JOURNAL OF CHEMICAL EDUCATION

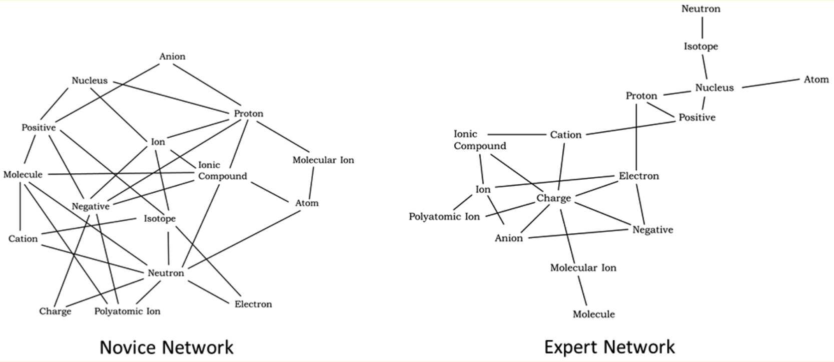
Establishing the Validity of Using Network Analysis Software for Measuring Students' Mental Storage of Chemistry Concepts

Kelly Y. Neiles,*[†] Ivy Todd,[†] and Diane M. Bunce[‡]

[†]Department of Chemistry and Biochemistry, St. Mary's College of Maryland, 18952 East Fisher Road, St. Mary's City, Maryland 20686-3001, United States

[‡]Department of Chemistry, The Catholic University, 620 Michigan Avenue, N.E. Washington, DC 20064, United States

Supporting Information



ABSTRACT: When designing a chemistry education research study it is important that the researcher select or create assessment instruments and methodologies that are valid. This ensures that correct interpretations can be made from the data collected using the instrument or methodology. The two studies described here introduce a novel method for investigating students' structural knowledge of chemistry topics using a program called Pathfinder. In these studies, the Pathfinder program is investigated for validity (content, construct, criterion, and concurrent), as well as its sensitivity to changes in students' structural knowledge. Once shown to be valid when used properly, the Pathfinder program can then be employed in future chemistry education research studies to investigate students' structural knowledge.

KEYWORDS: *Upper-Division Undergraduate, Chemistry Education Research, Testing/Assessment, General Public, Constructivism, Learning Theories*

FEATURE: Chemical Education Research

When designing a chemistry education research study, one of the most difficult tasks the researcher will encounter is the selection or creation of valid assessment instruments. Many assessments that are appropriate and useful for assigning grades in the classroom may not provide the more nuanced information often necessary for a research study. In recent years, researchers have started developing new assessments through rigorous instrument development.[1−5] These new assessments have included measurements of both sentiments (responses that do not have "correct" answers and reflect personal reactions, preferences, interests, attitudes, values, etc.) and judgments (responses which reflect student knowledge and understanding and have correct answers). Studies validating assessment instruments are invaluable to the chemistry education community as they provide quality instruments that can be reliably used in future research. In one example, Cooper, Underwood, and Hilley created an instrument to test students' beliefs about the information that can be obtained from Lewis structures.[1] Because of the rigorous development of this instrument, it can be used in various future studies where student understanding of Lewis structures is of importance. Fewer studies have provided novel *methods* of assessing student understanding that could be used by other researchers with different topics instead of only with the topic for which they were created. This study was designed to investigate how an alternative assessment method for measuring students' structural knowledge can be tested for validity in a variety of situations and on various chemistry topics.

THEORETICAL FRAMEWORK

Testing Theory

When chemistry education researchers are looking to measure some trait or psychological construct (such as proficiency, knowledge, beliefs, etc.) in their students, they often run into the problem that no single approach to the measurement of any

Received: September 14, 2015
Revised: March 3, 2016

DOI: 10.1021/acs.jchemed.5b00748

trait is universally accepted.[6,7] They will also quickly realize that measurements of psychological constructs are always indirect, based on behaviors that are perceived as relevant to the construct under study. This means that the researchers can only attempt to measure knowledge through the student's ability to perform smaller tasks, for example, problem solving.

In chemistry education research, the psychological construct of interest is often student's understanding of various chemistry topics. Traditionally, the method chosen for measuring student's understanding of a chemistry topic is problem-solving, usually in the form of multiple choice questions, short answer questions, and/or essay writing.[8] While these have proven to be valid methods of assessment, problems arise due to the limitation of possible questions. That is, the questions may not be measuring a large enough sample of the student's behavior to provide a true representation of the measured construct (knowledge of chemistry topic). For example, consider the question in Figure 1:

1. Select the best Lewis structure for NH_4^+

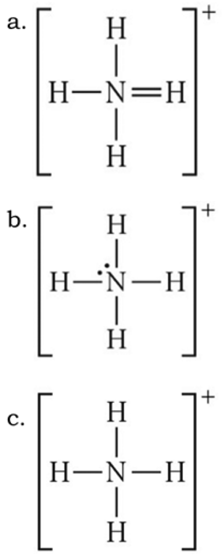


Figure 1. An example of a common general chemistry question. Reprinted with permission from Neiles.[9] Copyright 2010 American Chemistry Society.

A student who chooses the correct answer (c) may do so because they have an accurate understanding of the chemistry concept (drawing Lewis dot structures, polyatomic ions, formal charges, etc.). They may, however, select that answer because they believe that molecules will form certain structures because they 'like to be balanced or symmetrical', a common misconception identified in a study of Lewis structures by Cooper, Grove, Underwood, and Klymkowsky.[10] If this question were used in a research study to measure the student's knowledge on the subject, the student's answer may mislead the researcher to believe the student has an understanding of the subject, even though he/she still has some holes in their structural knowledge. A measurement method that includes additional information about how the student stores information structurally (as is proposed in this study), rather than a measure of the student's knowledge of individual facts may provide a more complete assessment of the student's structural knowledge of the chemistry topic.

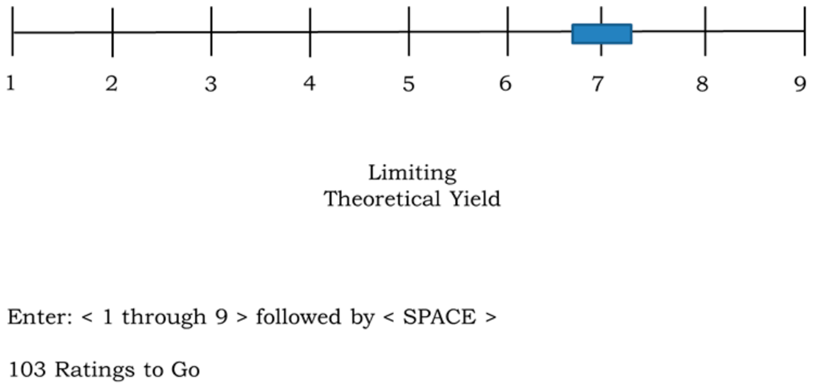
Structural Knowledge

For a student to be deemed competent in a chemistry topic, they must understand both the important concepts within that topic and the way those concepts interrelate. It has long been thought in cognitive and educational research that assessing the way a student structures their knowledge of the topic can indicate this competence.[11−13] The term structural knowledge refers to methods by which students file chemistry information (or any type of information) in their minds for later retrieval and use in a variety of situations.[14] It is a term based on the theory of schema described by numerous cognitive psychologists, which describes a person's knowledge through the way this knowledge is stored in the mind.[15−17] To acquire new information, the person must assimilate new material into his/her structural knowledge of the topic. The resulting structural knowledge can be thought of as an organized network of the information a student has learned, a topic that can be used in a wide variety of situations (solving problems, interpreting incoming information, etc.). This structure includes many complex connections between concepts within a topic and has a higher order hierarchical structure that influences how the information is stored and later retrieved for use.

Currently, there is no measure that can be employed by an education researcher to create an exact replication of a student's structural knowledge. A quick review of the *Journal* provides many examples of the traditional methods of measuring knowledge: multiple choice or open ended questions (either written or oral).[18−22] These methods provide information as to what facts or pieces of information a student has been able to store in their minds, but less information as to how the student has connected those pieces of information to one another (though open ended questions or essays do occasionally provide some of this information). In recent years, there has been a move to identify innovative assessment methods that go beyond the traditional tests.[23−25] In these studies, the researchers have utilized things like Web-based tutors, tablet-based PCs and pen-based technology, and assessment systems designed to recognize and respond to free-form student input.

One additional method of assessment utilized in the field of chemistry education and related to the new method presented in this paper is concept mapping.[9,26−28] A concept map is a graphical representation consisting of nodes and lines, where nodes are labels for important concepts (often keywords or terms) in a certain topic. As an assessment tool, researchers use concept maps created by the students as a measure of the students' understanding of chemistry concepts and/or how these understandings changed through the use of alternative study and assessment techniques. By using concept maps as assessments, researchers can often create more detailed pictures of a student's understanding of chemistry topics than a content test can provide. The use of concept mapping can, however, pose certain problems for the researcher. One issue is that the student must be able to evaluate his or her own understanding of a chemistry concept and try to reflect that understanding in the concept map task. This degree of reflection is often something that must be taught to the student through modeling by the instructor. The student also has to subjectively evaluate his or her own understanding and recreate that understanding in some type of concept map format.[27] This introduces a degree of subjectivity into the data collection even if that subjectivity is coming from the students themselves.

Another point of subjectivity comes from the researchers' interpretations in scoring the concept map created by the



Limiting
Theoretical Yield

Enter: < 1 through 9 > followed by < SPACE >

103 Ratings to Go

Figure 2. Rating program used to collect relatedness judgments.

students.[27] The researcher must infer the meaning behind the student's choices in the concept map and place value on the connections the student chooses to create. A more detailed review of concept mapping in chemistry education research and how it relates to the use of this new instrument presented in this study can be found in Neiles.[9] The difficulties in the use of concept maps described here have led to a search for a more objective measurement of students' structural knowledge. One result of this search was a method of creating network representations of students' understanding involving the use of something called proximity data.

Alternative Method

Pathfinder is a computer program that was developed by Roger Schvaneveldt.[29−31] The use of Pathfinder to create representations of students' structural knowledge is described in detail in *Measuring Knowledge* by Neiles.[9] In Pathfinder, the underlying organization of the person's schema is captured through relation and similarity judgments. These judgments of the relatedness between the members of pairs of concepts within a topic have been shown to capture the underlying organization of the person's schemas of that topic.[32] This procedure produces a matrix of proximity values that represent the degree of relationships between a pair of concepts.

The proximity data is gathered by a corresponding rate program.[31] In the rate program, a student is asked to judge the relatedness or similarity between members of concept pairs (key terms selected for a given topic). The rate program presents the student with a full list of the key terms, then presents each term pairwise and asks the student to rate the two terms based on their relatedness on a scale of 1−9, where 1 is the least related and 9 is the most related as seen in Figure 2.

To ensure that the student is utilizing the full scale, before the student completes the rate program he/she is presented with the full list of terms from the chemistry topic. The student is then asked to look through the list of items to find the two he/she believes are most related and set that as their 9, and find the two most unrelated and set that as 0.

After the student has been given instructions, the program presents every term against every other term, resulting in about 105 relatedness judgments for a topic containing 15 key terms. This process provides a matrix of proximity values in which each value in the matrix represents the degree of a relationship between a pair of concepts. The algorithm in the Pathfinder program is then used to reveal the underlying dimensions of the student's structural knowledge on which the student's judgments were based and create a Pathfinder network representation of the structural knowledge.

A weight associated with the strength of the relationship between two key terms, referred to as nodes, is associated with each link.[19] This weight reflects the distance between the nodes. For example, consider the Pathfinder network in Figure 3 created by a student on the subject of animals.

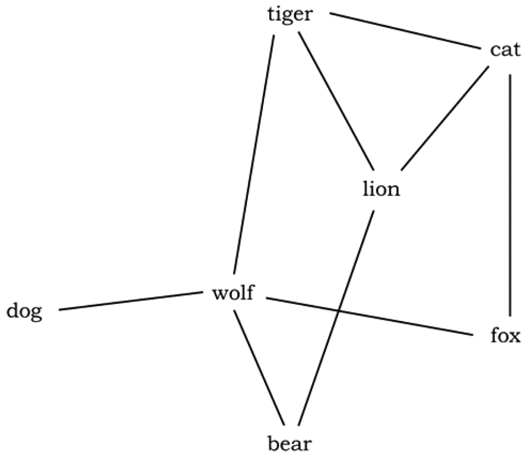


Figure 3. An example Pathfinder network created on the topic of animals.

Nodes in this network like lion and cat may be located closer to one another than lion and bear for various reasons. For example, many students may think about species similarities (lions and cats are both felines), while other students may consider the connections due to a certain well-known movie (lions, tigers, and bears!) and Pathfinder can accommodate both. The appeal of Pathfinder is that the researcher need not impose any of these reasons on the student. Pathfinder can find the underlying relationships in a student's mind regardless of what those relationships are based on. As can be seen through this example, the Pathfinder program provides evidence as to the existence of connections but does not provide the researcher any information as to the criteria a student uses to make their relatedness judgments (a limitation discussed further in the Limitations section of this paper). Nodes may be linked together directly (lion−bear) or through an indirect path (bear−lion−cat). The Pathfinder algorithm searches through the nodes to find the closest direct path between nodes or concepts. A link remains in the network only if it is the most direct path between two concepts. The most direct path could be a direct node to node link (which will result from a high relatedness judgment), or a multinode path depending

DOI: 10.1021/acs.jchemed.5b00748

on the student's responses during the rating task (which would result from a low relatedness judgment). All other links between those two key terms are removed from the network by the algorithm in the computer program.

The Pathfinder network is created by the algorithm using a q and r parameter equal to $n - 1$ and infinity, respectively, (where n = the number of concepts).[31] A path in the Pathfinder network consists of a number of nodes and connecting links. The length of a path defined by the r parameter is a function of the weights associated with the links in the path. This is calculated from the ratings provided by the student. As r decreases, links are added to the network. When the r is set to infinity, the number of links in the network is maximally reduced. The parameter q defines the maximum number of links in a path and also affects network density. The parameters $r = \infty$ and $q = n - 1$ generate the simplest Pathfinder network, which was desired for this study.

Since its development, Pathfinder has been used as an assessment tool in studies with people of all educational levels (elementary school through college/professional) and in numerous fields, including nursing, reading, patients with Alzheimer's, and many more.[33−37] Though the program has been tested for validity and reliability in many fields, it has not been widely used in chemistry education research and thus should be introduced and tested before use in this field.

Pathfinder networks can be analyzed both quantitatively, by mathematically comparing the number matrices created from the rate program, and qualitatively, by comparing the visual representations created by the Pathfinder program. Both methods provide important insights to the student's structural knowledge and understanding of the chemistry concepts. Unfortunately, while qualitative analysis of the results presented here would undoubtedly be interesting, providing a description of both quantitative and qualitative analysis would be too lengthy. The focus of this study is therefore mainly on quantitative analysis and leave qualitative analysis to subsequent papers.

The studies presented here looked at two variables used as comparison measures for the networks: Path Length Correlation (PLC) and Neighborhood Similarities (NS).[31] PLC is a measure of the presence and strengths of the links between the nodes in a network. While a network itself does not have an intrinsic score, comparing two or more networks can give researchers numerical data regarding the networks. Here, we compared the student network to an averaged expert network. Two networks can be similar or different in terms of the number and weights of the connections between nodes. This scoring provides evidence as to whether two networks have the same connections between nodes and whether those connections are of the same strength. PLC is determined by calculating the correlation between the distances in two networks, usually student vs expert.[31] Two networks with high PLC (a value close to 1) are from two people who rated concepts with similar strengths, whereas a low PLC (a value close to 0) would indicate dissimilar ratings.

Neighborhood Similarity (NS) is a measure of the way the nodes are grouped (or in "neighborhoods"). It is measured by determining whether a set of nodes surrounding a specific node in one network is the same set as around that specific node in another network.[31] If two networks have a high NS, it means that a specific node is surrounded by similar nodes in the two Pathfinder networks being compared. As in PLC, the NS measure ranges from 0 (low similarity) to 1 (identical).[31] Although these two measures are not the only two possibilities for determining the similarity of two networks, they are the most widely used since they provide quantitative numbers representing the quality and/or "expert-likeness" of student's mental storage of chemistry concepts.

Validity Testing

Prior to use in any research studies, a new assessment measurement or method, such as Pathfinder, must first be tested for validity to ensure its appropriateness for use in research. The validity of an assessment is the degree to which the assessment actually measures what it is designed to measure and whether the interpretation of the scores is appropriate.[38,39] There are four main types of validity generally evaluated: content, construct, criterion-related, and concurrent (though others exist, these are the four most widely used).

Content validity determines whether the assessment adequately samples the target domain; in Pathfinder's case, does it adequately test the chemistry topic of interest. For Pathfinder assessments, the content validity comes from the careful selection of key terms to be used as the nodes in the network to ensure that the key terms selected are relevant and representative of the chemistry topic of interest. Content validity must be determined for every new list of key terms used in a study.

Construct validity determines whether the assessment measures the construct (or trait) that it is said to measure. In a Pathfinder task, construct validity would be tested by evaluating the networks created by students and/or experts to determine whether they provide plausible connections between the key terms in a chemistry topic. Construct validity must be determined for each new list of key terms used in a study.

Criterion-related validity determines whether the test adequately predicts performance on the criterion of interest. To determine whether the Pathfinder task has criterion-related validity, its ability to predict student performance on a given chemistry topic should be evaluated.

Concurrent validity is whether the assessment correlates with another test that is designed to measure the same content or trait. To determine whether Pathfinder has concurrent validity, the correlation of its scores should be compared with another measure of the students' understanding of the chemistry topic.

When these four measures of validity are evaluated, a new assessment instrument or method can be deemed valid. The purpose of the two studies presented in this paper was to provide the reader with the process used to test the Pathfinder method for validity to show how Pathfinder data can be deemed to be valid using a variety of chemistry concepts.

■ STUDY 1

Study 1 was conducted to test the Pathfinder method for content validity, construct validity, and criterion-related validity through the evaluation of the following research questions:

1. Do expert's Pathfinder networks on chemistry topics adequately represent the chemistry topics? (Content and Construct Validity)
2. Can Pathfinder distinguish between low, medium, and high performing students based on their knowledge of general chemistry topics? (Construct, Concurrent, and Criterion-Related Validity)

DOI: 10.1021/acs.jchemed.5b00748

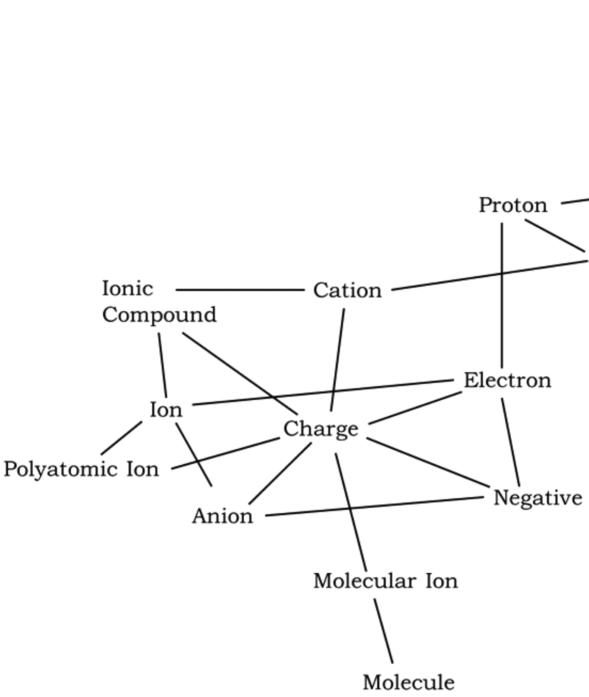


Figure 4. Averaged expert referent network for Atoms, Ions, and Molecules topic.

Methodology

Four topics were selected for use in this study and then narrowed down to two through the creation of referent expert networks. The original four topics were Atoms, Ions, and Molecules; Stoichiometry; Chemical Bonding; and Thermodynamics. An average of 15 key terms were created for each of the four original topics (a full list of the key terms for the studies presented here can be found in the Supporting Information). These were selected by cross referencing the list of vocabulary words found in several widely used general chemistry texts in the corresponding chapter for each of the four topics.[40−42] The resulting lists contained between 20 and 30 key terms which were then reduced to the target number of 15. Fifteen key terms results in 105 relatedness judgments made by each expert (or later with student participants). This number has been shown in previous studies to be the lowest number of key terms and relatedness judgments that results in accurate Pathfinder networks.[30] The set of topics and lists of key terms were reviewed by two chemistry instructors who had taught the topics recently. These instructors verified that the key terms used were both important to each topic and represented a wide range of important concepts in the topic.

Expert Networks. To give students' Pathfinder networks Path Length Correlation (PLC) and Neighborhood Similarity (NS) scores, they have to be compared to a referent expert network. To create this referent network, seven chemistry experts were involved in the study. These experts included undergraduate chemistry professors at a midsize private institution or chemists currently working in chemical research. The experts were asked to make relatedness judgment in the Pathfinder program for each list of key terms. The ratings were then used to create a Pathfinder network for each topic for a total of four networks per expert. The expert networks were then compared to one another to select the best two performing topics from the original four. They were compared on three criteria: coherency, PLC, and NS.

To determine which topics were best for use with subsequent phases of the study, each expert's four networks were first checked for coherency. Coherency is a reflection of the consistency of the data. The coherency of a set of proximity data is based on the assumption that the relatedness between a pair of items can be predicted by the relationships of the items to other items in the set. Very low coherency (below 0.20) may indicate that the participant is not an expert in that content area and should not be used as such.[43] Two experts' ratings were found to be below the necessary level for coherency on the Thermodynamics topic. This coupled with the fact that only six experts had completed the Pathfinder rating for this topic (one of the seven did not complete it due to time constraints) resulted in only four acceptable expert networks in the Thermodynamics topic. This topic was therefore not chosen as one of the two topics to be further used in the study. In the Chemical Bonding topic, one expert's network showed a low coherency score and was removed for the remainder of the analysis resulting in the Chemical Bonding topic having 6 expert networks.

The expert's Pathfinder networks were next averaged by the Pathfinder program to produce an overall average expert network for each topic. Figure 4 shows the visual representation of the average expert network for the Atoms, Ions, and Molecules topic.

Once the averaged network is created, the Pathfinder program can then mathematically compare each expert's network to this averaged referent network to determine whether an individual expert differs greatly from the group of experts on PLC and/or NS. This is done to ensure that no one expert would unduly influence the averaged network's structure.

The first measure, PLC, is similarity of path lengths between the expert's individual network and the referent experts' network. It was found that all expert/average expert correlations were large based on the correlation values generally accepted as large (r = 0.50−1.0).[44] The two topics with the highest average correlations were Atoms, Ions, and Molecules

(r = 0.800), and Stoichiometry (r = 0.662). The second measure, NS, assesses the similarity of the networks as a measure of the neighborhoods, or clusters of nodes, found within the two networks' corresponding nodes. The topics with the highest average NS's were Atoms, Ions, and Molecules (NS = 0.435), and Stoichiometry (NS = 0.438). These topics show NS values comparable to those reported as high in the literature.[30] Table 1 provides an overview of the topics' performance on the three measures of interest.

Table 1. Comparison of Topics to Determine Top Performing Two

Topic	Number of Participants with Appropriate Coherency	Average PLC Coefficients	Average NS	Selected Yes/No
Atoms, Ions, and Molecules	7	0.800	0.435	Yes
Stoichiometry	7	0.662	0.438	Yes
Chemical Bonding	6	0.656	0.312	No
Thermodynamics	4	X	X	No

The topics Atoms, Ions, and Molecules and Stoichiometry performed best in the number of experts with high coherency values, their average PLC, and their average NS values. On the basis of the results, these two topics were chosen for use in the student portion of the study. The two topics were also selected for further use in the study because they represented both a math based (Stoichiometry) and a nonmath based (Atoms, Ions, and Molecules) topic. These average referent experts' networks were used to determine the quality of each student's Pathfinder network.

Finally, the averaged networks for the two topics selected were given back to the chemistry experts for them to evaluate. To determine construct validity, the experts were asked to evaluate the averaged network as to whether the connections of the nodes in this averaged network adequately represented the topic. While every expert found small discrepancies in how the key terms were connected and how they felt they *should* be connected, on the whole they felt the average network was a good representation of the chemistry topic. This process was then repeated with a separate group of chemistry experts (faculty and industry chemists) with the results being the same. This indicates that the networks created by the Pathfinder program have construct validity.

In reference to RQ1, we find from the expert methodology and data analysis described above that the Pathfinder networks on chemistry topics do adequately represent the chemistry concepts, and thus, construct validity of this data was is established. We also find that through the careful selection and evaluation of key terms, the content validity of the subsequent networks can also be ensured.

Student Networks. The student participants were from a mid-sized private university in the mid-Atlantic. The students who participated were enrolled in or had been enrolled in any chemistry course offered at the university. This included introductory chemistry courses all the way up through forth-year advanced courses. With the use of a wide range of students, a varying range of structural knowledge was evaluated. Participants represented an age range of 18–23, as well as a variety of majors and both genders.

During the first session, students completed a consent form prior to beginning the study as per IRB protocol. The students were informed that participating in the study would not affect their grades in any chemistry class in which they were enrolled, but that it was very important for them to do their best on the task. Following the consent process, the students were briefed on the Pathfinder program. They first performed a practice relatedness task using animal key terms in order to get used to the rate program. After completing the practice task, the students performed the rate task on the relevant chemistry topics.

To evaluate RQ2 and establish concurrent validity, 43 students from a mid-sized private institution participated in the study. These students performed the Pathfinder task as part of a larger study to investigate students' use of chemistry texts. The Pathfinder tasks students completed covered the two topics selected through the expert portion of the study: (1) Atoms, Ions, and Molecules and (2) Stoichiometry. Figure 5 is an example of a student's pathfinder network on the Atoms, Ions, and Molecules topic.

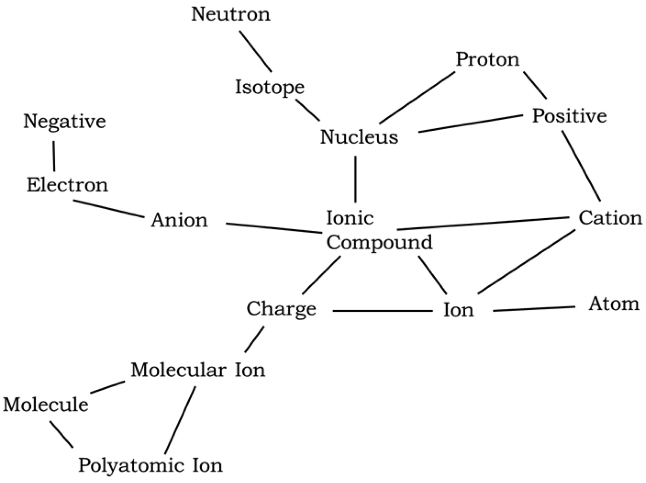


Figure 5. Student Pathfinder network on the topic of Atoms, Ions, and Molecules.

Each student's pathfinder networks were compared to the averaged expert networks on Path Length Correlation (PLC) and Neighborhood Similarity (NS) variables described previously.

As part of the larger study, students also completed a multiple choice test on the same two topics. The students' scores on the multiple choice tests were used to group them into low, medium, and high performing students. The multiple choice tests were created by pulling questions on the relevant topic from retired ACS general chemistry exams.[45] Ten questions were selected for each topic and then reviewed by chemistry faculty to determine whether the questions were relevant to the content area and/or covered the area sufficiently. A Multivariate Analysis of Variance (MANOVA) statistical analysis was completed with the high, medium, and low performing groups (based on multiple choice test score) used as an independent grouping variable and the PLC and NS evaluated as separate continuous dependent variables.

Results

MANOVA. Prior to running the MANOVA, the necessary assumptions were tested to ensure the data was appropriate for this statistical analysis. These assumptions included sample size, normality, linearity, homogeneity of covariance, and multi-

DOI: 10.1021/acs.jchemed.5b00748

colinearity.[46] In each case, the data passed the established test for that assumption. The mean scores for PLC and NS are provided in Table 2.

Table 2. Mean Scores for PLC and NS

Groups	Mean PLC Score	Mean NS Score
Low	0.255	0.222
Medium	0.507	0.348
High	0.685	0.385

The Wilk's Lambda multivariate test of significance was used to evaluate whether there was statistically significant differences among the groups (low, medium, and high) on a linear combination of the dependent variables (PLC and NS). Table 3 provides the results of this test.

Table 3. Results from MANOVA

Model Variables	Multivariate Test	Significance	Partial Eta Squared
Overall Model (two dependent variables combined)	Wilk's Lambda $F_{(4,78)} = 14.697$	$p < 0.001$	0.430
PLC	$F_{(2,40)} = 36.190$	$p < 0.001$	0.644
NS	$F_{(2,40)} = 13.986$	$p < 0.001$	0.412

The value of Wilk's Lambda ($F_{(4,78)} = 14.697$) and its associated significance level ($p < 0.001$) indicate that there is a significant difference among the groups. The effect size (partial η^2) indicates that 43.0% of the variance in students' scores was described by the linear combination of PLC and NS and was a large effect size.[44] Since significance was found in the Wilk's Lambda test, the dependent variables could next be investigated further individually.

To evaluate each of the dependent variables, the Tests of between Subjects Effects was considered (results also reported in Table 3). Looking at Path Length Correlation (PLC), we see that the test value ($F_{(2,40)} = 36.190$) is significant ($p < 0.001$). The Neighborhood Similarity (NS) test value ($F_{(2,40)} = 13.986$) was also found to be significant ($p < 0.001$). These values mean that significant differences between the groups (low, medium, and high) were found in both their PLC and NS scores. The effect sizes were also found to be large; 64.4% of the variance in students' scores described by PLC and 41.2% described by NS.

The between-subject significance found for PLC and NS means that Post Hoc analysis can be employed to determine where the significant differences lie. Tukey's test was selected for this analysis which required an additional assumption to be tested, the assumption of equal variances for the three groups. The data was analyzed for this assumption and passed. When the PLC was evaluated, all groups had significantly different PLC scores than all other groups ($p < 0.05$). With NS, significant differences ($p < 0.05$) were found between the low and medium groups' and also between the low and high groups. No significant difference was found between students' NS scores in the medium and high performing groups. These results indicate that the PLC and NS measures together are able to adequately differentiate between low, medium, and high performing students.

To test for concurrent validity, the PLC and NS scores were each evaluated in correlation with the chemistry content scores that students received through the multiple choice test. PLC was found to have a correlation of 0.847 with the multiple choice test and NS a score of 0.699. Both of these correlation scores are considered high.[44] This means that the Pathfinder networks correlate highly with the more traditional method, multiple choice tests, of assessing student knowledge and thus have concurrent validity.

The MANOVA and correlation results indicate that Pathfinder was able to distinguish between the chemistry content knowledge groups, and that both PLC and NS was correlated with students' multiple choice scores. This indicates that both construct and concurrent validity where validated; that is, Pathfinder measures the construct of interest and the assessment correlates with another test that is designed to measure the same content or trait.

Regression. A regression analysis was also performed to determine which, if either, of the two dependent variables was a better predictor of the student's performance on the multiple choice test (the grouping variable). This would indicate the assessment's criterion-related validity. The difference between the MANOVA test and the regression test is that in the MANOVA the multiple choice test is used as a grouping variable, while in the regression analysis, the multiple choice scores will be used as a continuous variable. With the use of the multiple choice scores as a continuous variable, more of the variance in students' scores may be captured in the statistical model. The assumption for the regression analysis is essentially the same as that for the MANOVA, and the data was determined to be appropriate for this statistical analysis.

First separate regressions of each dependent variable were evaluated to determine what percent of the students' multiple choice scores is predicted by each dependent variable. In the PLC regression, the test was found to be significant ($F_{(1,41)} = 17.606$, Sig < 0.001) and the PLC variable was found to have an adjusted R^2 value of 0.798. This means that the amount of variance in students' multiple choice scores predicted by the PLC variable, 79.8%, is a significant amount. In the NS regression, the test was found to be significant ($F_{(1,41)} = 24.156$, Sig < 0.001) and the NS variable was found to have an adjusted R^2 value of 0.609. This means that 60.9% of the variance in the students' multiple choice test scores is described or predicted by the NS variable. These results indicate that both variables are good predictors of the students' multiple choice test scores, indicating criterion-related validity.

Next, a combined regression was conducted to determine which, if either, of the variables was a better predictor of the students' multiple choice scores. In the combined regression, both variables were entered into the model simultaneously. As in the separate regressions, the combined regression was found to be significant ($F_{(2,40)} = 37.162$, Sig < 0.001). This means the variance in students' multiple choice scores predicted by the combined PLC and NS variables was a significant amount. The β score for PLC was found to be significant ($\beta = 0.974$, sig < 0.001), while the β score for NS was not significant ($\beta = 0.020$, sig = 0.232). This means that much of the variance predicted by PLC and NS overlaps between the variables and that while either variable can be considered a good predictor, if only one variable were to be used, PLC would likely be the better choice of the two. These regression results indicate that the Pathfinder measures, PLC and NS, adequately predict the students' multiple choice chemistry scores and thus have criterion-related validity.

DOI: 10.1021/acs.jchemed.5b00748

■ STUDY 2

Study 2 was conducted to test the Pathfinder method for content validity, construct validity, criterion-related validity, and concurrent validity through the evaluation of the following research questions:

1. When used as an assessment instrument, is the Pathfinder program sensitive enough to detect changes in general chemistry students' structural knowledge due to instruction? (Content and Construct Validity) (RQ1)
2. Is Pathfinder a good predictor of general chemistry students' understanding of a chemistry concept? (Criterion-related Validity) (RQ2)
   a. Is Pathfinder a better predictor of students' understanding than traditional multiple choice questions? (Criterion-related Validity) (RQ2a)

Methodology

Study 2 was conducted at a small public liberal arts college in the mid-Atlantic region. The students who participated were enrolled in one of three introductory chemistry courses, General Chemistry I and II or Contemporary Chemistry. General Chemistry is an introductory two-semester chemistry sequence mostly populated by science majors, including chemistry, biochemistry, and biology, but also taken by neuroscience minors. Contemporary Chemistry is a one semester introductory chemistry course taught to nonscience majors for a liberal arts general education requirement. Participants represented an age range of 18–23, as well as a variety of majors and both genders. A total of 42 students participated in Study 2.

During the fall, the study was conducted with the General Chemistry I population, and during the spring, students from both the General Chemistry II and Contemporary Chemistry populations participated. Student participation consisted of two sessions: one taking place before the lecture on a specific topic and one postlecture on that same topic. The topics were chosen to test the students on a broad range of introductory chemistry topics both math-based and conceptual (for example, colligative properties, acid–base equilibrium, phase diagrams, etc.). A total of 11 chemistry topics were utilized. Approximately 15 key terms were selected for use with each topic and validated in the same way described in Study 1 to ensure content validity (a full list of the key terms used in this study can be found in the Supporting Information).

The averaged expert referent networks were created for each chemistry topic through the same process described in Study 1. The experts consisted of faculty members from multiple institutions, all of whom had taught the topics of interest in the 2 years prior to the study. Construct validity was also determined in the same way by having additional experts evaluate the averaged referent networks for appropriateness and completeness. All chemistry topics except one passed this test and were thus determined to have construct validity and appropriate for further use in the study.

During the first session, students completed a consent form prior to beginning the study as per IRB protocol in the same way described in Study 1. Following the consent process, the students were briefed on the Pathfinder program. They first performed a practice relatedness task using animal key terms in order to get used to the rate program. After completing the practice task, the students performed the rate task on the relevant chemistry topic.

The first session usually took place the day before the relevant lecture and took between 15 and 25 min. Session two took place within 24 h after the lecture. Students were asked not to do any additional studying prior to their second session. During the second session, students performed the same rate task again on the relevant chemistry topic, along with a multiple choice task and an open ended question followed by a spoken explanation of their thought processes when solving the open ended question. The multiple choice tests were created by pulling questions on the relevant topic from retired ACS general chemistry exams.[45] Five questions were selected for each topic and then reviewed by chemistry faculty to determine whether the questions were relevant to the content area and/or covered the area sufficiently. The open ended questions were created to illicit student responses that spoke to their understanding of the chemistry concepts of interest. These were also reviewed by faculty members to ensure their relevance to the topic of interest.

The researchers took notes during the spoken explanation and scored the explanations and the open ended question answers themselves based on a rubric. The rubrics used to evaluate the open ended questions included scoring for the correct answer as well as the proper reasoning for that answer. The combined score of these was used as a measure of student understanding.

Results

To investigate research question one, a paired-samples *t* test was conducted to evaluate whether each of Pathfinder's two measures, PLCs and NSs, were sensitive enough to detect changes in general chemistry students' structural knowledge due to instruction. The paired-samples *t* test was selected because one group of students was utilized but data were collected from them on two different occasions in a pre-test/post-test experimental design. In statistical terms, this analysis tests the probability that the two sets of scores (prelecture vs postlecture) came from the same population. If the *t* test shows a significant difference, then the scores came from statistically different populations and thus would indicate a significant difference in students' Pathfinder scores before vs after they encounter the information in lecture.

PLC scores had a statistically significant increase in scores prelecture ($M = 0.248$, SD = 0.153) to postlecture ($M = 0.355$, SD = 0.187), $t_{(41)} = -3.642$, $p < 0.05$. The η^2 statistic (0.250) indicated a large effect size when compared to guidelines from Cohen.[44] This increased PLC score indicates that students' Pathfinder networks became significantly more like expert networks after attending lecture and the Pathfinder rate program was able to identify this change in students' structural knowledge.

NS had a statistically significant change in scores prelecture ($M = 0.210$, SD = 0.071) to postlecture ($M = 0.240$, SD = 0.091), $t_{(41)} = -2.034$, $p < 0.05$. The η^2 statistic (0.094) indicated a moderate to large effect size. This increased NS score indicates that students' Pathfinder networks became significantly more like expert networks in terms of the groupings of the nodes after attending lecture. It also indicates that the Pathfinder rate program was able to identify this change in students' structural knowledge.

In reference to RQ1, it was thus found that Pathfinder was sensitive enough to detect changes in general chemistry students structural knowledge due to instruction. It was also determined from the careful selection and validation of the key

DOI: 10.1021/acs.jchemed.5b00748

terms and resulting Pathfinder networks that the program, when used in this manner, has content and construct validity.

To investigate RQ2, regression analysis was utilized to determine which of the independent variables, namely multiple choice (MC), path length correlation (PLC), and neighborhood similarity (NS), best predicts the variance in the measure of students' understanding (open ended question combined with explanation). This analysis was chosen because both the dependent variable and the multiple independent variables were measured on continuous scales.

The first model evaluated was a standard multiple regression with all three independent variables (MC, PLC, and NS) entered into the model simultaneously as predictor variables. This allows us to evaluate the significance of the prediction of the dependent variable, as well as evaluate each independent variable. This first model resulted in a significant prediction of the students' overall understanding ($R = 0.488$, $F_{(3,38)} = 3.955$, $p < 0.05$). The percentage of variance in students' overall understanding may also be determine by evaluating the R^2 value, in this case 0.238 or 23.8%. This indicates that the model including all three dependent variables predicts about 23.8% of the variance in students' understanding (a percentage on par with other education research and cognitive psychology studies). Since the overall model is indeed significantly predicting the dependent variable, we can also evaluate each independent variable to determine which variables contribute the most to the prediction by evaluating their standardized coefficient or β score. In this model, MC had a $\beta = 0.216$ ($p = 0.149$), PLC had a $\beta = 0.222$ ($p = 0.238$), and NS had a $\beta = 0.203$ ($p = 0.288$). This indicates that no one independent variable accounts for a significant portion of the 23.8% variance predicted by the model when all three independent variables are entered into the model simultaneously. This ability to predict students' performance on a separate chemistry knowledge test also ensures the Pathfinder program has criterion-related validity. To understand exactly what portion of the variance in students' understanding scores is accounted for by each independent variable, stepwise multiple regressions were evaluated.

In the first stepwise regression evaluated, MC was entered into the model first and then PLC and NS were brought in which allows one to determine if the addition of PLC and NS contributed significantly to the model. The addition of PLC and NS was found to significantly change the amount of variance predicted by the model (R^2 change = 0.140, $F_{(2,38)} = 3.480$, $p < 0.05$). This indicates that PLC and NS contribute a significant amount to the prediction of students' understanding scores. In the second stepwise regression evaluated, PLC and NS were entered into the model first and then MC was brought in to determine if the addition of MC contributed significantly to the model. The addition of MC was found to not significantly change the amount of variance predicted by the model (R^2 change = 0.043, $F_{(2,38)} = 2.166$, $p = 0.149$). This indicates that MC does not contribute a significant amount beyond that of the combined PLC and NS to the prediction of students' understanding scores. These combined results indicate that the Pathfinder measures are better predictors of students' understanding than the traditional multiple choice tests.

■ DISCUSSION AND CONCLUSIONS

Study 1

The results of the MANOVA and regression analysis indicate that Pathfinder measures (PLC and NS) are able to identify differences between students' chemistry understanding at least as well as the multiple choice exams. This is evidenced in the ability of the variables to distinguish between the groups (MANOVA) and predict the variance in students' scores (regression). PLC and NS were also shown to correlate highly with another established method of testing student knowledge: multiple choice questions. The careful selection and validation of the key terms and averaged expert referent networks, MANOVA results, and correlation results indicate that the Pathfinder program's use with these key terms have content, construct, and concurrent validity. The regression analysis has shown that both measures of the quality of students' Pathfinder networks (PLC and NS) are good predictors of the students' performance on the chemistry topics and thus have criterion-related validity. It was also found that while either variable is a good predictor of students' chemistry knowledge (as measured by a multiple choice test), PLC may be a better choice since much of the variance in students' scores overlapped between the two variables and a greater amount was predicted by the PLC. However, if for some reason it is necessary to use only one variable, a researcher should evaluate the questions they are investigating to determine which variable they use. If, for instance, the researcher is particularly interested in how students are grouping the categories, then Neighborhood Similarity would provide more relevant information.

Study 2

Through the above analysis, it was found that the Pathfinder program was indeed sensitive enough to detect changes in general chemistry students' structural knowledge due to instruction as evidenced by the *t* test results which showed that both PLC and NS significantly increased postlecture. These results, along with the careful selection and validation of key terms and averaged expert referent networks, indicate that the Pathfinder program used with these key terms has content and construct validity. It was also found, through the regression analysis, that both PLC and NS were good predictors of students understanding of a chemistry concept and that this use of Pathfinder, therefore, has criterion-related validity. It was also found through the regression analysis that these Pathfinder measures were *better* predictors of students' understanding than a traditional multiple choice test.

The purpose of these studies was to ensure that when developed properly, the method of using rating tasks to create representations of students' structural knowledge can be a valid method and could be used confidently in further studies. This involved checking the method for content, construct, criterion-related, and concurrent validity. It should be further emphasized here, however, that this method can only be deemed valid when the careful creation of the lists of key terms, an implementation of the method, has taken place as described in these studies.

■ IMPLICATIONS

With the use of the methods described in the two studies here, the Pathfinder program has been found to be a valid method with these chemistry topics. This means that the Pathfinder program can be used in future chemistry education research as

DOI: 10.1021/acs.jchemed.5b00748

long as the steps necessary are taken to ensure the proper creation and implementation of the program. Unlike other measurement instruments presented in this *Journal*, this instrument is not restricted to any single chemistry topic. It can be used with any topic where key-terms can be identified and validated. The method may be used, for example, to test the students' understanding of a chemistry topic before and after a new lab to determine the lab's effectiveness. Or perhaps Pathfinder networks of important topics, such as acid−base chemistry, can be measured over the course of a student's academic career to monitor changes in his/her understanding. Another benefit of this method is the added information one can gain from evaluating Pathfinder networks over more traditional methods such as multiple choice. Researchers are better able to see the connections between concepts in a student's mental storage of a chemistry topic. Pathfinder also allows researchers to qualitatively evaluate the visual representations of the students' networks, though that evaluation was not described in the studies presented here. This new measurement method provides a different and perhaps more nuanced method of measuring student understanding in chemistry.

LIMITATIONS

The studies presented here could benefit from replication with larger numbers of students and at other institutions. The large size effects indicate the strength of our findings, but there is always room for replication with larger sample sizes. Also, there are some chemistry topics that were found not to perform well with the expert population and were thus not used in these studies. We are unsure why these topics did not perform. It may be that a larger number or different pool of experts could rectify the situation, or that there is no unified way experts organize their knowledge of those topics.

Pathfinder was used in these studies as a research tool to measure students' structural knowledge in general chemistry. It is not intended, as presented here, to be used in a classroom setting to assign students grades in a summative assessment capacity. However, it could certainly be used in a formative assessment to provide students with feedback on their progress in class.

As was stated earlier, an additional limitation of the Pathfinder program is that this method does not provide any information to the researcher about why or how the student chose to relate two key terms. Thus, the researcher has no way of knowing what criteria the student used to make their relatedness judgments. No one instrument can be expected to collect everything about how a student stores information for later retrieval and use. Additional data such as interviews or think-aloud protocol would be necessary if this information were needed by the researcher.

A consideration that should be investigated in addition to validity is the reliability of the assessment instrument. Reliability of an assessment instrument or method refers to the consistency of scores across replications of a measurement procedure.[47] This would be measured in Pathfinder by evaluating whether students' PLC or NS scores change over time. This process would likely be accomplished by an in depth qualitative study of the student results to determine whether changes over time were present or not. As the focus of this study was quantitative in nature, the reliability analysis has not been included, although it could be evaluated in future studies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.5b00748.

Complete list of the Pathfinder key terms used for each chemistry topic in these studies (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: kyneiles@smcm.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors wish to acknowledge the student and expert participants at both institutions involved in these studies. We also would like to thank Dr. Schvaneveldt for providing a copy of Pathfinder for use in these studies.

REFERENCES

(1) Cooper, M. M.; Underwood, S. M.; Hilley, C. Development and validation of the implicit information from Lewis structures instrument (IILSI): Do students connect structures with properties? *Chem. Educ. Res. Pract.* **2012**, *13*, 195−200.

(2) Chandrasegaran, A. L.; Treagust, D. F.; Mocerina, M. The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chem. Educ. Res. Pract.* **2007**, *8*, 293−307.

(3) Bowen, C. W. Development and score validation of a chemistry laboratory anxiety instrument (CLAI) for college chemistry students. *Educ. Psychol. Meas.* **1999**, *59* (1), 171−185.

(4) Cooper, M. M.; Sandi-Urena, S. Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *J. Chem. Educ.* **2009**, *86* (2), 240.

(5) Mulford, D. R.; Robinson, W. R. An inventory for alternate conceptions among first-semester general chemistry students. *J. Chem. Educ.* **2002**, *79* (6), 739−744.

(6) American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, 1999.

(7) Thorndike, R. L.,Ed.; *Educational Measurement*, 2nd ed.; American Council on Education: Washington, DC, 1971.

(8) Goldsmith, T. E.; Johnson, P. J.; Acton, W. H. Assessing Structural Knowledge. *Journal of Educational Psychology* **1991**, *83* (1), 88−96.

(9) Neiles, K. Y. Measuring Knowledge: Tools To Measure Students' Mental Organization of Chemistry Information. In *Tools of Chemistry Education Research*; Bunce, D. M.; Cole, R., Eds.; American Chemistry Society: Washington DC, 2014; Chapter 10.

(10) Cooper, M. M.; Grove, N.; Underwood, S. M.; Klymkowsky, M. W. Lost in lewis structures: An investigation of student difficulties in developing representational competence. *J. Chem. Educ.* **2010**, *87* (8), 869−874.

(11) Bower, G. H. A Selective Review of Organizational Factors in Memory. In *Organization of Memory*; Tulving, E.; Donaldson, W., Eds.; Academic Press: New York, 1972; pp 93−137.

(12) Collins, A. M.; Quillian, M. R. Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior* **1969**, *8*, 240−247.

(13) Shavelson, R. J. Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology* **1972**, *63*, 225−234.

(14) Goldsmith, T. E.; Johnson, P. J.; A Structural Assessment of Classroom Learning. In *Pathfinder Associative Networks: Studies in Knowledge Organization*; Schveneveldt, R. W., Ed.; Ablex: Norwood, NJ, 1990; pp 241−254.

(15) Bartlett, F. *Remembering*; Cambridge University Press: Cambridge, 1932.

(16) Mayer, R. E. *Thinking, Problem Solving, Cognition*; W.H. Freeman and Co.: New York, 1983.

(17) Anderson, J. R. *Cognitive Psychology and Its Implications*, 7th ed.; Worth Publishers: New York, 2010.

(18) Lewis, S. E.; Lewis, J. E. Departing from lectures: An evaluation of a peer-led guided inquiry alternative. *J. Chem. Educ.* **2005**, *82* (1), 135−139.

(19) Cooper, M. M.; Williams, L. C.; Underwood, S. Student understanding of intermolecular forces: A multimodal study. *J. Chem. Educ.* **2015**, *92* (8), 1288−1298.

(20) Boudreaux, A.; Campbell, C. Student understanding of liquid-vapor phase equilibrium. *J. Chem. Educ.* **2012**, *89*, 707−714.

(21) Bowen, C. W. Think-aloud methods in chemistry education. *J. Chem. Educ.* **1994**, *71* (3), 184−190.

(22) Roecker, L. Using Oral Examination as a technique to assess student understanding and teaching effectiveness. *J. Chem. Educ.* **2007**, *84* (10), 1663−1666.

(23) Stevens, R.; Soller, A.; Cooper, M.; Sprang, M. Modeling the Development of Problem Solving Skills in Chemistry with a Web-Based Tutor. In *Intelligent Tutoring Systems*; Springer: Berlin, Heidelberg, 2004.

(24) Bryfczynski, S. P.; Underwood, S. M.; Grove, N. P.; Pargas, R. P.; Cooper, M. M. OrganicPad as a Research Tool: Investigating the Development of Representational Competence in Chemistry. In *The Impact of Tablet PCs and Pen-Based Technology on Education*; Purdue University Press: West Lafayette, IN, 2010.

(25) Bryfczynski, S.; Pargas, R. P.; Cooper, M. M.; Kylmkowsky, M. BeSocratic: Graphically Assessing Student Knowledge n *IADIS International Conference Mobile Learning Proceedings*; IADIS: Berlin, Germany, 2012.

(26) Markow, P. G.; Lonning, R. A. Usefulness of Concept Maps in College Chemistry Laboratories: Students' Perceptions and Effects on Achievement. *J. Res. Sci. Teach.* **1998**, *35*, 1015−1029.

(27) McClure, J. R.; Sonak, B.; Suen, H. K. Concept Map Assessment of Classroom Learning: Relibaility, Validity, and Logistical Practicality. *J. Res. Sci. Teach.* **1999**, *36*, 475−492.

(28) Francisco, J. S.; Nahkleh, M. B.; Nurrenbern, S. C.; Miller, M. L. Assessing Student Understanding of General Chemistry with Concept Mapping. *J. Chem. Educ.* **2002**, *79*, 248.

(29) Schvaneveldt, R. W.; Durso, F. T. General Semantic Networks. Paper Presented at the Annual Meeting of the Psychonomic Society, Philladelphia, PA, 1981.

(30) Schvaneveldt, R. W.; Durso, F. T.; Goldsmith, T. E.; Bree, T. B.; Cooke, N. M.; De Maio, J. C. Measuring the structure of expertise. *Int. J. Man-Mach. Stud.* **1985**, *23*, 699−728.

(31) Schvaneveldt, R. W. *Pathfinder Associative Networks*; Ablex Publishing Corporation: Norwood, NJ, 1990.

(32) Gonzalvo, P.; Canas, J. J.; Bajo, M. T. Structural representation in knowledge acquisition. *Journal of Educational Psychology* **1994**, *86* (4), 601−616.

(33) Aronoff, J. M.; Gonnerman, L. M.; Almor, A.; Arunachalam, S.; Kempler, D.; Andersen, E. S. Information content versus relational knowledge: Semantic deficits in patients with Alzheimer's disease. *Neuropsychologia* **2006**, *44* (1), 21−35.

(34) Azarello, J. Use of the Pathfinder scaling algorithm to measure students' structural knowledge of community health nursing. *J. Nurs. Educ.* **2007**, *46* (7), 313−318.

(35) Barb, A. S.; Clariana, R. B.; Shyu, C. R. Applications of PathFinder Network Scaling for Improving the Ranking of Satellite Images. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* **2013**, *6* (3), 1092−1099.

(36) Clariana, R. B.; Rysavy, M. D.; Taricani, E. M. Text signals influence team artifacts. *Educational Technology Research and Development* **2015**, *63* (1), 35−52.

(37) Clariana, R. B.; Wallace, P. E.; Godshalk, V. M. Deriving and measuring group knowledge structure from essays: The effects of anaphoric reference. *Educational Technology Research and Development* **2009**, *57* (6), 725−737.

(38) Nunnally, J. *Psychometric Theory*, 2nd ed.; McGraw-Hill: New York,1978.

(39) Thorndike, R. L., Ed. *Educational Measurement*, 2nd ed.; American Council on Education: Washington, DC, 1971.

(40) Brown, T. L.; LeMay, H. E.; Bursten, B. E.; Burdge, J. R. *Chemistry: The Central Science*, 9th ed.; Prentice Hall: Upper Saddle River, NJ, 2003.

(41) Kotz, J. C.; Treichel, P. M.; Townsend, J. R.; Treichel, D. A. *Chemistry and Chemical Reactivity*, 9th ed.; Cengage Learning; Stamford, CT, 2012.

(42) Silberberg, M. S. *Chemistry: The Molecular Nature of Matter and Change*, 2nd ed.; McGraw Hill: New York, 2000.

(43) *Pathfinder 6.3 software*. Software for network analysis , 2011; http://interlinkinc.net/ (accessed Mar 2016).

(44) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*: Erlbaum: Hillsdale, NJ, 1988.

(45) Holme, T. ACS Exams: Past, Present, and Future. Paper from 2006 Fall Conference, ACS Division of Chemical Education Committee on Computers in Chemical Education, Online Conference, 2006.

(46) Pedhazur, E. J. *Multiple Regression in Behavioral Research*, 3rd ed.; Harcourt Brace College Publishers: Fort Worth, TX, 1997.

(47) Brennan, R. L. An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement* **2001**, *38*, 295−317.