

**Kennesaw State University**  
**DigitalCommons@Kennesaw State University**

---

Faculty Publications

---

2011

# Establishing Open-Ended Assessments: Investigating the Validity of Creative Exercises

Scott E. Lewis

*Kennesaw State University, slewis57@kennesaw.edu*


Janet L. Shaw

*Kennesaw State University, jshaw22@kennesaw.edu*

Kathryn A. Freeman

*Kennesaw State University*

Follow this and additional works at: <https://digitalcommons.kennesaw.edu/facpubs>

 Part of the [Chemistry Commons](#), [Curriculum and Instruction Commons](#), and the [Science and Mathematics Education Commons](#)

---

## Recommended Citation

Lewis SE, Shaw JL, Freeman KA. 2011. Establishing open-ended assessments: Investigating the validity of creative exercises. *Chemistry Education Research and Practice* 12(2):158-66.

This Article is brought to you for free and open access by DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Faculty Publications by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact [digitalcommons@kennesaw.edu](mailto:digitalcommons@kennesaw.edu).

# Establishing open-ended assessments: investigating the validity of creative exercises

Scott E. Lewis, Janet L. Shaw and Kathryn A. Freeman

Received 21st September 2010, Accepted 25th February 2011

DOI: 10.1039/C1RP90020J

Open-ended assessments, defined as assessments with a large set of possible correct answers, by nature lend themselves to concerns regarding accurate and consistent grading. This article describes one particular open-ended assessment, named Creative Exercises (CE), designed for promoting students' interconnection of concepts in a college general chemistry setting. The article presents evidence concerning several aspects of validity, including the extent scores represent chemistry knowledge and the extent scoring is consistent across three graders. The evidence is also presented in the context of what is known about concept maps, a commonly employed open-ended assessment in chemistry. Implications for the administration of CEs and the appropriateness of measuring students' hierarchical organization of knowledge are also discussed as a result of this comparison.

**Keywords:** assessment, validity, concept maps, general chemistry, misconceptions

## Introduction

Proponents of curricular reform have developed numerous methods to incorporate pedagogical changes into the chemistry classroom though assessment techniques have remained relatively unchanged (Bowen and Phelps, 1997; Wright *et al.*, 1998). Modifying assessment practices is particularly important as it has been suggested that our assessment practices have a strong influence over how students direct their efforts in a class (Trigwell and Sleet, 1990; Scouller, 1998; Biggs, 2001). It seems likely that the development of new assessment techniques have the potential to effect student motivations and ultimately improve student learning within a course. Such a new assessment would first have to be demonstrated to be a valid and reliable measure of students' chemistry knowledge. This article describes an investigation into several aspects of validity for an assessment technique that is designed to measure students' ability to form connections across the material within a course.

## Creative exercises

The assessment investigated in this study is termed Creative Exercises (CE) and was originally proposed by Trigwell and Sleet (1990). In a CE, students are given a brief prompt, for example '7.5 g of NaBr', and are asked to write down as many distinct, correct and relevant statements that pertain to the original prompt. Students receive credit for each statement that is correct, relevant to the material presented in the course and the original prompt, and distinct from the other statements for which they have received credit.

Students are not penalized for any incorrect statements, in order to spur creativity in their responses. The instructor (or panel of instructors) sets a maximum number of statements allowed for which students can receive points, and creates a rubric of likely student CE responses prior to grading. During grading, if there are responses that are not indicated on the rubric, they are decided on a case-by-case basis, and if credit is given the response is added to the rubric so that any similar response can also receive credit. As a result, CEs are considered open-ended assessments, as there is a large range of possible correct answers for students. A sample CE assignment, including a scoring rubric is available in Appendix A.

The goal of this study is to investigate the validity of CEs as a measure of chemistry knowledge in a first semester General Chemistry setting. Several aspects of validity are examined, including the relationship to a traditional assessment method and the consistency in scoring CEs. The investigation is designed to contrast the validity of two administrative methods for CEs: as homework assignments where students may consult additional resources, versus in-class assignments where students are timed and not permitted access to resources. This undertaking has the potential to support an instructor's use of CEs as an assessment technique, and may also offer a validated assessment technique to evaluate constructivist based learning reforms (Holme *et al.*, 2010).

## Assessments in chemistry

Common traditional assessment techniques in chemistry include multiple choice questions and short answer questions that have a clearly defined answer. There are benefits to these forms of assessment. Because they are written, they may be administered to a large class in one sitting, they can be graded consistently across students and

Kennesaw State University, Department of Chemistry and Biochemistry, 1000 Chastain Rd.; MB 1203, Kennesaw, GA 30144; USA; e-mail: [slewis57@kennesaw.edu](mailto:slewis57@kennesaw.edu)

in a timely manner, and they can measure individual student performance, which is often necessary for grading decisions. One critique of multiple choice questions and short answer questions is that they are largely instructor-centered in that instructors determine which information is to be included on the test. This can be mitigated by soliciting information from students on the topics to include, but this relies on the quality of student feedback. In addition, multiple choice questions and short answer questions tend to measure information in discrete pieces, where each piece of information is modeled as independent of other information in the class.

In contrast, under the constructivist paradigm the learner actively incorporates new information within existing schemas or mental structures. In the cognitive constructivist framework, students “*construct knowledge by transforming, organizing, and reorganizing previous knowledge*” (Santrock, 2008). Constructivism has been successfully incorporated into science teaching and learning, and into chemistry knowledge acquisition in particular (Bernal, 2006; Cakir, 2008). With CEs, students are assessed based on their ability to connect and incorporate new course information with information that was presented previously in the course. The act of students incorporating new course information with their own existing information maps onto the process of organization in Abraham’s description of the constructivist theory of learning (2008). As a result, the use of CEs in a classroom matches constructivist learning; as students approach new information they can make linkages to existing schemas and can subsequently use these connections to support their answers to the CEs.

CEs also allow students to choose which information to include in responses. For example, for the prompt ‘7.5 g of NaBr’, students may choose to focus on mathematical statements including moles, molar mass, or mass percent. Conversely, students may choose to name the compound, categorize the compound as ionic, or indicate properties of ionic compounds. In this way, CEs are more open-ended than traditional assessments where there is only one or a small set of correct answers. CEs are also more student-centered than traditional assessment, as CEs have the ability to value all the relevant information students can present, rather than assessing an instructor-defined objective as measured in a multiple choice question. To facilitate their implementation, CEs are similar to conventional assessments in that CEs are written, and can be administered to students individually, and scored in a timely manner.

As an alternative to conventional assessment techniques, CEs are most similar to concept maps as a technique that measures students’ ability to form connections (Francisco *et al.*, 2002). In a concept map students are asked to map a sequence of propositions, where each proposition is two concepts connected by an arrow and a linking word. In many cases, one central concept can be involved in many propositions. As an assessment technique, many variants of concept maps are available in the literature (Stoddart *et al.*,

2000). For example, scoring procedures may emphasize the organization of concepts (Novak and Gowin, 1984), interconnectedness of concepts (McClure *et al.*, 1999), or the validity of the propositions used (Francisco *et al.*, 2002). Students’ organization of concepts has alternatively been referred to as their hierarchy of knowledge or structure of knowledge. For clarity purposes, the term organization will be used throughout this article to represent any type of overall structure of knowledge relationships, and hierarchy reserved for the description of a pyramid type relation of knowledge.

Scoring procedures that emphasize organizations are dependent on assumptions concerning how knowledge is organized, which can be hierarchical, associative, or cyclical (Safayeni *et al.*, 2005; Derbentseva *et al.*, 2007). Additionally, the choice of how knowledge is organized may be domain or content specific, so that a hierarchical structure and grading scheme may be appropriate for some topics, while an associative scheme would be appropriate for others (Ruiz-Primo and Shavelson, 1996). For example, a concept map on naming covalent versus ionic compounds may be viewed as a hierarchical organization, while the variables in a gas law may be viewed as a cyclical organization. This variety hinders the development of a uniform scoring process for concept maps, and may serve to hinder adoption of concept maps as a practical assessment technique. Additionally, assessing the organization a student uses may also be problematic, as a variety of mental organizations may lead to a successful understanding of a chemistry topic, just as research has considered a variety of organizations to explain student understanding (Barenholz and Tamir, 1992; Markow and Lonning, 1998; Jones *et al.*, 2000).

In contrast, CEs provide credit for students to form relationships among content, but do not require students to describe the network of relationships. As a result, CEs have a relatively simple grading scheme that involves determining the number of correct, related concepts. In this sense, CEs reward students for forming connections regardless of the nature of the connection itself. CEs are similar to concept maps in that both seek to evaluate the number of correct propositions and amount of interconnectedness. In a CE, any correct statement that links related concepts (similar to propositions) or unrelated concepts (similar to interconnectedness) is valued. Therefore, both CEs and concept maps measure students’ completion of the organization process of constructivism. CEs are distinct from concept maps by rewarding connections without requiring an assessment decision on the organization of knowledge (since many organizations may be plausible). Additionally, in our experience CEs do not require explicit student training as concept maps do (Regis and Albertazzi, 1996; Stoddart *et al.*, 2000; Francisco *et al.*, 2002).

Other forms of open-ended assessment techniques in chemistry education have been developed and utilized, but not discussed to the extent of concept maps. Zoller has

**Table 1** CE prompts used

Assignment	Maximum statements	Prompt	Topic
CE HW 1	5	An atom of Germanium-72	a) Structure of the atom
CE HW 2	7	7.5 g of $\text{CaBr}_2$ is dissolved in a 1.50 L solution of excess $\text{Li}_2\text{CO}_3$ in the reaction: $\text{CaBr}_2(aq) + \text{Li}_2\text{CO}_3(aq) \rightarrow \text{CaCO}_3 + \text{LiBr}$	d) Stoichiometry e) Solubility
CE HW 3	7	In the reaction below 23.0 g of $\text{FeCl}_2$ undergoes the reaction in 5.15 L of water initially at 25.0 °C (assume 1.0 g/mL). $\text{FeCl}_2(s) \rightarrow \text{Fe}^{2+}(aq) + 2\text{Cl}^-(aq)$ $H_f(\text{FeCl}_2) = -341.8 \text{ kJ/mol}$ $H_f(\text{Fe}^{2+}) = -87.9 \text{ kJ/mol}$ $H_f(\text{Cl}^-) = -167.46 \text{ kJ/mol}$	g) Thermodynamics
CE HW 4	5	$\text{H}_2(g) + \text{Cl}_2(g) \rightarrow 2\text{HCl}(g)$	i) Periodic trends c) Covalent bonds
CE HW 5	5	The ion $\text{SF}_5^-$	h) Molecular shapes and polarity
Assignment	Maximum statements	Prompt	Topic
CE Exam 1	8	33.5 g of $\text{CaCl}_2$	b) Ionic bonds
CE Exam 2	9	Reacting 223 mL of 0.15 M of HCl with excess magnesium results in the reaction below: $\text{Mg}(s) + 2\text{HCl}(aq) \rightarrow \text{H}_2(g) + \text{MgCl}_2(aq)$ This reaction occurs at 1.25 atm and 24°C	d) Stoichiometry f) Properties of ideal gases
CE Exam 3	8	In the reaction below 28 g of $\text{Cl}_2$ react with excess $\text{BF}_3$ : $2\text{BF}_3(g) + 3\text{Cl}_2(g) \rightarrow 2\text{BCl}_3(g) + 3\text{F}_2(g) \quad \Delta H = 1466.4 \text{ kJ/mol}$	g) Thermodynamics
CE Exam 4	8	$\text{COH}_2$ where C is the central atom Electronegativity values: C = 2.5, H = 2.1, O = 3.5	h) Molecular shapes and polarity c) Covalent bonds

developed and extensively explored the use of Higher Order Cognitive Skills (HOCS) questions in chemistry. HOCS questions are designed to require students to apply previous knowledge, theories and capabilities to unfamiliar situations (Zoller *et al.*, 1995), an ability described as critical thinking. Teaching strategies designed to promote HOCS type understanding led to student improvement on a HOCS based assessment (Zoller, 1993) supporting the use of HOCS to measure critical thinking skills. Scoring of HOCS questions for university students was found to have a correlation of 0.413 between two HOCS questions (Zoller *et al.*, 2002) providing an indication of consistency between the two measures and potentially indicating an underlying trait of critical thinking that HOCS oriented questions were designed to measure. The focus of HOCS on critical thinking based on chemistry knowledge instead of the formation of connections among the chemistry content within a course make it an unsuitable comparison to CEs.

Bowen (1997) described the use of chemical demonstrations to precede student assessment, and demonstrated learning gains as a result. Students were given open-ended questions pertaining to the demonstrations, but a scoring scheme for the open-ended questions was not described, except an indication that students were scored based on drawing reasonable conclusions from the data presented. Roecker (2007) described the utility of oral examinations as an assessment technique in chemistry. A four-point scoring scheme for assessing students' oral responses was proposed and it was found that students scored higher on this measure than on written examinations. While both the open-ended questions on chemical demonstrations and the scoring of oral examinations were designed to measure chemistry knowledge, neither study featured an investigation into the validity or consistency of the assigned scores and therefore cannot be used as a comparison.

## Setting and procedures

This study focused on first semester General Chemistry at a medium sized public university in the southeastern United States. Five of the twelve classes offered over the course of an academic year employed CEs as a form of assessment. Three of the five classes used primarily lecture based instruction and two of the five classes implement a hybrid, peer-led team learning (PLTL) with lecture reform in the class. The reform classes were included in the study to offer support for the generalizability of the validity results across different types of pedagogy.

Each class that employed CEs used five as homework assignments and four as in-class questions incorporated into a conventional exam. The inclusion of CEs within a conventional exam offered the benefits of ensuring that learning objectives were met via the conventional questions while also providing an opportunity for students to present their understanding of related concepts in the CE question. The homework (CE HW) and exam CEs (CE Exam) were spread throughout the semester. The prompts used with each CE are listed in Table 1. This study was approved by the university's Institutional Review Board and informed consent was administered to the five classes of General Chemistry receiving the CEs. 276 of the 350 enrolled students (78.9%) consented. Of those who consented, 66.7% were Caucasian, 6.5% were Asian, 5.8 % African American and 4.0% Hispanic with the remaining 17.0% unknown. The gender split within the sample was 61.8% female and 38.2% male.

In this study, each student's CE responses were photocopied and graded by each of the three authors. Two of the authors are regular instructors of General Chemistry and one is an upper-level chemistry student with career plans to be a secondary-level chemistry teacher. Prior to administering each CE, the three graders created a common

rubric for each assignment based on expected answers. Then, during grading, each grader added to the rubric independently as they came across additional, viable student answers. CEs were scored by each grader independently, leading to 100% overlap among graders, and no grader had knowledge of the scores assigned by the other graders. At the end of the semester the students took the First-Term General Chemistry from the American Chemical Society (ACS) Examinations Institute (2002) as the final exam for their course. This exam is externally constructed and nationally available through the ACS Examinations Institute and has a Cronbach's  $\alpha$  of 0.85 with the sample of interest. The ACS exam is also readily available for critique by other researchers and is generally recognized as an appropriate measure of chemistry knowledge.

## Results and discussion

An important issue with any proposed chemistry assessment is the question of validity. Messick (1995) describes six different aspects of validity to be considered, which are summarized as:

- Content – to ensure all parts of the content domain are represented
- Structural – the scoring criteria match the theory of the construct
- Generalizability – correlation of assessed tasks with other tasks and generalizable across time or observers
- External – correlation with other assessments reflects the expected relationship
- Substantive – the respondents are engaged in the intended process
- Consequential – evaluating the consequences of score interpretation

The purpose of this article is to present evidence pertaining to four of the six aspects: content, structural, external and generalizability.

### Content

The content aspect of validity seeks to ensure that all parts of the content domain are represented. The content domain is defined as the topics covered during the first semester General Chemistry sequence:

- a. Structure of the atom
- b. Nature of and naming of ionic chemical bonds
- c. Nature of and naming of covalent chemical bonds
- d. Stoichiometry and mass relationships in chemical reactions
- e. Characteristics of solubility, acid/base and redox reactions
- f. Properties of ideal gases
- g. Thermodynamics and heat relationships
- h. Molecular shapes and molecular polarity
- i. Periodic trends

Table 1 shows each CE prompt and the link with the relevant learning topics. Since CEs are deliberately open-ended, students can use additional topics that were presented in class. The topics listed in the table were the intended target for each CE, and the time of the CE

administration corresponded with the introduction of these topics. The match between CE prompts and topics in the course is indicative of the content validity of CEs.

### Structural

The scoring criteria for CEs are designed for the intended tasks of promoting students' connections of the content covered in class. First, students are prompted, and receive credit based on the number of statements that are correct, distinct and relevant to the prompt given in the question. Several threats exist to students' circumventing the intended goal, which are addressed in the scoring scheme. First, students may repeat a similar calculation several times in an attempt to receive full credit while making a relatively limited number of connections. For example, in any of the prompts featuring the mass of a compound and a chemical reaction, students could solve for the mass of every other compound. To address this possibility, students are informed that the distinct criterion means that performing several similar operations will only count as one statement. The scoring procedure also follows suit on this prompt, as shown in the grading of Student 1 in Appendix B.

Second, students may choose to include information from outside of the class content to reach full credit. This threat primarily exists on the homework CEs as the exam CEs are in-class, and students' are not permitted access to outside information during the exam CEs. To address this threat, students are informed on the homework CEs: "*Each statement you use should refer to material that has been presented in this course. You can use outside information (such as other reference material) but that will only count as one statement, regardless of how much information is presented from other sources.*" The scoring procedure then follows suit, ensuring that the strong majority of points awarded are for presenting information that has been presented in class.

Finally, students may use logic schemes to create additional statements without the use of additional content, namely by providing overly-general statements. For example, on CE HW 2, a student may write that the 7.5 grams of  $\text{CaBr}_2$  in the prompt is less than 10.0 grams of  $\text{CaBr}_2$ . Here the distinct and relevant criteria are employed, and students are informed that many samples are less than 10.0 grams and no relevant information to the original prompt has been added. Another example, from the same prompt, is for a student to indicate that the reaction is not single replacement. Again, as many reactions are not single replacement, students do not receive credit for this response. Another example would be students covering all bases using contradictory statements, such as 'this reaction is endothermic' and 'this reaction is exothermic' in the same response, where students would receive credit for neither.

As a result of the use of the criteria of correctness, distinctness and relevance, students' scores on CEs reflect the amount of information that students' can present that is both relevant to the course material and to the given prompt. For structural validity, the scoring criteria provide credit

when students relate material presented throughout the course to the given prompt, thereby promoting students' recognition of connections throughout the course.

### External

External validity was examined through investigating the relationship between student performances on CEs with a separate, distinct assessment. The assessment chosen was the ACS Exam, because all the instructors at the setting agreed that it was a suitable measure of student knowledge of the topics described above. As a result, there is an expectation for convergence between scores on CEs and scores on the ACS Exam, as both are designed to measure chemistry knowledge pertaining to the same sequence. The convergence is expected to be moderate owing to the different methods by which the exams measure knowledge; where CEs reward students for forming connections in topics throughout the course, the ACS Exam measures student knowledge in separate multiple choice questions.

As a result there is an expected correlation of 0.50 between CEs and ACS Exams, which would represent moderate agreement. Correlations much larger may represent a redundancy between the two measures, while correlations much smaller may call into question whether student scores on CEs do reflect student chemistry knowledge. It is noteworthy that the value of 0.50 represents the typical correlations witnessed in reviews of the literature between concept maps and conventional science assessments (though the range is considerable) (Liu and Hinchey, 1996; Ruiz-Primo and Shavelson, 1996; Rice *et al.*, 1998). The correlations between student performance on CEs and on the ACS Exam for each grader are shown in Table 2.

By examining Table 2 it is apparent that CEs administered as homework questions have a markedly lower correlation with the ACS Exam compared to the CEs administered as part of in-class exams. This is likely a result of the differences in administration. As a homework assignment, students had access to resources including class notes, textbooks and other students, as opposed to the exams where no other resources were permitted. In addition, students had many days to complete the homework CEs, but the exam CEs were included as a component of a timed test. The in-class CEs regularly had correlations at or near the 0.50 benchmark. This would suggest that, as an alternative assessment, CEs administered in-class feature a similar claim to external validity as concept maps.

Another explanation for the difference in correlations is that the homework CEs had a possible ceiling effect, where too many students reached the perfect score. Across the five homework CEs, students reached the maximum number of statements possible, respectively, 70.0%, 51.5%, 29.4%, 58.7% and 74.7% of the time. In contrast, across the four in-class exam CEs, students reached the maximum number of statements possible 1.9%, 0.8%, 3.1% and 10.5% of the time. The ceiling effect hinders correlations with other variables, as there is no discrimination among the large group of students with the perfect score on the assignment.

**Table 2** Correlations of CEs with ACS Exam

Assignment	Grader 1, ACS Exam	Grader 2, ACS Exam	Grader 3, ACS Exam
CE HW 1	0.145	0.142	0.088
CE HW 2	0.257	0.258	0.316
CE HW 3	0.424	0.393	0.329
CE HW 4	0.213	0.120	0.039
CE HW 5	0.121	0.199	0.157
Homework CEs average correlation	0.232	0.222	0.186
CE Exam 1	0.503	0.486	0.542
CE Exam 2	0.562	0.584	0.522
CE Exam 3	0.515	0.491	0.435
CE Exam 4	0.505	0.464	0.434
Exam CEs average correlation	0.521	0.506	0.483
Overall average correlation	0.361	0.349	0.318

This effect is particularly evident with CE homework assignments 1, 4 and 5 which had both the most students at the maximum number of statements and the lowest correlation with the ACS Exam. One suggestion may be to increase the maximum number of statements required for the homework CEs, which would make it more difficult for students to achieve the maximum number of statements. The use of a maximum score for each CE was in place so that instructors could limit the points awarded to any individual assignment. Another suggestion is to view the homework CEs as an assessment where students are rewarded for effort, where the students' score does not necessarily reflect their chemistry knowledge. Then, in-class exam, CEs would be merit-based and reflect a student's chemistry knowledge. That said, it seems prudent to retain the homework CEs, as they provide practice for the students prior to exposure with the in-class exam CEs. Otherwise, removing the homework CEs might affect the correlations witnessed among the in-class CEs.

### Generalizability

The next aspect of validity examined is whether scores are generalizable across graders. Owing to the open-ended nature of CEs, a variety of correct student responses is possible. This raises the question as to whether CEs can be graded consistently across different graders. This is an important question for a new assessment, as students' grades should not be dependent on who performed the grading. Our first concern was for the ranking of students within a class by CEs. This ranking is frequently the determining factor in assigning grades, and therefore it is important to examine whether each of the three graders in this study were consistent in how the students were placed relative to their peers. To measure the consistency in rankings, the intra-class correlations between graders were examined. In concepts maps rater agreement was found to have a generalizability coefficient ranging from 0.23 to 0.76 (McClure *et al.*, 1999). The generalizability coefficient is similar and comparable to the intra-class correlation

**Table 3** Inter-rater correlations

Assignment	Grader 1, Grader 2	Grader 2, Grader 3	Grader 1, Grader 3
CE HW 1	0.569	0.877	0.527
CE HW 2	0.906	0.87	0.893
CE HW 3	0.889	0.86	0.878
CE HW 4	0.720	0.747	0.772
CE HW 5	0.734	0.865	0.747
Homework CEs average correlation	0.764	0.844	0.763
CE Exam 1	0.898	0.887	0.894
CE Exam 2	0.884	0.860	0.882
CE Exam 3	0.899	0.855	0.863
CE Exam 4	0.861	0.778	0.791
Exam CEs average correlation	0.886	0.845	0.858
Overall average correlation	0.818	0.844	0.805

coefficient (Shavelson *et al.*, 1989). As a result, intra-class correlations above 0.800 were considered as an indication that CEs could be scored consistently to the extent that has been shown with concept maps. The average intra-class correlations, using the two-way random approach (Shrout and Fleiss, 1979), among each pair of graders are shown in Table 3. The decision to evaluate each pair of graders facilitates a direct comparison with the values found with graders of concept maps.

CE HW 1 has the lowest correlations among graders, with values at 0.527 and 0.569. This is likely a result of being the first CE to be graded and represents inexperience on the part of the graders. This would seem to suggest that a grader new to CEs may benefit from a practice run, either grading CEs that have already been collected or by assigning students a low-stakes CE assignment to give both the grader and students practice with the new assessment. CE HW 4 and CE HW 5 had correlations between 0.700 and 0.800, possibly indicative of the ceiling effect. Three of the in-class exam assignments (CE Exam 1, CE Exam 2, CE Exam 3) had correlations consistently over 0.800, with the majority above 0.880. The only in-class exam correlation that was below 0.800 was CE Exam 4 where two of the three correlations were below 0.800 (one was 0.778 and the other 0.791). Overall CE rater agreement matched or exceeded the top level of rater agreement that has been observed with concept maps and therefore indicates comparable generalizable validity.

Moreover, there is also interest in the absolute agreement between graders. This would be important in situations that have many graders within one class, for example, in large classes with teaching assistants sharing the work. Cohen's Kappa statistic was used to provide the measure of agreement between graders, with values greater than 0.400 representing good agreement between graders (Stoddart *et al.*, 2000). With rating concept maps, Kappa values have been found to range from 0.45 to 0.48 for individual variables (Stoddart *et al.*, 2000). Average Kappa values for each pair of graders in this study are given in Table 4.

**Table 4** Cohen's Kappa values

Assignment	Grader 1, Grader 2	Grader 2, Grader 3	Grader 1, Grader 3
CE HW 1	0.295	0.558	0.327
CE HW 2	0.484	0.398	0.398
CE HW 3	0.445	0.442	0.434
CE HW 4	0.315	0.378	0.367
CE HW 5	0.560	0.595	0.551
Homework CEs average Kappa	0.420	0.474	0.415
CE Exam 1	0.482	0.484	0.510
CE Exam 2	0.392	0.384	0.437
CE Exam 3	0.494	0.397	0.400
CE Exam 4	0.463	0.299	0.314
Exam CEs average Kappa	0.458	0.391	0.415
Overall average Kappa	0.437	0.437	0.415

Kappa values for the majority of the combinations were above the 0.400 threshold with most between 0.400 and 0.500. This level of inter-rater agreement, in particular with the in-class exam questions, is similar to the values observed with concept maps (Stoddart *et al.*, 2000). Still there were 12 of the 27 combinations (9 assignments and 3 comparisons between each grader) where Kappa values fell below the 0.400 threshold, although 4 of those 12 were between 0.390 and 0.400. When examining the disagreements between graders, it was found that the very strong majority of disagreements were a difference of one point. In short, except for CE HW 1 and CE Exam 4, over 85% of the grades between two graders were in either perfect agreement or differed by only one point. While the level of agreement for most assignments was high, CE HW 1 and CE Exam 4 may benefit by revising the assignment or rubric. Additionally, as has been mentioned, CE HW 1 may benefit by having the graders practise on a sample set of assignments first.

## Conclusions and future work

The validity of CEs in terms of assessing students' knowledge of general chemistry has been investigated across four aspects. CEs can be designed and scored to cover all the topics of a first semester general chemistry course, and to promote students' relating concepts within the course, indicating content and structural validity. The evidence resulting from this investigation demonstrates that in the research setting with this sample of students, CEs, particularly those given in-class, feature external validity by matching the expected correlation with a traditional measure of chemistry achievement. The relationship was consistent with that observed between concept maps, another assessment designed to promote students' connections within content, and traditional science assessment. Similarly, students' scores were shown to be generalizable across raters to a similar extent as concept maps.

The two remaining aspects of validity, substantive and consequential are suitable areas for future work. Substantive validity would focus on the extent CEs require students to connect concepts throughout the class and could be investigated by the qualitative analysis of students' responses. Cursory evidence indicating students' connecting concepts is presented elsewhere (Lewis *et al.*, 2010), but a more complete investigation is warranted to provide evidence for substantive validity. Consequential validity is also a suitable area for future investigation. Ultimately, the scores on CEs are used for determination of student promotion into follow-on courses in chemistry. Investigating whether CE scores relate to performance in the second semester of General Chemistry could establish consequential validity for this assessment.

The result that CEs, which measure the number of correct connections and interconnections, feature a correlation to traditional measures of science achievement on par with concept maps, has theoretical implications. It has been proposed here that evaluating the structure or organization of concept maps hindered concept maps as an assessment tool in that several organizations may be plausible. The external validity results of CEs lend some support to this assertion, though additional work to corroborate these results is necessary. In contrast, other researchers have proposed that the overall organization of concept maps should be the focus of assessment, with some evidence of validity in assessing the overall structure (Liu and Hinchey, 1996). Whether successful constructivist learning is best measured by the connections and interconnections or the overall organization of knowledge remains an open question.

This investigation also provided several recommendations for the administration of CEs. First, there is evidence that a practice run may benefit the CE graders, and it is suggested that either graders practice by grading a set of student responses previously collected or instructors use a one-time practice assignment to provide students and the instructor an opportunity to practice the assessment before it is used more formally. In this practice run, students may be given credit for simply completing the assignment and the instructor could provide feedback so students can become familiar with their instructors' expectations. Second, the evidence supporting the validity and reliability for CEs was stronger with in-class exam CEs than with the homework CEs. This may affect instructor decisions of how to weigh each type appropriately when determining contributions to an overall grade. It would not seem advisable to discard the homework CEs completely though, because the homework CEs provide experience for the students prior to encountering the more high-stakes exam CE question.

One of the goals of this project was that an investigation into validity could support an instructors' decision to implement CEs. As a result of this project, additional anecdotal findings also support the decision to implement CEs as an assessment technique. In scoring students' answers to CEs, there was clear evidence of students'

retaining concepts that were presented earlier in the course. This retention of earlier concepts was not clear in traditional assessments, and future work may want to examine whether CEs do in fact lead to retention of concepts. Scoring CEs also offered insights into incorrect links students made across concepts, which would not have been witnessed in conventional assessments. For example, in CE HW 4, some students attempted to fit the  $\text{H}_2 + \text{Cl}_2 \rightarrow 2 \text{HCl}$  reaction into the Born-Haber cycle description for ionic compounds. This insight can be used to guide instruction, in particular, in emphasizing the limits of models and equations. Finally, reported elsewhere, a survey of students' impressions indicated a positive response to CEs with a strong majority responding favorably (Lewis *et al.*, 2010). There is also anecdotal evidence of CEs altering students' study practices; in particular, as students were observed brainstorming possible answers for potential in-class CEs to be given. These findings suggest future research projects into the effect of CEs on students' motivation and study approaches.

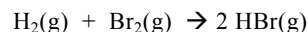
## Acknowledgements

Partial support for this work was provided by the National Science Foundation's Course, Curriculum, and Laboratory Improvement (CCLI) program under DUE-0941976. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors also wish to acknowledge the assistance of Dr. Angela Blaver for assistance in developing the theoretical framework discussed.

## Appendix A: Sample Assignment

### Assignment (CE HW 4)

Write down as many correct, distinct, and relevant facts you can about:



You'll receive three points for each statement. Five statements will get you full credit for the problem. Recall the information you use should be information that has been presented in class. All outside information, combined, will only count as one distinct fact toward your five.

### Scoring rubric

- All are covalent compounds
- HBr is polar covalent
- Calculation of  $\Delta \text{EN}$  or drawing dipole on molecule
- Lewis electron dot symbols of any species
- Ground state electron configurations or valence electron count
- Properties of species based on covalency of bonding (gases at room temperature, mp/bp/non-electrolyte)
- Any indication of single bonds
- Bond lengths, any of the following: 74 pm for H-H; 228 pm for Br-Br; 141 pm for H-Br
- Bond energies, any of the following: 432 kJ/mol for H-H; 193 kJ/mol for Br-Br; 363 kJ/mol for H-Br



- $\Delta H$  of reaction from bond energies: broken – formed = -101 kJ
- Exothermic reaction
- Label the reaction as either: synthesis reaction or redox reaction
- H is oxidized, Br is reduced
- Naming of any compound or HBr is an acid
- Number of protons or neutrons or place on periodic table
- Calculation of molar mass ( $H_2 = 2.02$ ,  $Br_2 = 159.8$ ,  $HBr = 80.8$ )
- Reaction represents heat of formation of HBr

## Appendix B: Sample student responses

Each student response is reported in verbatim, except for the inclusion of the line numbers and the descriptions given in square brackets.

### Student 1

Line 1:  $\Delta H = \text{bonds broken} - \text{bonds formed} = (432 + 193) - (2 \times 363) = -101 \text{ kJ/mol}$

Line 2: Hydrogen has 1 valence electron

Line 3: Br has 7 valence electrons

Line 4: [Correct Lewis structure for  $H_2$ ]

Line 5: [Correct Lewis structure for  $Br_2$ ]

Line 6: [Correct Lewis structure for HBr]

Line 7: HBr is an ionic bond

**Notes:** Student 1 received credit for correct bond energy (line 1) and for solving the  $\Delta H$  of reaction from bond energies (line 1). The student also received credit for number of valence electrons in line 2, but not line 3 owing to the distinct criteria. Similarly, the student received credit for line 4 but not for lines 5 and 6. Finally the student did not receive credit for line 7 as it is incorrect. In total, the student received credit for four statements (two from line 1, one each from line 2 and 4).

### Student 2

Line 1: This is a redox reaction

Line 2: The molar mass of  $H_2$  is 2.02 g/mol

Line 3:  $H_2$  is a nonpolar covalent compound

Line 4: HBr is polar

Line 5: HBr is 1.2% by mass hydrogen

Line 6:  $Br_2$  is being reduced

Line 7: H is losing electrons

**Notes:** Student 2 received credit for lines 1 through 5 as each represent distinct statements. The decision to treat line 3 and line 4 as distinct is evident in the scoring rubric and was made as the covalent label can be applied with just the non-metal to non-metal definition while the polar covalent label required the introduction of electronegativity. The student would have also received credit for line 6 but the maximum statements for full credit were five statements (see original assignment). The student would not have received credit for line 7 as it was not distinct from line 6.

## References

- Abraham M. R., (2008), Importance of a theoretical framework for research, in D. M. Bunce and R. S. Cole (eds.), *Nuts and bolts of chemical education research*, Washington: Oxford University Press, pp. 47-66.
- Barenholz H. and Tamir P., (1992), A comprehensive use of concept mapping in design instruction and assessment, *Res. Sci. Technol. Educ.*, **10**, 37-52.
- Bernal P. J., (2006), Addressing the philosophical confusion of regarding constructivism in chemical education, *J. Chem. Educ.*, **83**, 324-326.
- Biggs J. B., (2001), The revised two-factor Study Process Questionnaire: R-SPQ-2F, *Brit. J. Educ. Psychol.*, **71**, 133-149.
- Bowen C. W. and Phelps A. J., (1997), Demonstration-based cooperative testing in general chemistry: a broader assessment-of-learning technique, *J. Chem. Educ.*, **74**, 715-719.
- Cakir M., (2008), Constructivist approaches to learning in science and their implications for science pedagogy: a literature review, *Int. J. Env. Sci. Educ.*, **3**, 193-206.
- Derbentseva N., Safayeni F. and Canas A. J., (2007), Concept maps: experiments on Dynamic Thinking, *J. Res. Sci. Teach.*, **44**, 448-465.
- Examinations Institute of the American Chemical Society Division of Education, (2002), *First term general chemistry*, Milwaukee, WI: University of Wisconsin-Milwaukee.
- Francisco J. S., Nakhleh M. B., Nurrenbern S. C. and Miller M. L. (2002), Assessing student understanding of general chemistry with concept mapping, *J. Chem. Educ.*, **79**, 248-257.
- Holme T., Bretz S. L., Cooper M., Lewis J. E., Paek P., Pienta N., Stacy A., Steven R., Towns M. H., (2010), Enhancing the role of assessment in curriculum reform in chemistry, *Chem. Educ. Res. Pract.*, **11**, 92-97.
- Jones M. G., Carter G. and Rua M. J., (2000), Exploring the development of conceptual ecologies: communities of concepts related to convection and heat, *J. Res. Sci. Teach.*, **37**, 139-159.
- Lewis S. E., Shaw J. L. and Freeman K. A., (2010), Creative exercises in general chemistry; a student-centered assessment, *J. Coll. Sci. Teach.*, **40**, 48-53.
- Liu X. and Hinchey M., (1996), The internal consistency of a concept mapping scoring scheme and its effect on prediction validity, *Int. J. Sci. Educ.*, **18**, 921-937.
- Markow P. G. and Lonning R. A., (1998), Usefulness of concept maps in college chemistry laboratories: students' perceptions and effects on achievement, *J. Res. Sci. Teach.*, **35**, 1015-1029.
- McClure J. R., Sonak B. and Suen H. K., (1999), Concept map assessment of classroom learning: reliability, validity, and logistical practicality, *J. Res. Sci. Teach.*, **36**, 475-492.
- Messick S., (1995), Validity of psychological assessment; validation of inferences from persons' responses and performance as scientific inquiry into score meaning, *Amer. Psychologist.*, **50**, 741-749.
- Novak J. D. and Gowin B. B., (1984), *Learning how to learn*, Cambridge: Cambridge University Press.
- Regis A. and Albertazzi P. G., (1996), Concept maps in chemistry education, *J. Chem. Educ.*, **73**, 1084-1088.
- Rice D. C., Ryan J. M. and Samson S. M., (1998), Using concept maps to assess student learning in the science classroom: must different methods compete?, *J. Res. Sci. Teach.*, **35**, 1103-1127.
- Roecker L., (2007), Using oral examinations as a technique to assess student understanding and teaching effectiveness, *J. Chem. Educ.*, **84**, 1663-1666.
- Ruiz-Primo M. A. and Shavelson R. J., (1996), Problems and issues in the use of concept maps in science assessment, *J. Res. Sci. Teach.*, **33**, 569-600.
- Safayeni F., Derbentseva N. and Canas A. J., (2005), A theoretical note on concepts and the need for cyclic concept maps, *J. Res. Sci. Teach.*, **42**, 741-766.
- Santrock D. H., (2008), *Adolescence* (12th ed.), Boston: McGraw-Hill Higher Education.

- Scouller K., (1998), The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay, *High. Educ.*, **35**, 453-472.
- Shavelson R. J., Webb N. M. and Rowley G. L., (1989), Generalizability theory, *Am. Psychol.*, **44**, 922-932.
- Shrout P. E. and Fleiss J. L., (1979), Intraclass correlations: uses in assessing rater reliability, *Psychol. Bull.*, **86**, 420-428.
- Stoddart T., Abrams R., Gasper E. and Canaday, D., (2000), Concept maps as assessment in science inquiry learning – a report of methodology, *Int. J. Sci. Educ.*, **22**, 1221-1246.
- Trigwell K. and Sleet R., (1990), Improving the relationship between assessment results and student understanding, *Assess. Eval. Higher Educ.*, **15**, 190-197.
- Wright J. C., Millar S. B., Kosciuk S. A., Penberthy D. L., Williams P. H. and Wampold B. E., (1998), A novel strategy for assessing the effects of curriculum reform on student competence, *J. Chem. Educ.*, **75**, 986-992.
- Zoller U., (1993), Are lecture and learning compatible? Maybe for LOCS: unlikely for HOCS, *J. Chem. Educ.*, **70**, 195-197.
- Zoller U., Lubezky A., Nahkleh M. B., Tessier B. and Dori Y., (1995), Success on algorithmic and LOCS vs. conceptual chemistry exam questions, *J. Chem. Educ.*, **72**, 987-989.
- Zoller U., Dori Y. J. and Lubezky A., (2002), Algorithmic, LOCS and HOCS (chemistry) exam questions: performance and attitudes of college students, *Int. J. Sci. Educ.*, **24**, 185-203.