed to approach the molecular-level "picture questions" differently than other questions, leading investigators to wonder whether encountering the molecular-level picture question (P) early in the inventory might

[a] Department of Chemistry, University of Nebraska-Lincoln, Lincoln, USA. E-mail: mstains2@unl.edu
[b] Department of Natural Sciences, Oregon Institute of Technology, USA. E-mail: Travis.Lund@oit.edu
[c] School of Education, University of Colorado, Boulder, USA. E-mail: laurie.langdon@colorado.edu

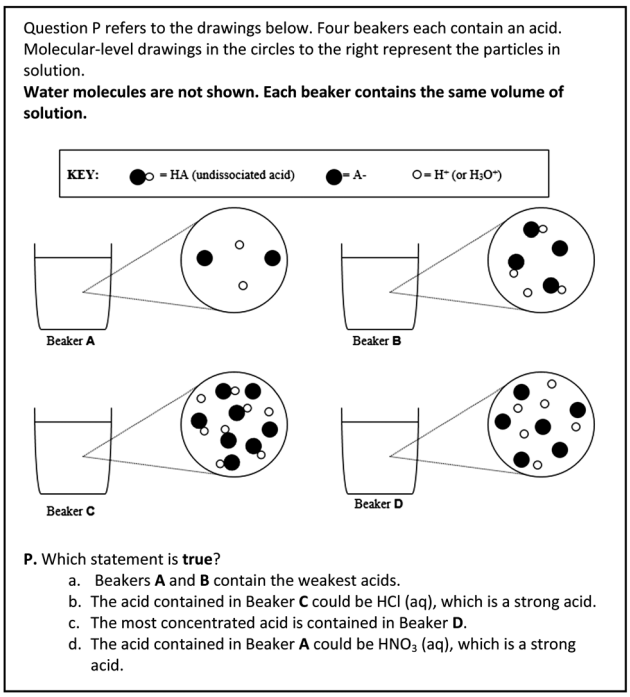

Fig. 1 Two questions were constructed to assess student knowledge of acid strength, concentration, and pH relationships. Question P requires students to interpret a particulate-level diagram. Question V is verbal, *i.e.* based on text.

help students with the verbal question (V) later in the inventory, or *vice versa*. Thus two slightly different versions of the instrument were administered at the end of the semester, in which the pictorial and verbal items related to acid strength/concentration/pH were reordered to probe a possible question order effect. Initial analyses of the post-assessment indicated that students performed better on the verbal question when it was ordered after the pictorial question. Due to some limitations with how the post-assessment was administered, we repeated the study at a different institution with an online version of the instrument. Analyses of results from both institutions will be presented below. The overall goals of this study are to examine whether altering the order of pictorial and verbal items related to acid concentration/strength/pH relationships affects student performance on those particular items, and if so, to also determine the extent to which these effects are consistent across institutional and instrument administration contexts.

Research question

The primary research questions being investigated are: How does question order affect student performance on conceptually isomorphic questions when students are presented with pictorial and verbal versions of the questions, and why are students impacted differently by this question order effect?

Background

Concept inventory development

In their review of concept inventory development methodologies, Lindell *et al.* (2007, p. 14) defined 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," usually consisting of 20 plus items constructed around the concepts of interest. The process for developing instruments that have strong validity arguments and high reliability involves statistical analyses of item and instrument characteristics as well as qualitative methods such as student interviews to generate questions, and plausible distractors, and to test and refine wording and representations used in each item. The chemistry education research community is increasingly engaged in developing tools for measuring student conceptual understanding in chemistry. Recent journal articles highlight the need for researchers to continually collect, analyse, and report sources of validity evidence as they develop their own tools and as they use existing instruments in their own contexts (Arjoon *et al.*, 2013).

It is useful to note that an assessment or test is never itself "validated." Rather, validation studies are conducted to provide evidence for interpreting the meaning of test scores, which themselves are a function of the items, respondents, and contexts in which the assessment is given (Messick, 1995). Threats to validity also need to be considered. For instance, construct-irrelevant variance may be an important factor when items become more difficult or easier for some individuals or groups based on features that are irrelevant to the construct under measure (Messick, 1995). This is of interest in the current study as we investigate whether features of either a verbal-based question or pictorial question help or hinder respondents in subsequent related items, depending on the order they encounter them. If one question order produces an advantageous outcome over the other, that has implications for instrument design and score interpretations, especially in cases where item order is varied across administrations.

Question order effect

The idea that the order of questions on a test or concept inventory could affect the context in which students answer successive questions is called the question order effect (Oldendick, 2008). Indeed, there are numerous options for ordering questions on a test: they can be ordered sequentially (*i.e.*, following the order of the chapters learned during the course) or randomly; they can also been ordered based on question difficulty (*e.g.*, easiest to most difficult or *vice versa*). The question order effect was first studied by Mollenkopf in 1950 (Mollenkopf, 1950). In this study, Mollenkopf gave two versions of a verbal skills test and two versions of a mathematics test to high school junior and senior

students in the United States. Certain questions appeared early in one version and late in the other version. The different versions were also given under speed (*i.e.*, trying to complete as many questions as possible in a given time frame with the expectation that not all questions may be completed) and power (*i.e.*, trying to complete a small number of complex questions in the same time frame with the expectation of completing all questions) conditions. Mollenkopf found that of the four test conditions (verbal-power, verbal-speed, math-power, and math-speed), all but the math-power test yielded significant results between the two ordering versions, suggesting that question order effect plays a role in determining students' success on an exam. Since then, numerous studies have been conducted to test various factors around question order that could affect student exam performance (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). These studies have investigated similar factors to Mollenkopf's original study such as placement of items based on difficulty (Monk and Stallings, 1970; Coniam, 1993). Others have studied different types of test conditions such as the order of the test items relative to the order the material was taught in class and also whether the question order effect depended on subject matter (Hodson, 1984; Gohmann and Spector, 1989; Coniam, 1993).

Leary and Dorans (1985) created a summary of the results of major studies on question order effect completed to date. The results were inconclusive as to whether or not a question order effect existed when questions are ordered from difficult to easy. However, they did not find a question order effect based on content order (*i.e.*, sequenced *versus* randomly ordered tests). Overall, they found as many studies with significant test results as studies with insignificant results. They concluded that no firm conclusion could be reached as to the conditions under which the question order effect exists. Table 1 includes several additional studies on question order effect either completed since the summary by Leary and Dorans or not included in their summary. The results corroborate the conclusions made in Leary and Dorans that while some studies show significance for different factors, there is considerable disagreement in whether the question order effect exists.

Researchers have also been concerned about how question order could affect a test's reliability, validity, and difficulty, as well as students' motivation and post-exam evaluation (*e.g.*, Monk and Stallings, 1970; Hodson, 1984; Balch, 1989; Carlson and Ostrosky, 1992; Coniam, 1993; Pettijohn and Sacco, 2007). Pettijohn and Sacco (2007) and Balch (1989) found no significant differences in the time it took students to complete different versions of their exams. Monk and Stallings (1970) and Carlson and Ostrosky (1992) found that question order did not seem to affect question validity or exam reliability.

Studies conducted to date on the question order effect thus provide inconclusive results. However, these studies had limitations regarding their testing methods and participants which render their comparisons difficult. For instance, although random assignment of participants to different version of the test is preferred, certain studies could only ''randomize'' the test population by handing out different testing packets as students walked through the test centre door, or by pre-assigned test packets ''randomly'' following alphabetical order of the class roster (Plake, 1980; Balch, 1989). Participants' demographics also differed: some studies have been conducted with high school students (Mollenkopf, 1950), while others with undergraduate students (Gohmann and Spector, 1989), and still others with adults in the workforce (Bradburn and Mason, 1964). Perhaps two of the most critical differences limiting the comparability of these studies are in the physical testing conditions and the subject being tested. Until tests began to be administered digitally, these studies conducted their tests on paper, and as such the researchers could not explicitly control whether students were answering questions in the order expected, and thus were not strictly testing the question order effect (Dean, 1973). In addition, these studies cover a wide range of topics including psychology (Balch, 1989; Neely *et al.*, 1994), math (Mollenkopf, 1950; Leary and Dorans, 1985), verbal skills (Mollenkopf, 1950; Coniam, 1993), geography (Monk and Stallings, 1970), psychiatric nursing

Table 1 Examples of heavily cited studies on question order effect. s = significant; ns = not significant

Item order	Investigator	Test conditions[a]	Significance of results
Sequenced/content-ordered *vs.* Random content	Balch (1989)	Power	s
	Gohmann and Spector (1992)	Power	ns
	Neely *et al.* (1994)	Power	ns
	Pettijohn and Sacco (2007)	Power	ns
	Tal *et al.* (2008)	Power	ns
Variations in placement of a section of a questionnaire	Bradburn and Mason (1964)	Power	ns
Sequenced/content-ordered and Easy to Hard *vs.* Random content and Random difficulty	Carlson and Ostrosky (1992)	Power	s
Easy to Hard *vs.* Hard to Easy	Coniam (1993)	Power	ns
	Hodson (1984)	Speed	ns

[a] A power test assesses a student's ability with no regard for how long it takes to complete the test; an ideal power test would give students all the time necessary for them finish the entire test. A speeded test ideally would contain homogenously simple tasks so that if students were given an unlimited amount of time, they should be able to get 100%, but because of the time limit, it is testing the students' processing speed. Most tests fall somewhere in between these two; most tests commonly fall under the category ''timed power test'' such as the SAT, ACT, and GRE because not all speed tests can have items that are ''trivially easy'' and most power tests are restricted to some kind of time constraint (Mead and Drasgow, 1993).

(Plake, 1980), physics (Gray *et al.*, 2002), general social science (Crano, 1977), job related interviews (Bradburn and Mason, 1964), business and economics (Dean, 1973; Gohmann and Spector, 1989; Carlson and Ostrosky, 1992) and relatively few in chemistry (Hodson, 1984). As Bradburn and Mason stated: "it is impossible to generalize with any degree of confidence to other situations. . .the effect of a particular question or topic on later questions can only be determined empirically within the context of a particular questionnaire" (Bradburn and Mason, 1964, p. 61).

In summary, the literature across several disciplinary domains is inconclusive on whether or not a question order effect actually exists; many types of ordering and other factors have been tested, yet the empirical evidence shows mixed results. It is thus not yet clear whether researchers are introducing unwanted variability to their data if they utilize alternate versions of a concept inventory.

Format of assessment question: pictorial *versus* verbal

The design of assessment tools has been extensively researched. Specific attention has been provided to the qualities of the verbal and pictorial representations in questions. For example, Haláková and Prokša (2007) found that three of the most important qualities to consider in order to minimize confusion are the length of verbal questions, which could cause fatigue, word choice in verbal questions, as well as picture captions and clarity of the picture for the sake of understanding what it is trying to communicate to the student. Several characteristics have been identified as critical to the effectiveness of a picture: color, realism, relevance, interactivity, and animation (Phillips *et al.*, 2010). Phillips *et al.* (2010) suggested the following guidelines to maximize the use of visuals:

- The visuals must be relevant and not distracting to the text.
- The content of the visuals is more important than color, simplicity, or realism.
- The point of using visuals is to supplement, not replace text.

It is especially important to keep these guidelines in mind in the design of assessment tools because "the main risk of including images in the context of examining is that an image may lead to the formations of a mental representation of a question that does not match the meaning intended by the question setter" (Phillips *et al.*, 2010, p. 141). Indeed, studies have found that students do not use the same visuals the same way (Angeli and Valanides, 2004; Crisp and Sweiry, 2006). For example, Duran and Balta (2014) found that having visuals within test questions had a significant effect on student scores for students who did not already excel at science, but no effect for students who already did well in science. In chemistry, studies have also found that students struggle in their analysis of visual questions (Nurrenbern and Pickering, 1987; Sanger and Phelps, 2007), in part because of the limitations associated with static representations (*e.g.*, velocity of atoms and molecules are typically not represented even though this information may be critical to selecting the correct answer) (Sanger and Phelps, 2007). Duran and Balta (2014) thus suggested that it is not necessarily always better to have a visual for every test question but that some questions would be more effective with just text.

The balance between visuals and text has been an important topic in the assessment literature. Studies by Holliday (1975), Kapıcı and Savaşcı–Açıkalın (2015), Mayer (1989), Mayer and Anderson (1991), Mayer *et al.* (1996), and Phillips *et al.* (2010) have all concluded that pictures would actually be almost useless without an appropriate, small amount of text to accompany them in the form of captions or supporting text. Of course, the text must also be relevant to the visual to have any impact (Mayer and Anderson, 1991). In other words, with either too much or too little accompanying text or instructions, students tend to undervalue and ignore the visuals, in which case a "good picture" can actually fail to serve its purpose (Weidenmann, 1989, p. 163).

One of the most widely used theories to explain the cognitive processing of visuals and text is dual coding theory (Paivio, 1990; Clark and Paivio, 1991; Paivio, 2013). This theory states that we are cognitively capable of encoding both visual and verbal forms of information. Verbal information is only coded verbally, while imagery can be coded using both verbal and visual encoding. The dual coding theory suggests that using both representational and referential processing can aid in recall and recognition of learned information (Mayer and Anderson, 1991). Many researchers use dual coding theory to explain the benefits of using visuals in education (Winn, 1987; Weidenmann, 1989; Clark and Paivio, 1991; Mayer and Anderson, 1991). However, some are more skeptical. For example, Schnotz and Bannert (2003) claimed that dual coding theory is not sufficient because it does not take into account that students can encode visuals incorrectly, and therefore the visuals can have a negative effect on the learning. They concluded that it was not appropriate to assume that pictures have a generally beneficial effect on learning. In fact, Schnotz (2002) suggested that when presenting pictorial and verbal test questions, it may be better to present the visuals first because they require less working memory space, and then the verbal portion should follow.

In summary, the literature is in general agreement that the judicious use of visuals in test questions is generally beneficial, particularly in science education. The goal of the current study is to test for question order effect within the realm of verbal *versus* pictorial questions.

Methodology

The study was conducted in two phases. The first phase was conducted at a large research university in the Rocky Mountain region of the United States (labelled thereafter as Western University), in the early stages of development of a concept inventory on solubility and acid–base concepts. The second phase was conducted at a large research university in the Central region of the United States (labelled thereafter as Midwestern University) in order to determine whether question order effects persist across institutional contexts and in varied

testing conditions. Participants and data collection for each phase of the study are described below. Data analysis methods and results follow. Human subjects research protocols were submitted, reviewed, and approved prior to data collection at each institution. Students provided their informed consent for each aspect of the study in which they participated.

Phase 1: Western University

Participants were students enrolled in the second semester of a General Chemistry course (GCII). A paper version of the 20-item acid/base and solubility concept inventory was administered in recitation sections as a pre-test in week 1 of Spring semester, before the relevant material had been covered in class. Each recitation section consisted of 15 to 20 students.

Following the pre-test, researchers conducted think-aloud interviews with seven students on the entire 20-item instrument. In this version, students encountered pictorial questions regarding acid concentration/strength/pH relationships early and responded to the verbal question on the same concept later. Upon noticing that some students approached the pictorial question differently than the related verbal question during the interview, the researchers created two versions of the instrument to administer at the end of the semester. One version (labelled PV for purposes of this study) was identical to the pre-test, with pictorial question (P) as Question 9, and the verbal question (labelled V) as Question 18. The second version (labelled VP for purposes of this study) switched the placement of these questions. The first 8 questions remained the same across both versions and were followed by either P or V, depending on the inventory version. The remaining questions were kept in the same order between the two versions.

Post-test versions were administered in the last week of the semester to students based on their regular recitation section; the PV version was given to students who attended their weekly recitation sections on Monday (five sections) and Tuesday (nine sections), while the VP version was given to students in recitation sections on Wednesday (five sections), Thursday (eleven sections), and Friday (two sections). Students were allotted 3 points out of 1000 total course points for completing the inventory, regardless of their score. Of 640 students in the class, a total 553 post-tests were completed.

All Graduate Teaching Assistants were instructed on how to administer the inventory to their recitation sections. During the weekly lab/recitation preparation meeting, the researcher emphasized the importance of collecting good data, meaning that all students needed to be encouraged to give their best efforts with also knowing their score would not affect their grades in any way. Students were allowed to take as long as they needed to complete the instrument, and most students finished within 25 minutes of the 50-minute recitation period.

Phase 2: Midwestern University

The second portion of this study was conducted at the Midwestern university during Spring and Fall semesters using a mixed-methods explanatory design. Students were enrolled in the second semester of General Chemistry (GCII) or the first semester of Organic Chemistry (OCI). The concept inventory is relevant for use in Organic Chemistry, since acid base chemistry is central to the subject. Both PV and VP versions of the instrument were distributed *via* Qualtrics, an online surveying software. Students were offered extra credit points by their professor for taking the concept inventory. The professor distributed the Qualtrics link, and students subsequently had one week following distribution of the link to take the concept inventory on their own time (similar to power conditions). The inventory was distributed at the end of the semester. The Qualtrics software randomly presented each student with either version PV or VP. The software also forced students to complete the inventory in sequential order, restricting them from returning to questions after they were completed. A total of 768 online inventories were collected, of which 643 were usable after removing responses from students who self-reported using outside resources while completing the inventory, which they had been instructed not to do.

Qualitative data were collected through 19 student interviews conducted after the collection of the inventories. Students received the same version of the concept inventory that they took online. Participants were selected based on how they answered P and V. If students took version PV, they were contacted if they answered P correctly and V correctly or incorrectly. If students took version VP, they were contacted if they answered V correctly and P correctly or incorrectly. Each interview consisted of two parts. First, students were asked to think-aloud as they solve questions P and V. This type of interview was chosen as it is one of the most effective strategies to capture students' thinking processes while they perform a task (Ericsson and Simon, 1980). Second, students were engaged in a semi-structured interview, in which they were probed about their preference of whether P or V was presented first, whether or not the initial question was helpful in answering the following question, and whether a change in the order of the two questions would have helped them answer the second question. Each interview transcript was read and annotated by the first author. The first and last author independently classified each transcript based on interviewee's preferences for seeing question V or P first and coded reasons for their choices. Upon comparisons of codes, few inconsistencies were found and resolved through discussion (Saldaña, 2015).

Data selection and analysis

Several measures were taken to ensure that non-valid data were excluded from the analyses. Since a paper and pencil version was used at the Western university, the researcher initially removed scantrons in which students obviously completed fewer than 60% of the items or whose filled-in bubbles made a pattern (for instance, repeating A-B-C-D or all C's). Both universities used a demographic question at the end of the inventory asking for self-reported effort on a scale of 1 to 4, with 1 representing "I gave it my best effort", and 4 representing "I didn't take it too seriously." Students who reported a low effort of 4 were removed from the data set in order to eliminate responses which did not reflect a reasonable consideration of

Table 2 Western university results; percentages of students answering indicated question correctly

GCII post-instruction, moderate/high effort						GCII post-instruction, high effort					
	Test version		Statistics				Test version		Statistics		
Question	PV (n = 205) (%)	VP (n = 291) (%)	p value	Φ	$1-\beta$	Question	PV (n = 165) (%)	VP (n = 235) (%)	p value	Φ	$1-\beta$
P	46.3	57.0	0.024	0.106	0.446	P	47.9	60.4	0.017	0.124	0.498
V	38.5	30.6	0.081	0.083	0.260	V	40.0	34.5	0.306	0.057	0.090

Table 3 Midwestern university results

GCII post-instruction, moderate/high effort						GCII post-instruction, high effort					
	Test version		Statistics				Test version		Statistics		
Question	PV (n = 101) (%)	VP (n = 101) (%)	p value	Φ	$1-\beta$	Question	PV (n = 70) (%)	VP (n = 61) (%)	p value	Φ	$1-\beta$
P	39.6	49.5	0.157	0.100	0.144	P	47.1	57.4	0.242	0.102	0.094
V	43.6	53.5	0.159	0.099	0.141	V	38.6	57.4	0.032	0.188	0.370
OCI post-instruction, moderate/high effort						OCI post-instruction, high effort					
	Test version		Statistics				Test		Statistics		
Question	PV (n = 82) (%)	VP (n = 81) (%)	p value	Φ	$1-\beta$	Question	PV (n = 39) (%)	VP (n = 40) (%)	p value	Φ	$1-\beta$
P	29.3	30.9	0.824	0.017	0.015	P	35.9	40.0	0.707	0.042	0.019
V	35.4	38.3	0.701	0.030	0.020	V	38.5	42.5	0.715	0.041	0.019

the inventory questions. This left a population of students who self-reported an effort level of 1, 2, or 3, referred to as ''moderate/high effort.'' We also extracted a population subset from this data by removing students who responded with an effort of 3, leaving only student who reported a ''high effort'' of 1 or 2. Tables 2 and 3 report the number of participants in each population for each institution.

Cleaning of the Midwestern data was done by removing the scores of participants who self-reported using resources even though they were instructed not to at the beginning of the online concept inventory. No strict data cleaning could be done based on the amount of time spent taking the concept inventory online because only a start and stop time were included in the derived Qualtrics data. If the two time stamps indicated an unreasonably short time spent between opening the inventory and submission, this data was excluded, although this applied to very few participants. Because of the nature of the online medium, we could not control how long students left the inventory open on their computer. Technically students could have left the inventory open on their computer for the whole week while working on it a little bit at a time throughout the week, making the Midwestern data collection resemble an almost ideal power test, whereas the Western university's data collection resembles more of a timed power test since the questions are not trivially easy, but the inventory time was limited to the written lab quiz time (Mead and Drasgow, 1993). We were not concerned about the difference in mediums due to the fact that Mead and Drasgow (1993) found in their study that for power tests (not speeded) online *versus* paper-and-pencil medium did not affect participants' performance.

Concept inventory data were analysed using the statistics software SPSS. This software was used to compute a 2×2 contingency test based on whether the students' answer choices were correct or incorrect. We applied the Bonferroni correction to minimize Type I error; this led to a threshold level of significance of 0.013 for each population. In addition, SPSS was used to calculate t-tests for the total inventory scores and for the students' total scores on the first 8 questions of the inventory.

Since the first 8 questions of the inventory were always presented in an identical order on all versions of the inventory, they functioned as a set of control questions that enabled the student performance on the first portion of the instrument to be compared universally. Student performance on these initial control questions was compared across all inventory sections, and no statistically significant differences were observed (see Appendix).

Statistical analysis of student demographics

In order to ensure that the populations between the two concept inventory versions were in all other ways comparable, demographic information was collected at each university. The Western university collected information on class level, grade in the current class, and major. The Midwestern university collected information on class level, major, whether or not the student is repeating the class they are currently in, their grade in their previous chemistry class, gender, and the lecture section (instructor) they were currently enrolled in. Participant populations were compared using Fisher or Chi Square tests. All populations were determined to be demographically comparable.

Results

Concept inventory results

Table 2 shows the quantitative results on the two questions targeted in this study (P and V) for the concept inventories taken at the Western university, where the concept inventories

were completed on paper. Table 3 shows the same results for the inventories taken at the Midwestern university, where the concept inventories were taken online.

We found no statistical significance for any of the comparisons at both institutions indicating an absence of the question order effect in our populations. After no differences were found, we calculated the statistical power for each test using the observed effect sizes (shown in Table 3). These results indicate relatively low probabilities (between 0.015 and 0.498) that the sample sizes used in this study would be large enough to identify the observed effect sizes even if such effects existed (Type II error rate was between 0.502 and 0.985). Therefore, we identified two trends that were present in our data set that warrant further investigation with larger sample sizes. First, we found at both institutions that GCII students who took version VP, in which they saw the pictorial question P after the verbal question V, performed approximately 10% better on P than the students who took version PV, in which they saw P prior to V. At the Midwestern university, we found that the High effort GCII student population performed almost 20% better on the verbal question when they saw it first (inventory version VP) than when they saw it second (inventory version PV). This trend was also observed in the moderate/high effort group. This suggests that students who take the inventory more seriously may be more affected by the order of questions. All these trends should be investigated further since currently they can only be interpreted as noise.

Interview results

Table 4 shows the results for the interviews from the Midwestern university. When asked about their use of the first question to help them answer the second isomorphic question, students interviewed were split: seven of the 18 students indicated making use of the first question while 11 of the 18 students indicated not using it. Of the seven students who had a definitive preference for which question was presented first, four preferred to see the verbal question (V) first. These students described V as acting like a definition for strong and weak acids:

"*Yeah because [V] was kind of like a definition almost and that kind of thing and [the diagram] was kind of more applied so it built off of it. . .[having V first] made me more sure of my answers.*"

Three out of the seven students preferred seeing the pictorial question (P) first because it helped them "*visualize what was going on with the strong and weak acids.*" Interestingly, the benefits advanced by these students for these preferences did not always materialize. Indeed, three out of the seven students who claimed that seeing one question before the other helped them answer the second question provided incorrect answer on this latter question.

Several reasons were advanced by the 11 students who did not use the first question to answer the second question. Five indicated dealing with questions independently of each other when taking any kind of tests. Each question is thus treated by these students in isolation. Four believed that there may have been some kind of subconscious effect having seen similar questions previously but did not consciously use the first question. Three indicated that they felt they knew the concept well enough that they could answer the questions regardless of their order. Another three students felt that P and V worked together as a package and that it was helpful to see them next to each other regardless of which question came first.

Table 4 Interview results

		Interviewee				
		GCII (n = 7)		OCI (n = 12)		
Interview question	Choice provided by interviewee	PV	VP	PV	VP	Example of quote providing justification for choice
Did you prefer seeing the [verbal/pictorial] question first? Did it matter to you?	Pictorial	2	0	1	0	"I thought it [P1] kind of helped to visualize the dissociatedness because you can tell the stronger ones and they tell you here that's undissociated and you know it's a weak acid. So yeah that kind of helped me to see [P1] first."
	Verbal	0	1	0	3	"Yeah because [V] was kind of like a definition almost and that kind of thing and [the diagram] was kind of more applied so it built off of it. . . [having V first] made me more sure of my answers."
	No preference	2	0	4	4	". . .they kind of work in a package where like no matter which order you put them in they all kind of influence the other one the ones that follow. . ."
	N/A	1	1	0	0	
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	2	1	1	3	"I did think that [V] influenced my answer because I, if I wasn't 100% certain on the behavior of the strong and weak acid, I leaned back on my answer for [V] to answer [P], so by choosing an answer here in [V], I carried that information forward to [P]."
	No	2	1	4	4	"I just kind of went through them. I didn't really think about the other questions. I guess that's just kind of how I take tests."
	Subconscious	1	1	1	1	"I guess subconsciously it did [influence my answer], but like I wasn't aware of it."
	N/A	1	0	0	0	

Discussion and conclusions

In this study, we examined question order effects within a concept inventory on solubility and acid/base topics among chemistry students enrolled in a General Chemistry and/or Organic Chemistry courses at two different universities in the United States. The quantitative analyses at both institutions revealed no statistically significant question order effects among all populations investigated. Some trends pointing to an order effect for specific population were observed and should be investigated further with a larger population size in order to achieve a higher statistical power than the one accomplished in this study. These results corroborate with the current literature, in which question order effects are often observed, but not with any great consistency or predictability (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).

From a practical perspective, the lack of ordering effects observed in this study indicate to instructors that generation of alternate test versions for use in the classroom should not favour a particular group of students. However, the implications are more cautionary for educational researchers using concept inventories as a research tool. This study indicates that specific questions on an inventory may exhibit an ordering effect and that this effect may be more prevalent across different types of students. Researchers planning to run question-level psychometric analyses on concept inventory responses should use particular care in generating differently-ordered inventory versions. At times, question order effects may disrupt analyses of validity, reliability, or difficulty of particular questions.

Some studies suggest that other factors besides question order can have as much or more impact on student performance (*e.g.*, answer choice order, item difficulty) inasmuch that having several complex questions in a row could cause cognitive overload as the student proceeds to successive questions (Tellinghuisen and Sulikowski, 2008; Schroeder *et al.*, 2012). There is thus a critical need to further study which factors matter most when developing assessment tools intended to measure students' learning as accurately as possible.

Limitations of the study

Although this two-part study was intended to address methodological weaknesses of prior investigations of the question order effect, limitations are still present.

First, this study investigated the question order among second semester general chemistry and first semester organic chemistry students in the United States. It is possible that students enrolled in lower-level chemistry undergraduate courses did not show any question order effect with pictorial and verbal questions, but upper level students, students from other science disciplines, or students from other countries might experience this effect. We are currently conducting a similar study with students enrolled in lower level geoscience courses to test whether a disciplinary effect exists.

Second, students from the two different institutions took the concept inventories under different conditions and incentives. For example, the students at the Western university were required to take the inventory in person and on paper during recitation and were offered a certain number of points towards their grade for taking the inventory. On the other hand, the Midwestern students took the inventory online under voluntary circumstances and were offered extra points toward their laboratory or lecture grade at the instructor's discretion. These differences in testing conditions could have affected how the students performed overall on the inventory. We did attempt to control for these differences in implementation by separating out students who indicated providing a high effort in answering the inventory from those indicating providing moderate effort.

Third, the sample size of the Midwestern university study was not as large as the one from the Western university, which diminished the identification of statistically significant results and resulted in low statistical power. The lower response rate at the Midwestern university may be due to two factors: first, the voluntary nature of the study at the Midwestern school could have dissuaded students from investing extra time and effort if they did not feel that they needed the extra credit; data had to be cleaned extensively based on self-reported effort levels and usage of external resources. Second, since names were attached to the inventory collected at the Western university and the collection of inventory was conducted in class, students may have felt more compelled to take the task seriously.

Appendix

Tables 5–8.

Table 5 Results of the comparison of students' performance on the first eight questions between the PV and VP versions for the Midwestern institution

Student population		Average score on the first eight questions		Significance		
Course	Student effort level	PV version	VP version	*p*-value	Eta-squared	$1-\beta$
GCII post	Moderate/ high effort	3.91	4.19	0.298	0.0054	0.013
	High effort	4.17	4.34	0.588	0.0023	0.013
OCI post	Moderate/ high effort	3.66	3.74	0.787	0.0005	0.013
	High effort	4.05	4.35	0.512	0.0056	0.013

Table 6 Results of the comparison of students' performance on the first eight questions between the PV and VP versions for the Western institution

Student population		Average score on the first eight questions		Significance		
Course	Student effort level	PV version	VP version	*p*-value	Eta-squared	$1-\beta$
GCII post	Moderate/ high effort	5.17	5.24	0.610	0.0231	0.026
	High effort	5.19	5.43	0.156	0.0703	0.140

Table 7 Results of the comparison of students' performance on the concept inventory as a whole between the PV and VP versions for the Midwestern institution

Student population		Average total score on the concept inventory (/18)[a]		Significance		
Course	Student effort level	PV version	VP version	*p*-value	Eta-squared	$1-\beta$
GCII post	Moderate/ high effort	7.50	8.05	0.196	0.0083	0.013
	High effort	7.93	8.43	0.332	0.0073	0.013
OCI post	Moderate/ high effort	6.90	7.12	0.652	0.0013	0.013
	High effort	7.74	8.38	0.405	0.0090	0.013

[a] Two questions are not included in the calculations of the total score because they were not distributed to all courses because of a programming error in Qualtrics.

Table 8 Results of the comparison of students' performance on the concept inventory as a whole between the PV and VP versions for the Western institution

Student population		Average total score on the concept inventory (/20)		Significance		
Course	Student effort level	PV version	VP version	*p*-value	Eta-squared	$1-\beta$
GCII post	Moderate/ high effort	10.49	9.90	0.0574	0.0860	0.285
	High effort	10.62	10.35	0.4318	0.0399	0.046

Acknowledgements

We would like to acknowledge the University of Nebraska-Lincoln's Undergraduate Creative Activity and Research Experience (UCARE) Grant for providing funding to support the first author and the University of Colorado Boulder Science Education Initiative (SEI) for financially supporting part of the project. We would also like to thank all the students and instructors whose cooperation made this study possible. We would like to thank Dr Jordan Harshman for helpful guidance on the statistical analyses.

Notes and 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, *Educ. Technol. Res. Dev.*, **52**, 23–36.

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: examining the psychometric evidence, *J. Chem. Educ.*, **90**, 536–545.

Balch W. R., (1989), Item order affects performance on multiple-choice exams, *Teach. Psychol.*, **16**, 75–77.

Bradburn N. M. and Mason W. M., (1964), The effect of question order on responses, *J. Marketing Res.*, 57–61.

Carlson J. L. and Ostrosky A. L., (1992), Item sequence and student performance on multiple-choice exams: further evidence, *J. Econ. Educ.*, **23**, 232–235.

Clark J. M. and Paivio A., (1991), Dual coding theory and education, *Educ. Psychol. Rev.*, **3**, 149–210.

Coniam D., (1993), Does the ordering of questions in a test affect student performance, *Educ. Res. J.*, **8**, 74–78.

Crano W. D., (1977), Primacy versus recency in retention of information and opinion change, *J. Soc. Psychol.*, **101**, 87–96.

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

Dean M. L., (1973), The impact of exam question order effects on student evaluations, *J. Psychol.*, **85**, 245–248.

Duran M. and Balta N., (2014), The influence of figured and non-figured questions on secondary students' success at science exams, *Pak. J. Stat.*, **30**, 1279–1288.

Ericsson K. A. and Simon H. A., (1980), Verbal reports as data, *Psychol. Rev.*, **87**, 215.

Gohmann S. F. and Spector L. C., (1989), Test scrambling and student performance, *J. Econ. Educ.*, **20**, 235–238.

Gray K., Rebello S. and Zollman D., *The effect of question order on responses to multiple-choice questions*, Boise, Idaho, 2002.

Hake R. R., (1998), Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.*, **66**, 64–74.

Haláková Z. and Prokša M., (2007), Two kinds of conceptual problems in chemistry teaching, *J. Chem. Educ.*, **84**, 172–174.

Hestenes D., Wells M. and Swackhamer G., (1992), Force concept inventory, *Phys. Teach.*, **30**, 141–158.

Hodson D., (1984), The effect of changes in item sequence on student performance in a multiple-choice chemistry test, *J. Res. Sci. Teach.*, **21**, 489–495.

Holliday W. G., (1975), The effects of verbal and adjunct pictorial-verbal information in science instruction, *J. Res. Sci. Teach.*, **12**, 77–83.

Kapıcı H. Ö. and Savaşcı-Açıkalın F., (2015), Examination of visuals about the particulate nature of matter in Turkish middle school science textbooks, *Chem. Educ. Res. Pract.*, **16**, 518–536.

Leary L. F. and Dorans N. J., (1985), Implications for altering the context in which test items appear: a historical perspective on an immediate concern, *Rev. Educ. Res.*, **55**, 387–413.

Lindell R. S., Peak E. and Foster T. M., (2007), *Are they all created equal? A comparison of different concept inventory development methodologies*.

Mayer R. E., (1989), Systematic thinking fostered by illustrations in scientific text, *J. Educ.Psychol.*, **81**, 240.

Mayer R. E. and Anderson R. B., (1991), Animations need narrations: an experimental test of a dual-coding hypothesis, *J. Educ. Psychol.*, **83**, 484.

Mayer R. E., 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, *J. Educ. Psychol.*, **88**, 64.

Mead A. D. and Drasgow F., (1993), Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis, *Psychol. Bull.*, **114**, 449.

Messick S., (1995), Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *Am. Psychol.*, **50**, 741.

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. and Stallings W. M., (1970), Effects of Item Order on Test Scores', *J. Educ. Res.*, **63**, 463–465.

Mulford D. R. and Robinson W. R., (2002), An inventory for alternate conceptions among first-semester general chemistry students, *J. Chem. Educ.*, **79**, 739–744.

Neely D. L., Springston F. J. and McCann S. J., (1994), Does item order affect performance on multiple-choice exams?, *Teach. Psychol.*, **21**, 44–45.

Nurrenbern S. C. and Pickering M., (1987), Concept learning versus problem solving: Is there a difference?, *J. Chem. Educ.*, **64**, 508.

Oldendick R. W., (2008), Question Order Effect, http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n428.xml, accessed 12/14, 2015.

Paivio A., (1990), *Mental representations: a dual coding approach*, New York, NY: Oxford University Press.

Paivio A., (2013), *Imagery and verbal processes*, Psychology Press.

Pettijohn T. F. and Sacco M. F., (2007), Multiple-choice exam question order influences on student performance, completion time, and perceptions, *J. Instr. Psychol.*, **34**, 142.

Phillips L. M., Norris S. P. and Macnab J. S., (2010), *Visualization in mathematics, reading and science education*, Springer Science & Business Media.

Plake B. S., (1980), Item arrangement and knowledge of arrangement on test scores, *J. Exp. Educ.*, **49**, 56–58.

Saldaña J., (2015), *The coding manual for qualitative researchers*, Thousand Oaks, CA: Sage.

Sanger M. J. and Phelps A. J., (2007), What are students thinking when they pick their answer? A content analysis of students' explanations of gas properties, *J. Chem. Educ.*, **84**, 870.

Schnotz W., (2002), Commentary: towards an integrated view of learning from text and visual displays, *Educ. Psychol. Rev.*, **14**, 101–120.

Schnotz W. and Bannert M., (2003), Construction and interference in learning from multiple representation, *Learn. Instr.*, **13**, 141–156.

Schroeder J., Murphy K. L. and Holme T. A., (2012), Investigating factors that influence item performance on ACS exams, *J. Chem. Educ.*, **89**, 346–350.

Tal I. R., Akers K. G. and Hodge G. K., (2008), Effect of paper color and question order on exam performance, *Teach. Psychol.*, **35**, 26–28.

Tellinghuisen J. and Sulikowski M. M., (2008), Does the answer order matter on multiple-choice exams?, *J. Chem. Educ.*, **85**, 572.

Weidenmann B., (1989), in *Knowledge acquisition from text and pictures: Advances in Psychology*, ed. Levin H. M. J. R., Amsterdam: Elsvier.

Weinstein Y. and Roediger III H. L., (2012), The effect of question order on evaluations of test performance: how does the bias evolve?, *Mem. Cognition*, **40**, 727–735.

Winn B., (1987), in Houghton D. M. W. H. A. (ed.), *The psychology of illustration*, New York: Springer-Verlag, vol. 1, pp. 152–198.

