2010-07-16

Estimating the Reliability of Concept Map Ratings Using a Scoring Rubric Based on Three Attributes

Laura Jimenez
Brigham Young University - Provo

Estimating the Reliability of Concept Map Ratings

Using a Scoring Rubric Based on

Three Attributes of Propositions

Laura Jimenez Snelson

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R Sudweeks, Chair
Gary M. Booth
Paul F. Merrill
Andrew S. Gibbons
Timothy Morrison

Department of Instructional Psychology and Technology

Brigham Young University

April 2010

ABSTRACT

Estimating the Reliability of Concept Map Ratings

Using a Scoring Rubric Based on

Three Attributes of Propositions

Laura Jimenez Snelson

Department of Instructional Psychology and Technology

Doctor of Philosophy

Concept maps provide a way to assess how well students have developed an organized understanding of how the concepts taught in a unit are interrelated and fit together. However, concept maps are challenging to score because of the idiosyncratic ways in which students organize their knowledge (McClure, Sonak, & Suen, 1999).

The construct a map or C-mapping" task has been shown to capture students' organized understanding. This "C-mapping" task involves giving students a list of concepts and asking them to produce a map showing how these concepts are interrelated. The purpose of this study was twofold: (a) to determine to what extent the use of the restricted C-mapping technique coupled with the threefold scoring rubric produced reliable ratings of students conceptual understanding from two examinations, and (b) to project how the reliability of the mean ratings for individual students would likely vary as a function of the average number of raters and rating occasions from two examinations.

Nearly three-fourths (73%) of the variability in the ratings for one exam and (43 %) of the variability for the other exam were due to dependable differences in the students' understanding detected by the raters. The rater inconsistencies were higher for one exam and somewhat lower for the other exam. The person-to-rater interaction was relatively small for one exam and somewhat higher for the other exam. The rater-by-occasion variance components were zero for both exams. The unexplained variance accounted for 19% on one exam and 14% on the other.
The size of the reliability coefficient of student concept map scores varied across the two examinations. A reliability of .95 and .93 for relative and absolute decision was obtained for one exam. A reliability of .88 and .78. for absolute and relative decision was obtained for the other exam. Increasing the number of raters from one to two on one rating occasion would yield a greater increase in the reliability of the ratings at a lower cost than increasing the number of rating occasions. The same pattern holds for both exams.

Keywords: concept maps, reliability, scoring rubric, rating concept maps propositions, Biology, connected understanding, assessment, psychometrics, concept map ratings.

ACKNOWLEDGMENTS

This dissertation is dedicated to Dr. Richard Sudweeks, professor at Brigham Young University for his exemplary mentoring; to Dr. Robert Patterson, former Dean of the Education Department at Brigham Young University for his financial assistance; and to my son Henry Jimenez for his sacrifice. I am grateful to each of these individuals for constantly encouraging me to climb to a place I had thought impossible to reach. Their support and friendship took an unsure neophyte student and made her a confident and successful doctor.

I want to express my deepest gratitude and appreciation to my dear son Henry who has stuck with me through rough times in the last five years as I've pursued my academic dreams. My gratitude also goes to my parents—papá Santiago and mamá Escolastica—who worked hard to support me emotionally all the way from Peru. I would also like to thank Dr. Andrew Gibbons for opening the doors of opportunity for me by providing financial support to attend school. I'm grateful to Dr. Paul Merrill for selflessly giving his time to teach me the art of writing, which I hope to someday master. To Dr. David Williams for working with me and keeping me focused on the dissertation—thank you.

I also wish to convey my deep appreciation to Dr. Gary Booth, to his graduate student Jessica Rosenvall, and to the four teaching assistants who essentially made my research possible. My thanks go to Jeff Moore, the Biology expert, for his hard work in the creation of the original concept maps for the pilot study. Thank you to Ken Plummer for his guidance and unconditional support throughout the whole process of the research study.

To all the members of my committee—Dr. Richard Sudweeks, Dr. Tim Morrison, Dr. Andrew Gibbons, and Dr. Paul Merrill—who contributed to help me create a presentable final document: thank you.

For their continual support and encouragement in finishing my dissertation, I'd like to say thank you to my friends in the Instructional Psychology and Technology department, especially to Barbara Culatta, Asunta Hardy, Michelle Baron, Tonya Trip, Cary Jonson, and Ken Plummer. Thanks also to our secretary, Michelle Bray, for her kindness and assistance.

My gratitude goes to my friends in the Office of Institutional Assessment and Analysis: to Janae Balibrea for her special friendship and for freely giving her time to help with the writing of this document, and to Eric Jenson, Steve Wygant, Danny Olsen, and Tracy Keck for their friendship and continual support and encouragement in finishing my dissertation.

My heart overflows with appreciation for Mr. Alva Merkley for being my angel friend on earth. Thank you for your time and emotional and financial support during my academic career. Without your support, I would not have been able to make it through college.

To my husband, Terry, for adding meaning to my life, for his kind words, and for his encouragement in getting this work done: thank you so much.

Table of Contents

List of Tables

List of Figures

Introduction

Map making is an ancient practice that can be traced back many millennia but is still important in modern times. As a result of man's continuing exploration and mapping activities, maps of previously undocumented regions of the Earth are now available including the ocean floor, remote islands, inaccessible mountainous regions, polar areas, and the moon. Even though the procedures used to create and reproduce maps have changed greatly, maps still serve many of the same purposes that they did anciently.

Wandersee (1990) called attention to an insightful distinction in terminology. He claims that prior to the time a geographic region had been mapped, it was often referred to as *terra incognita*: an unknown land. However, once a particular area had been mapped, it came to be considered *terra cognita*: a known region. While the primary use of the term *terra incognita* refers to unknown geographic regions, another common use of this term refers to subjects or topics about which nothing is known (Mawson, 1975, p. 335). Hence, Wandersee concluded that the verb *to map* essentially means *to know*. He further asserted that creating a map means "to construct a bounded graphic representation that corresponds to a perceived reality" (p. 323). This generic definition encompasses procedures for graphically representing geographic areas as well as other areas of human thought and inquiry.

Definition of a Concept and a Concept Map

A *concept* is a mental representation of a category of objects, events, processes, roles, relationships, or situations (Murphy, 2002). *Chair* is an example of a concept because it is the mental idea that represents the category to which the word is attached (Sudweeks, 2004). A *concept map* is a graphic representation intended to reveal a students' understanding of how the concepts within a content domain are interrelated.

Concept maps are only one way of representing meaning graphically. Concept maps fit under the general heading of graphic organizers. A graphic organizer is an instructional tool used to illustrate a students' prior knowledge about a topic. Other types of graphic organizers include flow charts, organizational charts semantic networks, and predictability trees. A graphic organizer is a way of visually representing knowledge, structuring information, or arranging important aspects of a concept or topic into a spatial pattern using labels (Barron, 1969). The present study focuses on only concept maps.

Figure 1 provides an example of a concept map. Concept maps are diagrams that consist of four components: (a) nodes in the form of ellipses which contain a written word or phrase that represents a concept, (b) linking lines which represent relationships between related concepts,
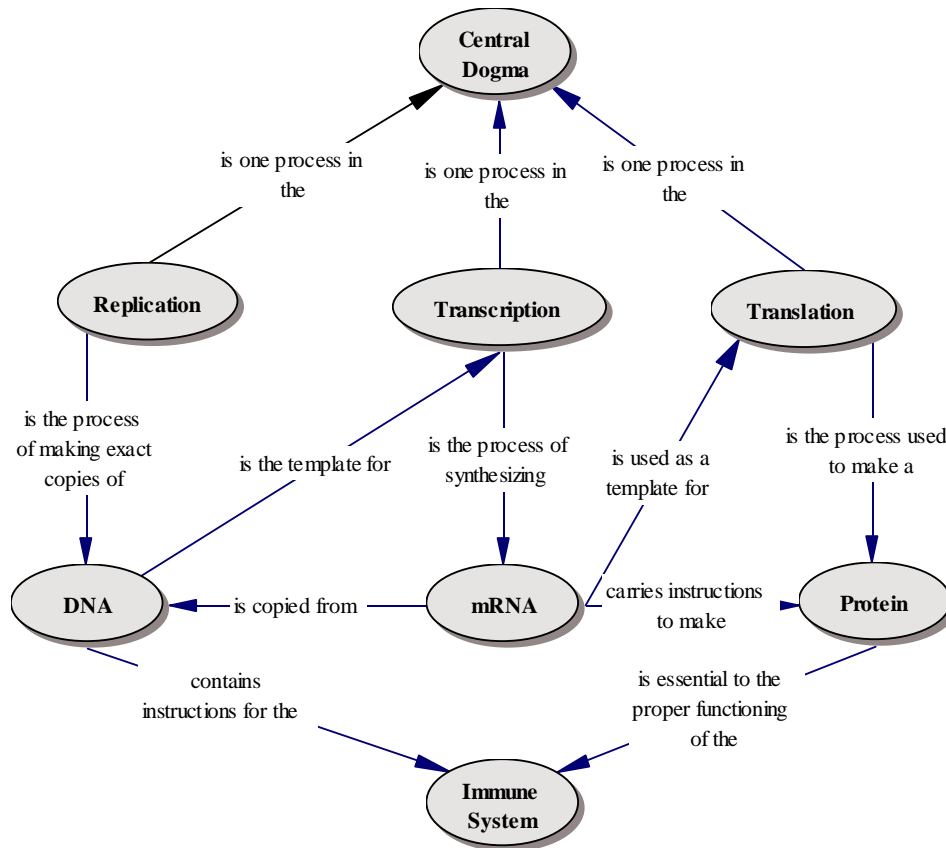
Figure 1. Example of a concept map.

(c) linking phrases that describe the nature of the relationship, and (d) propositions which connect a pair of concepts through a linking line and phrase.

The concept map in Figure 1 contains eight nodes representing concepts. The arrow of each linking line shows the direction of the relationship between concepts. In Figure 1, the linking phrase "is the process of making copies of" describes the relationship between the concepts "Replication" and "DNA." These two concept nodes along with their linking phrase represent the proposition "Replication is the process of making exact copies of DNA." Such propositions are the basic unit of meaning in a concept map.

Instructors have used concept maps to promote understanding by helping individual students to (a) organize their knowledge, (b) make explicit connections between concepts, (c) clarify the meaning of the relationship between various pairs of concepts, and (d) recognize how individual concepts fit together into a larger, interdependent network or conceptual framework.

Concept maps also provide a means of revealing students' conceptual frameworks and making them manifest so that others can observe and assess them in terms of the completeness and the correctness of the concepts listed and the accuracy of the propositions stated. From the 1980s to the 1990s, much research dealt with concept-map assessment tools. Much of that research dealt with challenges related to reliability and, to a much lesser extent, the validity of concept-map ratings as evidence of student's conceptual understanding.

The Use of Concept Maps as an Assessment Device

The use of concept maps for instructional and assessment purposes is consistent with constructivist theories of what learning is and how it occurs. Constructivist views are based on the assumption that the process of learning involves building knowledge structures by connecting new ideas to what one already knows and understands (Osborne & Wittrock, 1983; Palinscar,

1998; Phillips, 1997; Pope & Gilbert, 1983; Smith, diSessa, & Roschelle, 1993). The resulting cognitive structure constructed by a learner is presumed to be somewhat idiosyncratic to that learner.

Many traditional assessment formats such as multiple-choice, alternative response, matching, and short-answer can be reliably scored but often test recall or recognition of facts without regard to how students organize these facts or concepts within a larger conceptual framework (Ruiz-Primo & Shavelson, 1996). Dissatisfaction with traditional forms of assessment has led many educators to seek alternative ways to assess students' learning.

Those assessments that do have potential to measure the degree of student conceptual-framework organization, such as essays or structured interviews (Southerland, Smith, & Cummins, (1998), are generally time consuming to administer and to evaluate. Concept maps provide an alternative to these traditional forms of assessment. Even though concept maps are time consuming to implement, they are more economical, and they are not as labor intensive to rate as extended essay. Studies have indicated that using concept maps is more economical to use than using structured interviews (Southerland, Smith, & Cummins 1998).

The rationale for using concept maps as assessment devices is that (a) they provide a window into the mind of a student and that (b) they are superior to tests that consist solely of selected-response items (e.g., multiple-choice, matching, or true-false items) as a means of assessing students' understanding of how a set of related concepts fit together into an organized, integrated whole. The individual items in a multiple-choice or true-false test are analogous to individual pieces of a jigsaw puzzle. These selected-response items typically focus on assessing isolated bits and pieces of knowledge. Each individual test item may focus on an important part of a greater whole, but until these individual, ideational components are linked together and

connected, the picture presented by a completed puzzle is not apparent. The inherent weakness associated with this approach is that a student may correctly answer all or most of the isolated questions in a selected-response test but fail to grasp how the component concepts fit together into a larger conceptual domain or network of interrelated ideas.

Concept maps provide a direct means of assessing students' holistic understanding of the big picture (i.e., the panoramic perspective of a conceptual domain that shows how the component ideas fit together into an integrated whole). Using a concept map instead of a multiple-choice test to assess students' understanding is analogous to using a camera with a wide-angle lens to take photographs of a landscape instead of using a telephoto lens that captures only a small portion of the scene. The telephoto lens will magnify one small part of the larger landscape, and will reveal details that would otherwise be inconspicuous in the limited conceptual territory that it captures. But it will not show how that small segment fits into the context of the broader conceptual domain. Although both of these lenses provides a useful perspective, they provide very different views. Both perspectives are useful and informative. Which one is most appropriate depends upon the purposes and goals of the teacher or researcher who is doing the assessment. Multiple-choice tests are useful for obtaining a close-up view of how well students understand specific individual concepts in isolation, but they generally fail to provide evidence of students' understanding of how the individual concepts fit together into a comprehensive, unified framework. In contrast, concept maps focus on the broader conceptual landscape consisting of a network of propositions that describe the various ways in which the individual concepts are linked together into an organized framework.

A well-constructed concept map reflects the psychological structure of a student's understanding of a conceptual domain (Wandersee, 1990). "If knowing is making a mental map

of the concepts one has learned and if people think with concepts, then the better one's map the

better one can think" (Wandersee, 1990, p. 926). Although concept maps provide a way to

assess the degree to which students have developed an organized understanding of how the

various concepts taught in a unit are interrelated and fit together into a meaningful whole, they

are challenging to score because of the idiosyncratic ways in which students organize their

knowledge (McClure, Sonak, & Suen, 1999).

Statement of the Problem

In spite of their potential advantages as a mode of assessment, concept maps have a

serious disadvantage that limits their usefulness for assessment purposes. Because of the

idiosyncratic nature of the ways in which individual students construct and organize their

understanding of a conceptual domain coupled with the idiosyncratic manner in which different

students graphically represent their understanding when asked to produce a concept map, the

scoring of the resulting maps is necessarily a rater-mediated process that is subjective and also

expensive in terms of time and effort (Kinchin, 2000).

A *rating* is a judgment made by a human about the quality or quantity of some property

or characteristic of an object or event. In the context of concept maps used for assessment

purposes, ratings are evaluative inferences about the adequacy of a student's understanding

based on evidence that is or is not present in the map produced by a particular student.

Consequently, explicit steps must be taken to standardize the criteria and minimize the

subjectivity of the rating process. Otherwise, the rating that an individual examinee receives may

depend more on who did the rating or when it occurred than on the quality of the examinee's

understanding.

Reliability is a matter of degree. In the context of a rating situation, reliability refers to the degree to which the ratings are free from inconsistencies. Ratings are subject to multiple sources of inconsistencies including: (a) differences between two or more raters who rated the same map on the same rating occasion (i.e., a lack of interrater reliability), and (b) differences in ratings from any given rater who rated the same map on two or more rating occasions (i.e., a lack of intrarater reliability). In addition, to these two common types of rater inconsistencies, other types or rater inconsistencies may involve various kinds of two-way interactions such as rater-by-occasion interaction.

Traditional procedures for estimating reliability are not capable of simultaneously estimating the impact of multiple sources of inconsistencies in ratings. However, Cronbach, Gleser, Nanda, & Rajaratnam, (1972) developed a framework and set of procedures known as Generalizability Theory (G-theory) that provides a way to simultaneously estimate the effects of multiple sources of error variability--including two-way and higher-order interactions--on the reliability of a set of ratings. Generalizability theory is derived from factorial analysis of variance and provides a way of partitioning the total variability in a set of ratings into multiple components each associated with a different source of true or error variance. Generalizability theory is particularly appropriate for use with concept maps because it defines reliability as a variable that takes on different values depending on the magnitude of the variance components associated with each source of error and depending upon the number of raters and rating occasions used. Shavelson and Webb (1993) and Brennan (2001) have done much to popularize the generalizability theory approach to conducting reliability studies.

The research described in this study was an attempt to solve the problems inherent in using concept maps for assessment purposes by (a) using the restricted C-mapping technique to

limit the scope of the domain students were expected to map, (b) specifying a single, central concept intended to serve as the focal point for the map the students produced, (c) providing a scoring rubric and set of guidelines for raters to use to judge the adequacy of students' understanding as manifest in the maps they generated, and (d) using generalizability theory to estimate the reliability of the ratings. The adequacy of the students' maps was defined in terms of three attributes of the propositions that the students' included in their concept maps: (a) importance, (b) accuracy, and (c) completeness.

Statement of Purpose

The purpose of this study was twofold: (a) to determine to what extent the use of the restricted C-mapping technique coupled with the threefold scoring rubric produced reliable ratings of students conceptual understanding across two examinations, and (b) to project how the reliability of the mean ratings for individual students would likely vary as a function of the average number of raters and rating occasions across two examinations.

Research Questions

This study focused on four research questions:

1. What percentage of the variability in the ratings for each examination is due to dependable differences in the students' conceptual understanding?

2. What percent of the variance in each examination is due to the following sources of measurement error?

    a. inconsistencies between raters (lack of inter-rater reliability)

    b. inconsistencies within individual raters across rating occasions (lack of intra-rater reliability)

    c. inconsistencies described by the three 2-way interactions that can be estimated from the two-facet, fully crossed design

    d. unexplained residual error that cannot be attributed to any of the identified sources of variability.

3. What is the reliability of the mean ratings (averaged across four raters and two rating occasions) for making relative and absolute decisions about students' understanding of the subject matter assessed by each of the two examinations?

4. To what extent will the reliability of the mean ratings from each examination likely be increased or decreased by varying the number of raters and/or rating occasions?

Review of Literature

This study aims to investigate the reliability of students' C-mapping scores obtained from an innovative rubric that accounts for three proposition attributes. To help conceptualize this study, we will briefly examine a variety of relevant literature that addresses the history of concept-map assessment, the components of concept-map assessment, and the basic concepts in generalizability theory.

Use of Concept Maps for Assessment Purposes

Since their inception in the early 1970s, concept maps have been mainly used as instructional tools (Novak & Gowin, 1984). In the late 1970s and early 1980s, researchers began to conceptualize ways to use concept maps as assessment tools. Concept maps were found to be superior to traditional assessment items in assisting researchers in their efforts to document students' conceptual change over time (Rowell, 1978). Research completed in the last 15 years provides evidence that concept maps are a defensible measure of students' organized understanding.

Ruiz-Primo, Schultz, Li, & Shavelson, (2001a, 2001b) and Yin et al. (2005) investigated several types of concept-map assessments and found that the most effective concept-map assessments were those that generated scores which evidenced an acceptable degree of reliability. There are several threats to score reliability relative to concept-map assessments. Several researchers have examined the reliability ratings of specific aspects of the concept map assessments including (a) the mapping task or activity (Ruiz-Primo, Schultz, Li, & Shavelson, 2001a, 2001b; Yin et al., 2005); (b) the response format, either paper-and-pencil or computer, (Baker, Niemi, Novak, & Herl, 1991; Fisher, 1990; Liu, 2002); and (c) the scoring system (Nicoll, Francisco, & Nakhleh, 2001; Rice et al. 1998; Rye et al., 2002)

One key issue that reoccurs in the literature is that concept map assessments tend to be very challenging to score. This happens because most students have idiosyncratic ways of organizing their knowledge. The better a concept map reflects an individual's knowledge organization, the more challenging it is to develop scoring criteria that captures the idiosyncratic representation of that organized knowledge (McClure et al., 1999).

Nature of Concept Mapping

Ruiz-Primo & Shavelson (1996) conceptualized a framework which decomposed concept-map assessments into component parts. Their framework characterizes a concept map assessment as including three components: (a) a task that invites students to provide evidence of their knowledge structure in a domain, (b) a format for the students' response, and (c) a scoring system by which students' concept maps can be evaluated accurately and consistently (p. 573).

Ruiz-Primo and Shavelson (1996) claimed that without all three of these components, the use of concept maps could not be considered an assessment. This framework served as a guide for most researchers during this period of time, including key work done by Jacobs-Lawson & Hershey (2001), McClure, Sonak, and Suen (1999), Rice, Ryan, & Samson, (1998), Rye & Rubba (2002), West, Park, Pomeroy, and Sandoval (2002), and Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, (2005). We define each of the three mapping components in the sections that follow.

Mapping tasks. Concept-map tasks are designed to communicate the nature of the task the examinee is expected to perform. According to Ruiz-Primo & Shavelson (1996). A concept-mappins assessment task is comoposed of three variables: (a) A task demand, (b) a task constraint, and (c) a task content structure which refers to the subject domain to be mapped.

The nature of the task demands has implications for many aspects of concept-map assessing activities including feasibility of administration and analysis as well as the reliability and validity of the resulting ratings. Different levels of prompts and directions provided with tasks cause students to draw upon different cognitive processes.

Task constraints are used to specify the task by placing restrictions or limitations on what the student is expected to do in a particular concept-mapping situation. A task that provides linking phrases is more restrictive than a task that directs students to create their own linking phrases. On the surface it may appear that task constraints and task demands are essentially the same. However, task constraints may or may not be impacted by the nature of the task demands. A task that directs students to construct a map from a topic would be less restrictive and more demanding than a task that directs students to construct a map from a list of concepts. Task content structures refer to the intersection of the task demands and constraints with the structure of the subject domain to be mapped.

Response formats. The response format is the second component of a concept-map assessment. This component refers to the format or medium by which a student responds to the concept-mapping task. For example, a student may (a) provide an oral explanation producing a transcription from which a concept map is constructed, (b) draw a map with paper and pencil, or (c) construct a map electronically using concept-map-generating software.

Ruiz-Primo & Shavelson (1996) identified three aspects of a response format from which variations of responses could be derived: (a) the response mode which refers to the medium in which the map is drawn; (b) the characteristics of the response format, which are tied closely to the task demands and constraints imposed by the assessment; and (c) the mapper, who is the person drawing the map. Students generally draw their own map; however, there are instances

when the map is drawn for the student by someone else based on an essay written by the student or a transcript of an interview obtained from the student.

Many possible response formats can be generated by these three elements. These elements include the following examples of concept map response formats (Ruiz-Primo, Schultz, Li, & Shavelson, 2001)

1. Select-the-link: students are asked to select the linking phrase from a provided list.

2. Select-the-node: students are asked to select a concept from a provided list.

3. Select-the-link and node: students are asked to select the link and node from a provided list.

4. Fill-in-the-links: students are asked to fill in a skeleton map with a description of the relationship of each pair of connected concepts.

5. Fill-in-the-node: students are asked to fill in a skeleton map with a concept that would complete the description of the relationship of each pair of connected concepts.

6. Fill-in-the-link and node: students are asked to fill the link and node in a skeleton map with a concept and a linking phrase that would state a proposition.

7. Construct-a-map by assembling concepts and linking phrases: students are asked to construct a map from a provided list of concepts and linking phrases.

8. Construct-a-map from the list of concepts provided: students are asked to construct a concept map from a provided list of concepts.

9. Construct-a-map from scratch: students are given a blank piece of paper and a main topic and asked to connect all of the key concepts subsumed under that topic.

10.   Construct-a-hierarchical-map: students are asked to construct a map taking into

consideration the hierarchy of the concepts.

What follows are two examples of concept-map response formats. Figure 2 is an

example of the response format characteristic (number 6 in the list above) of the fill-in-the link

and node concept-map assessment. Figure 3 is an example of the task instructions with its

corresponding format characteristic (concept map format number 8 in the list above) of the

construct-a-map from the list of concepts provided.

This C-mapping response format assessment has been called the gold standard of

concept-map assessments (Ruiz-Primo & Shavelson, 1996) because it has been shown to

provide valid information regarding how students organize their knowledge and shows high

reliability coefficients (Plummer, 2008; Ruiz-Primo, Schultz, Li, & Shavelson, 2001,). Since the

C-mapping task does not supply a diagram for the students, it has been considered problematic

for large-scale assessment because there can be such variability in the diagrams that students

produce. Students, on the one hand, need to be trained to use maps effectively before they can

create them, and the variability in the maps of untrained students makes scoring difficult and

time-consuming (e.g., Schau & Mattern, 1997). Ruiz-Primo & Shavelson (1996) has tried to

overcome these two problems by designing a 50-minute instructional program to teach students

how to construct concept maps.

The program has proved to be effective, producing the intended outcome when used with

more than 100 high school students. Map propositions were scored for accuracy and

comprehensiveness.

Concept map assessment research has shown that the more freedom students have to

visually depict their understanding of how concepts interrelate, the more likely their scores will
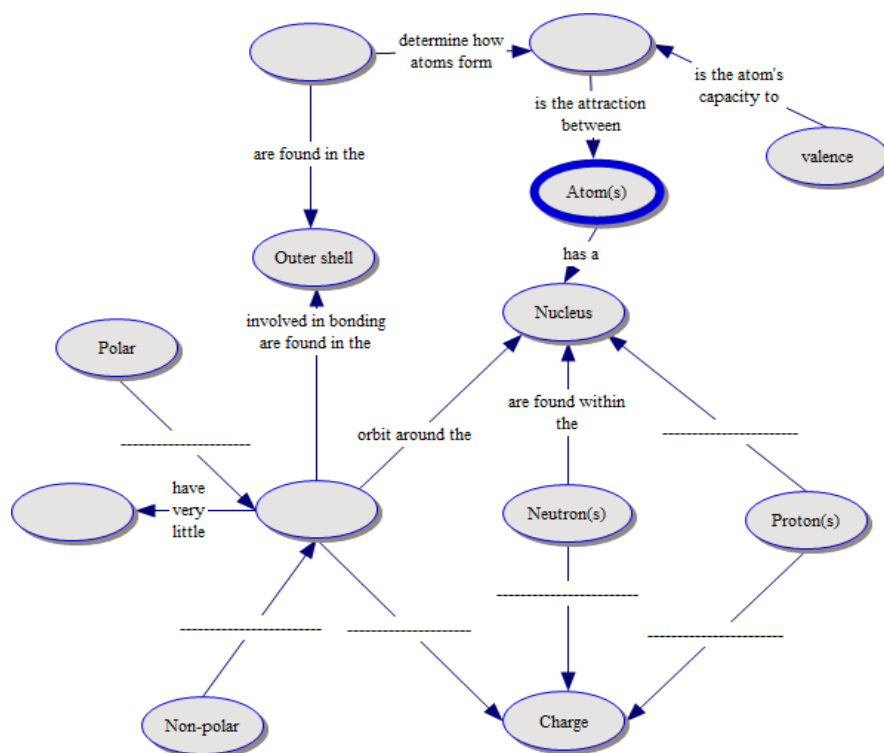
Figure 2. Fill-in-the-link and node concept-map response format assessment.

Instructions: Construct a concept map showing how ideas listed below are interrelated. Include examples of various concepts where appropriate.
1. Central Dogma
2. DNA
3. Immune System
4. mRNA
5. Protein
6. Replication
7. Transcription
8. Translation

Figure 3. Example of the C-mapping technique.

correlate with other valid measures of their connected understanding, such as essays (Schau et al., 1997). Yin et al. (2005) said that "The C-mapping technique, more accurately reflected differences of students' knowledge structures; provided greater latitude for demonstrating students' partial understanding and misconceptions; supplied students with more opportunities to reveal their conceptual understanding; and elicited more high-order cognitive processes, such as explaining and planning (p. 167)."

McClure et al., (1999) described the C-mapping task as the technique of choice for considering scoring method, and it has been favorably compared with other measures of connected understanding (Francisco, Nakhjleh, Nurrenbern, & Miller, 2002; Rice, Ryan, & Samson, 1998).

Scoring systems. The third and final component of a concept-map assessment is a scoring system by which student concept maps can be evaluated. Ruiz-Primo & Shavelson (1996) explain that a scoring system is a "systematic method with which students' concept maps can be evaluated accurately and consistently" (p. 581). A more detailed description of how Ruiz Primo & Shavelson categorize these scoring strategies is provided in the review of literature section.

Scoring Systems Used

Ruiz-Primo and Shavelson (1996) explain that a scoring system is a "systematic method with which students' concept maps can be evaluated accurately and consistently" (p. 581). They categorize scoring systems into three general strategies: (a) scoring the components of a map, (b) comparing the student's map with a criterion or master map, and (c) using a combination of these first two strategies. (d) using the holistic scoring method studied by McClure et al. (1999).

Scoring map components. The components of a concept map that have been scored include the (a) concepts, (b) propositions, (c) examples, and (d) map structure.

Concepts. Scoring the concepts of a map should occur only if the students rather than the assessor are directed to supply map concepts. For example, in the construct-a-map from scratch task, students are given a topic and asked to construct a map that depicts the key concepts and propositions of that topic. In this case students are directed to supply relevant concepts to the topic, and then their concept selection is scored accordingly. Because of the heavy cognitive load imposed upon students with such a task, in most cases concept selection is already done for the student and hence, not scored (Schau et al., 1997). Another task that would require the scoring of student-concept selection would be requiring the students to add relevant concepts to a list of assessor-selected concepts (e.g., Rice et al., 1998). This feature adds a level of complexity to the scoring method in that students may add an innumerable number of concepts with their resulting propositions, each of which must be accounted for by the raters.

Propositions. When scoring propositions, two strategies are generally considered: the scoring of individual propositions and the calculation of total map proposition scores. Individual propositions have generally been evaluated based on their level of correctness. In some instances the propositions are scored simply as correct or incorrect (Yin et al., 2005) and in other instances the propositions are rated based on degrees of correctness (Ruiz-Primo et al., 2001). Some scoring methods take into account proposition choice (McClure et al., 1999), and others consider the direction of the linking phrase arrow (Anderson & Huang, 1989).

At the very least, a proposition is scored based on how correct or accurate it is. A correct proposition simply communicates an idea that is accepted as valid by domain or content experts in a given context. Proposition choice is another proposition-scoring attribute that has been

included by some researchers in the concept-mapping assessment literature (e.g., Yin et al.,
2005; Rice et al., 1998). Ruiz-Primo & Shavelson (1996) suggest that when a student selects a
pair of concepts to be mapped, he connect concepts that vary in degrees in the strength of their
association. Propositions may be weighted based on their associated strength as well as their
relevance to the overall topic. Correctly choosing pairs of concepts to form key propositions is
essentially a function of the context or topic of the domain and the level of domain expertise
possessed by the mapper.

Another scoring attribute or property of individual propositions is proposition
completeness (Ruiz-Primo et al., 2004)**,** which describes the degree to which the information in
the proposition demonstrates a complete understanding of the relationship between two concepts.
A proposition can be accurate and vary in its degree of completeness. For example, the
proposition *reliability must be present in order to claim validity* is an accurate proposition;
however, a more complete expression of their relationship would be *reliability is necessary but
not sufficient in order to infer validity*. Notice that the first proposition essentially communicates
that reliability is a requisite of validity. The second proposition adds the idea that reliability is
requisite but not the only requirement to make a claim of validity.

Researchers such as Nicoll et al., (2001) have considered other scoring properties. In
their studies, they derived concept maps from student interview transcripts. They rated each link
based on the following:

1. Proposition utility, which is the degree to which a proposition is considered correct.

2. Proposition stability, which is the degree to which a student expresses a proposition with
   certainty.

3. Proposition complexity, which is the degree to which a proposition is useful in predicting or explaining a scientific phenomena or enhancing the understanding of other connections on the map.

Along with individual proposition scores, total proposition scoring schemes have also been conceptualized by researchers. Ruiz-Primo et al. (2004) describes three total proposition scores (a) total proposition accuracy—the sum of all the scored propositions in a map, (b) convergence score—the percentage or proportion of scores on a student map found on an expert or criterion map, and (c) salience score—the percentage or proportion of accurate propositions out of the total number of propositions in the student's map.

Summing the scores of all propositions in a map is a simple procedure, yet there are several issues to consider when doing so. As an extreme case, if every concept could be meaningfully connected to every other concept on a map, then the number of the total propositions that could be connected can be calculated using the formula *N(N-1)/2*, where N equals the number of concepts in the list. If the number of concepts were 10, then the number of total possible propositions would be 45. If a student constructs 45 propositions and each proposition can be scored on a scale of 0 to 2, then the highest total proposition accuracy score would be 90.

Of course, it is inconceivable that an instance could occur where every concept could be meaningfully linked with every other concept on a map. Additionally, most concepts within a given subject or discipline differ in the degree to which they meaningfully relate to one another. This scoring attribute, as described earlier, is called *proposition choice* or *proposition importance* and directs raters to give credit to students who connect those concepts that should be connected and no credit for those concepts that should not be connected. The challenge here

is for content experts to develop a list of strongly associated propositions that students should make along with a list of moderately and weakly associated propositions. Such an effort can be daunting because any one discipline may possess an exhaustive list of closely and moderately related concepts. If this can be accomplished, however, then a total possible proposition score is a more viable approach to evaluate student maps (Yin et al., 2005).

Some researchers have not pursued a total proposition score because of this challenge and have looked to other scoring approaches that reflect student concept-mapping performance such as convergence and salience scores.

Convergence scores are calculated by comparing the number of propositions shared by a student map and an expert map. This score is generally calculated as a percentage or a proportion. If a student constructs 90% of the propositions found on an expert map, then she would receive a .90

In the case of salience scores, this scoring technique is calculated in several ways (see Francisco et al., 2002; Ruiz-Primo et al., 1997). The most basic calculation is done by dividing the number of correct propositions by the total number of propositions on the map. If a student constructs ten propositions and five are correct, then his score would be .5 or 50% correct. A challenge with salience score calculations is that a student could conceivably score a 1.0 by constructing only one or two accurate propositions. Hence, a score of 1.0 may or may not represent a student who possesses a well-developed connected understanding of the material.

Ruiz-Primo et al. (1997) compared the results of concept-mapping scores calculated using all three methods and found total proposition accuracy and convergence scores to be more consistent than salience scores. They found that student differences were more pronounced when using total proposition accuracy and convergence scores than salience scores.

Examples. Citing examples provides evidence of a student's ability to instantiate abstract concepts. For example, it may be known that a fifteen-year-old boy knows that a Llama is an animal, but if he links Llama to the instance *K'ara Llama* with the linking phrase *is an instance of,* it would also be known that he could identify an instance of the concept Llama. When scoring these types of propositions, Novak & Gowin (1984) weighted propositions with examples and other propositions equally. The limitation here is that since concept maps can showcase a students' understanding of the essential relationships between key conceptual pairs in a given domain, it may be of less interest to depict an example of any one concept. Hence, examples of certain concepts may not evidence propositional or structural understanding but evidence more an understanding of an instance of a particular concept. If this is an outcome of interest to the assessor, then students should be directed to add examples where applicable in their maps.

Map structure. A hierarchical structure includes any structural pattern that transcends simple propositional relationships. Map structure can include subordinate /superordinate relationships between concepts as well as coordinate (coequal) relationships. Subordinate/superordinate relationships may be depicted with an all-inclusive superordinate concept placed at the top of the page and increasingly less inclusive subordinate concepts subsumed below it. For example, the concept *polygon* is a superordinate concept subsuming concepts such as *quadrilateral* and *triangle*. The concept quadrilateral in turn subsumes the concepts *rhombus* and *parallelogram* while the concept triangle subsumes the concepts *scalene* and *obtuse*.

Novak & Gowin (1984) designed a scoring formula that accounts for map structure by counting and weighting valid levels of hierarchy as well as cross-links connecting different

clusters of strongly associated concepts. They assumed that expressing hierarchical levels in a given domain provides evidence of student ability to differentiate concepts based on developed nuanced understanding of how they fit into a larger conceptual framework.

However, Ruiz-Primo et al. (1997) explain that few domains are purely hierarchical and that most manifest more or less of what they term *hierarchiness*. It appears that most content domains feature some hierarchical structure; however, hierarchical relationships do not generally account for the vast number of propositional relationships (Cohen, 1983). In other words, an assertion can be made that all domains have some hierarchical skeletal structure, but hierarchical structure generally accounts for a much smaller percentage of the total propositions that could be constructed from those domains. Scoring map structure is important if (a) there is a strong presence of hierarchical relationships in the content domain and (b) it is the explicit objective of a course to assist students in understanding the hierarchical nature of the content.

If, however, the spatial features of the map do not account for a conceptual framework, the individual propositions are the only map components left to score. This gives rise to the question, can the content structure of a domain be accounted for by analyzing solely the linking phrases expressed within each proposition of the map without its spatial features? Anderson (1995) makes the following point answering this question in the affirmative, paraphrased below.

The spatial location of elements in a network is totally irrelevant to the interpretation. A network can be thought of as a tangle of marbles connected by strings. The marbles represent the nodes, and the strings represent the links between the nodes. The network represented on a two-dimensional page is that tangle of marbles laid out in a certain way. We try to lay the network out in a way that facilitates its understandability, but any layout is possible. All that matters is what elements are connected to which, not where the components lie (p. 148).

One way to capture student knowledge structure without considering the spatial layout of the map is to consider two propositional attributes: proposition choice/importance and proposition completeness. If students are to pair concepts that have hierarchical relationships, then this would be a criterion for appropriate proposition choice. If the essential relationship between two concepts is hierarchical in nature, then students would be expected to express a hierarchical relationship in the linking phrase in order for the proposition to be considered complete.

This issue has important implications for scoring concept maps. While a few researchers continue to study the possibility of scoring map structure (e.g., Yin et al., 2005), more theoretical and empirical work needs to be done considering the viability of accounting for structures using methods that are reliable and valid.

Comparing students' maps with a master map. Another scoring option that has gained wide acceptance is to compare a student map with an expert, criterion, or master map. The criterion map functions as a standard to evaluate (a) the acceptableness of concept selection, (b) proposition choice, (c) proposition accuracy, (d) map structure, etc. Criterion maps are difficult to construct because of challenges highlighted in the study by Acton, Johnson, & Goldsmith (1994). In their study, criterion maps were constructed by field experts and a class instructor. They found that individual experts were highly variable in the specifics and, in some instances, the generalities of their maps. The course instructor, however, showed even greater map variability from the experts. To add to the complexness, the student maps correlated much less with the instructor map than with the expert maps. This finding has serious implications for the viability of comparing students' maps with a master map.

Combining strategies. The third strategy proposed by Ruiz-Primo & Shavelson (1996) is to score concept maps using a combination of strategies—scoring the components of a map while using the criterion map as a guide. McClure et al. (1999) investigated six scoring methods that focused on different aspects of student maps including a holistic, structural, and relational evaluation. The relational scoring method (scoring each proposition separately) guided by a criterion map proved to demonstrate the highest reliability ratings of the other five methods. Hence, a combination of strategies or a triangulated method may provide greater reliability of concept-map assessments.

Using a holistic scoring method. While not as common, the holistic method has been studied in a few investigations. As mentioned previously, McClure et al. (1999) studied the inter-rater reliability of raters rating concept maps with different scoring methods. One of those methods was the holistic scoring method where raters examined student concept maps and judged the mapper's overall connected understanding from the map on a scale of 1 to 10. This particular method was found to generate inconsistent ratings. The researchers reported that this might have in part been due to how cognitively taxing it is to account for map quality without a specific guide for scoring the detailed components of the map.

This research study built on the concept map research conducted by Plummer (2008) who recommends that four distinct scales can be developed measuring importance, accuracy, completeness, and relevance for each proposition on the map. We decided to develop a scoring rubric that accounted for three attributes of a proposition leaving out relevance. Plummer asserted that a scoring method developed in this vein is less cognitively loaded for raters and assists them in accounting for all three rating elements.

Reliability of Map Ratings

Since the primary focus of this study deals with estimating the reliability of the C-mapping technique, we include an introduction to G-theory. G-theory is a measurement theory that explicitly acknowledges the existence of different sources of measurement error and provides a way to simultaneously estimate the magnitude of these multiple sources of error that may affect the dependent variable

In writing this section, we assume that readers are familiar with classical reliability theory. We assume that they understand such concepts as true score, error, and reliability. We also assume that readers have some basic understanding of Analysis of Variance (ANOVA) and specifically how ANOVA partitions variability. Minimally, we provide basic definitions for those readers who may be unfamiliar with concepts related to G-theory. Our goal in the present section is to help readers understand the level of applicability of G-theory to the present study.

In G-theory, a behavioral measurement is considered to be a sample from a universe of all possible observations. This universe of possible observations may include one or more facets. The term "*facets*" is analogous to "*factors*" in the literature on experimental design and factorial analysis of variance. In short, a facet is a potential source of error. A universe of observations is said to consist of one facet if the generalizability of observations regarding one source of variation in the universe, say arithmetic questions of varying difficulty, is at stake. For instance, a particular arithmetic test includes a sample of items covering different addition, subtraction, multiplication, and division problems of one- and two-digit numbers. The decision maker is interested in general arithmetic achievement, and is indifferent to the particular questions on the test. In the one-facet design, one is interested in estimating the universe score of each person based on the sample of items included in the test.

A universe is said to have two (or more) facets if the generalizability of observations regarding two (or more) sources of variation in the universe—say items, raters, and occasions—is at stake. Reliability can be increased by increasing the number of items, raters, and occasions. However, there is a trade-off between increases in reliability and cost.

G-study. The first stage of a generalizability study is called a G-study. The purpose of this phase is to obtain estimates of the variance components. Based on the use of G-theory, it is possible to determine which of these facets contributes the most measurement error. This is done by partitioning the total variance into separate, additive components including the universe variance and the variance due to each facet and each possible interaction. According to Cronbach et al. (1972), the conceptual framework underlying G-theory is that "an investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs that is, he generalizes from sample to universe" (p.15). The question of reliability thus resolves into a question of accuracy of generalization.

The concept of universe score can be considered the heart of the G-theory. For any particular measurement it is possible to conceive of many different universes to which one may want to generalize. It is therefore essential that the universe the investigator wishes to generalize be defined by specifying which facets are likely to change without making the observation unreliable. Ideally, we would like to know an examinee's score (universe score, over all combinations of facets (e.g., all possible tasks, task forms, or all possible occasions). Unfortunately, the choice of a particular task, task format, or rating occasion will inevitably introduce error into the measurement procedure because the universe score can only be estimated. In most situations, persons are the objects of measurement. Variability among

persons is treated as true variance, whereas all other facets are treated as potential sources of error although G-theory, via the principle of symmetry, does permit the possibility that some other facets could be regarded as the objects of measurement, in which case persons would be treated as measurement error (Marcoulides, 1989).

Main effects. It is important to note that G-theory is analogous to random effects analysis of variance in that both are used to estimate the variance components associated with the main effects and interaction effects through the analysis of mean ratings. Instead of computing F-ratios to test hypotheses, the ANOVA in a G-study produces an estimate of the variance component for each main effect and each interaction. G-theory thus goes beyond ANOVA, in that it can be used to estimate the relative percentage of measurement error from each of these facets. In the section that follows, an explanation of each main effect variance component and each interaction effect will be explained.

Variance components for persons. Ideally, the variance component for persons (students in this study) should be larger than any of the other variance components. Students are the object of measurement and thus constitute the population of interest from which the researcher wishes to make inferences. The purpose of assessing the students' connected understanding of related concepts presupposes that the amount of this trait varies from student to student. It is one of the express goals of this study to investigate how sensitive C-mapping scores are for different students who manifest varying degrees of this trait. If the variance components for a person are high but there is no variance across rater and occasion facets, the score for each person varies, but each person is given the same score by each rater in each occasion. In other words, the person means vary, while the rater means and the occasion means are constant.

Variance components for raters. If all the raters in the universe of admissible raters were equally stringent in the way they rate concept maps, then the average rating for each rater would be the same and the variance component for raters would be zero. This variance component may be large for several reasons, including rater fatigue, disparate rater knowledge of the subject matter, a consistent tendency of some raters to be lenient or stringent in their ratings, and other sources of rater error. This happens when the mean ratings for each person vary from rater to rater, while person mean rating variability, and occasion mean rating variability remain constant across all raters. The mean ratings for raters are computed by averaging the scores for each rater across both occasion and persons.

Variance components for rating occasions. If the average ratings are unchanged on all occasions in the universe of admissible rating occasions, then the average rating on each occasion would be the same and the variance component for occasion would be zero. This would indicate that on the whole, each rater was consistent with themselves from one rating occasion to another. The error related to occasion effect may be a result of all raters collectively or individually making changes in the way they rate or of some outside experience that causes them to rate differently from one occasion to another.

 So far we talked about the analysis of the three main effects. Now we will talk about the analysis on the three 2-way interaction effect between persons, raters, and occasions in an effort to estimate the amount of measurement error from these sources.

Variance components for interaction effects. Two way and higher order interactions can both be estimated in generalizability studies.

Person-by-rater. If all of the ratings for persons were ordered by rank and it was found that each rater ranked persons differently, then the person-by-rater variance component will be

high, and this constitutes a source of measurement error. Generalizability theory is sensitive to

this rank ordering difference between raters. This source of measurement error may occur when

raters are not uniform in their understanding of the rating criteria, or if raters are partial to

unrelated aspects of the rating process, such as how neat and well-organized the students'

concept maps may appear.

Person-by-occasion. If a group of raters all ranked persons similarly on one occasion and

then ranked them as a group differently on another occasion, then the person-by-occasion

variance component would have a relatively large value.

Rater-by-occasion. In the case of the rater-by-occasion variance component, each rater

ranks all persons the same, or in other words uses exactly the same scoring methods. If on the

second occasion, each rater rates each person much higher or lower than they did on the first

occasion, then the rater-by-occasion variance component would be relatively large.

The residual variance. Ideally, the unexplained residual variance should be small relative

to the other variance components. These sources of variability is a composite of the two-way

interaction between the person-by-rater/person-by-occasion variance components and any other

random and unidentified events error, all these sources of variability cannot be disentangled.

This error cannot be explained because it represents variability beyond the scope of the analysis.

Mean ratings can only be computed on prespecified parameters and facets. In the case of this

study, any other source of measurement error such as ill-defined aspects of the rubric or

complete randomness in the way raters rate on each occasion would represent error that cannot

be detected with the prespecified facets of the study.

D-Study. After the estimates of the variance components are obtained, the estimated

values are then further analyzed in the second and final phase called a D-study. The purpose of

the D-study (Decision Study) is to make informed decisions about how many levels of each facet (mapping tasks, raters, and rating occasions) should be used to obtain acceptable reliability at a feasible cost (Shavelson & Webb, 1991).

In a D-study, the researcher must (a) define a *universe of generalization* (the number and breadth of facets to be generalized across, such as rater, occasion, task, scoring scheme, etc.); (b) specify the proposed interpretation of the measurement: *relative decisions to* rank order individuals standing relative to all others or *absolute decisions—*an individual's absolute score without regard to other student scores; and (c) use variance components estimates from the G-study to estimate the relative and absolute error variance and the generalizability coefficients for relative and absolute decisions. The variance components that contribute to measurement error are different for relative and absolute decisions. For relative decisions, all variance components that influence the relative standing of individuals contribute to error. These components are the interactions of each facet with the object of measurement, in this case, students. For absolute decisions, all variance components except the object of measurement contribute to measurement error. These components include all interactions and the facet main effects.

Method

G-theory was used to determine what percent of the variability in concept map ratings was due to dependable differences in the students' understanding and what percent was attributable to various sources of error. G-theory was also used to assess students' connected understanding by determining the reliability of students' concept map scores across different examinations. Four raters were used to rate concept maps using an innovative rubric that accounted for three attributes of concept map propositions. The sections that follow provide a description of the methodology using participants, the procedures of the study, the student training, the student assignments, the instrumentation, the study design, and the data collection and analysis.

Participants

A total of 120 freshman college students enrolled in a Biology 100 class at Brigham Young University participated in this study. The gender distribution was 60% males and 40% females. Approximately 90% of students were white and 10% were minority students.

Instrumentation

Mapping tasks. As mentioned earlier, this study was conducted within the context of a university beginning biology course. The course curriculum included four exams. For the purposes of this study a C-mapping task was included in each of the second and third course exams. Each C-mapping task included instructions on how to proceed with the construction of the map. The instructions for the two exams included a note indicating that the map should be focused around the central concepts "Cell" and "Evolution," respectively. The words *Cell* and *Evolution* were bold and underlined and placed at the top of the list. The test instrument also had two other notes that suggested that students include all the concepts listed in the test item. Please

refer to Appendix B for a full version of the test instruments for Exam 2 and Exam 3. Additionally, another text-note was given to students to let them know that their map should include at least 24 propositions. Then a list of 20 concepts for Exam 2 and 16 concepts for Exam 3 was given. The purpose for underlining the central concept was so that students would keep in mind when constructing the map that they should start around the underlined central concept.

At the bottom of the page, a description of the scoring rubric was provided in a text box. This description had the purpose of helping students know how their maps would be rated. A second box explained the number of points that each quality of proposition would receive based on the scoring rubric guidelines. Furthermore, a definition of the word *proposition* was provided as well.

Rater judgments. Four experienced teaching assistants of the Biology 100 class were used as raters to rate all the student concept maps using the innovative scoring rubric and the master maps corresponding to Exam 2 and Exam 3. Please refer to Appendix B to see the master maps.

Scoring rubric. This study used an enhanced rubric to account for three attributes of a concept map proposition. The rubric to score the maps was built based on the recommendations of Plummer (2007) that emphasized proposition accuracy. The ratings resulting of using the scoring rubric was the dependent measure used in this study.

The detailed description of the scoring rubric was included in the students' training packet. The scoring rubric description provided instructions on the purpose of concept mapping as well as a description of how to concept map and an example of how to write an appropriate linking phrase. The student training packet also included examples of the fill-in-the-blank nodes

and blank links concept map task as well as examples of the C-mapping technique. For a full version of the scoring rubric and student training packet, please refer to Appendix D.

Procedures

Prior to the beginning of the semester, a total of four teaching assistants prepared the concept map assignments and their corresponding master maps for each Biology 100 lecture to be taught during the coming semester. Each teaching assistant created a concept map for each lecture based on 10 to 12 concepts that the instructor of the class had predetermined to be of importance using the Inspiration software, a computer visual mapping tool. At this stage, the instructor also determined the central concept to which all the other concepts included in the map should be related. All four teaching assistants' brought their concept maps to a group work session, and then, based on the best contribution among the four maps, the examination map and a master answer key were created. This process was done with the participation of all teaching assistants and the researcher. This final map was reviewed and approved by the instructor of the class.

At the beginning of the study student participation was classified into two categories: activities and exercises that occurred as part of the course curriculum and additional activities that were introduced as part of this study. The activities that were already a part of the curriculum included concept map training, lab quizzes, homework assignments, and a total of four examinations. In the present study, a C-mapping test item was included in both the second and third course exams, here after referred to as Exam 2 and Exam3. Each student's concept map for Exam 2 was rated by four raters on two different rating occasions. The elapsed time between the two rating occasions was one week. The same raters were employed on each occasion. Five weeks later, this same process was repeated for Exam 3. That is, each student's

map for Exam 3 was rated on two different rating occasions by four different raters with one week elapsing between the first and second rating occasion.

All student participants were invited to sign a document of informed consent expressing their willingness to participate in this study for the additional events that were not part of the regular course (see Appendix A). Activities such as concept map assignments and lab training were not included in the informed consent because they were already part of the curriculum. In the section that follows, we proceed to explain the procedures we followed for rater training, student training, and assignments.

Training in concept mapping. During the first class session of the semester, students received a training packet. (A complete version of the concept-map training packet can be found in Appendix D.) The training packet contained information about the meaning of a concept, information about the purpose of concept mapping, and information about the types of concept maps they would be exposed to during the semester. The introduction was followed by training on the construction of concept maps. The researcher took 15-20 minutes of a one-hour lab session to train students on the construction of concept maps. During this session, the researcher explained to students that the objectives of the course included the goal that each student learn and understand biological facts, biological concepts, and how those concepts interrelate with one another. It was further explained that several methods could be employed to assist them in developing a useful, organized understanding of the content and that concept mapping was the tool of choice in their Biology 100 course.

Assignments. As part of the students' training, they were also given an assignment that consisted in filling the blank nodes and blank links of a skeletal concept map. This assignment had to be completed within a week. All four assignments prior to the first midterm exam were

the fill-in-the-blank-links or blank-nodes of a skeletal concept map. The purpose of giving this type of concept map as an assignment was to facilitate students in making a smooth transition in the process of learning the C- mapping skill.

Starting with Lecture 6, in addition to fill-in-the-blank-links and nodes concept map, Task I introduced the second type of concept mapping task known as the C-mapping task. Students received both types of concept mapping for the remainder of the semester. The weekly assignments corresponded with the weekly lectures.

Upon students' completion of concept map assignments, the maps were graded by their respective teaching assistants. All skeletal maps and C-maps assignments were rated using the innovative scoring rubric that looked at three attributes of a concept map proposition. Once assignments were graded, they were returned to the students, and students were given the opportunity to request a meeting with their teaching assistant to receive feedback on possible misconceptions. Students who met with their teaching assistant and corrected their concept maps were given full credit for the assignment. This process continued throughout the course of the semester. The feedback and reward process was implemented because of the positive effect that this intervention had in students towards the end of the course in prior research studies we conducted using concept maps. This feedback procedure did not have the impact we expected in this study. Very few students took the time to request a meeting with their rater to get full credit. The reasons for this are unknown and beyond the scope of this study.

Plummer (2008) used a list of about thirty concepts when giving students the assignment to construct the C-mapping task. In this study we decided to significantly reduce the number of concepts provided to a maximum of 20 in the second examination and 16 in the third examination. By reducing the number of concepts and by predetermining a central concept, we

assumed that the rating process would go faster and raters would not experience as much fatigue.

Assignment scores were not included in the data collection/analysis of the present study.

Exams. There were a total of four exams in the course of the semester. Three out of four

exams consisted of a 50 multiple-choice questions plus a concept mapping task. The fourth

exam did not include the concept map question. The first examination took place approximately

four weeks after the semester started. The remaining three exams were schedule at different

times with four weeks gap from exam to exam during the semester. Students took their test any

day during the week that an exam was scheduled. All examinations took place at the university's

testing center. At the end of the each examination period, all multiple-choice questions were

machine scored. Completed concept maps were returned to the course teaching assistants for

their corresponding rating. The same concept map rating procedure was used for the first three

exams. Data for this study was collected from Exam 2 and Exam 3.

Design

A total of 115 students in the class constructed a concept map as part of Exam 2 and a

another map as part of Exam 3. The maps produced by each participating student in response to

Exam 2 were rated by all four raters on the first rating occasion. These ratings were used for two

purposes: (a) to provide the course instructor with ratings of the students' understanding of the

concepts taught in the associated part of the course, and (b) to provide data to be analyzed in this

study. In order to reduce the time demands on the raters and the costs of conducting the study,

systematic random sampling was then used to select a random half of the maps to receive a

second rating.   Fifty-seven of the Exam 2 maps were rated on the second rating occasion. This

same procedure was used in rating the concept maps collected as part of Exam 3. Again 57 of

the Exam 3 maps were chosen to be rated a second time, but they did not necessarily represent the same 57 students whose maps were selected for a second rating in conjunction with Exam 2.

The design of this study was a two-facet, fully crossed P x R x O design where P designates Persons (the object of measurement), R designates Raters, and O represents rating occasion. This same P x R x O design was subsequently used to rate the concept maps generated by the students as part of Exam 3.

Data Collection and Analysis

Estimates of the variance components for the two exams were computed using the GENOVA software (Crick & Brennan, 1982, 1983). A D-study for each exam was conducted to predict how varying the number of facets used in the study would affect the size of the variance components and the reliability estimates.

Results

Estimated Variance Components

Research question 1 focuses in the percent of variability in the ratings for each examination. Table 1 reports the results of the G-studies for both Exam 2 and Exam 3. The table reports an estimated variance component for each of the seven sources of variability that can be estimated in a fully crossed P x R x O design.

Table 1

Estimated Variance Components by Exam and Source of Variation

Sources of Variation	Degrees of Freedom	Exam 2			Exam 3		
		Estimated Variance Component	Percent of Total Variation	Standard Error	Estimated Variance Component	Percent of Total Variation	Standard Error
Persons (P)	56	10.5208	73%	2.0577	4.9748	43%	1.0484
Occasion(O)	1	0.0000	0%	0.0069	0.0151	0%	0.0214
Raters (R)	3	0.7267	5%	0.4853	2.9211	25%	1.8763
PO	56	0.0506	0%	0.1160	0.0000	0%	0.1204
PR	168	1.1037	8%	0.2529	1.4648	13%	0.3072
OR	3	0.0068	0%	0.0268	0.0000	0%	0.0254
POR,e	168	2.0041	14%	0.2174	2.2626	19%	0.2454

The entries in the Standard Error column provide an index of how precisely each of the corresponding variance components was estimated. The variance components for persons for each of the two exams is reported in the first line of Table 1. This variance component provides an estimate of the degree to which the mean ratings for the different students vary about the grand mean of all the students in the population as shown in Figure 4. The relative size of this variance component is indicative of the degree to which the raters were able to make dependable distinctions in their ratings of the student's conceptual understanding. Since the variance

component for persons represents desirable variance, ideally it should be large relative to the other variance components reported in Table 1.

Generalizability theory provides a way of partitioning the total variance in the ratings from each exam into component parts. The percentages reported inTable 1 were computed by applying the heuristic suggested by Shavelson and Webb (1991). The sum of the variance component estimates for each comprehension measure was computed first. Then each variance component estimate was divided by this total, and the quotient was multiplied by 100%. In the context of this study, the variance due to students is considered to be a universal score variance, which is analogous to true score variance in classical test theory.

The percentage of variability associated with each variance component reflects its relative magnitude. The variance component for students is described as universe score variance in G theory and is analogous to true score variance in classical test theory. In other words, the variance component for students summarizes the degree to which the variability in the ratings is indicative of dependable differences in the students' understanding of the conceptual domain being assessed. The other variance components in Table 1 are indicative of the degree to which the overall ratings are influenced by the different sources of measurement error that could be expected to influence the ratings.

Ideally, the variance component for persons should be large compared to the combined sources of error in the ratings. From this perspective the ratings for Exam 2 are much closer to the ideal than the ratings for Exam 3. Less than half (43%) of the variance in the ratings for Exam 3 is due to dependable differences in the students compared to 73% for Exam 2.
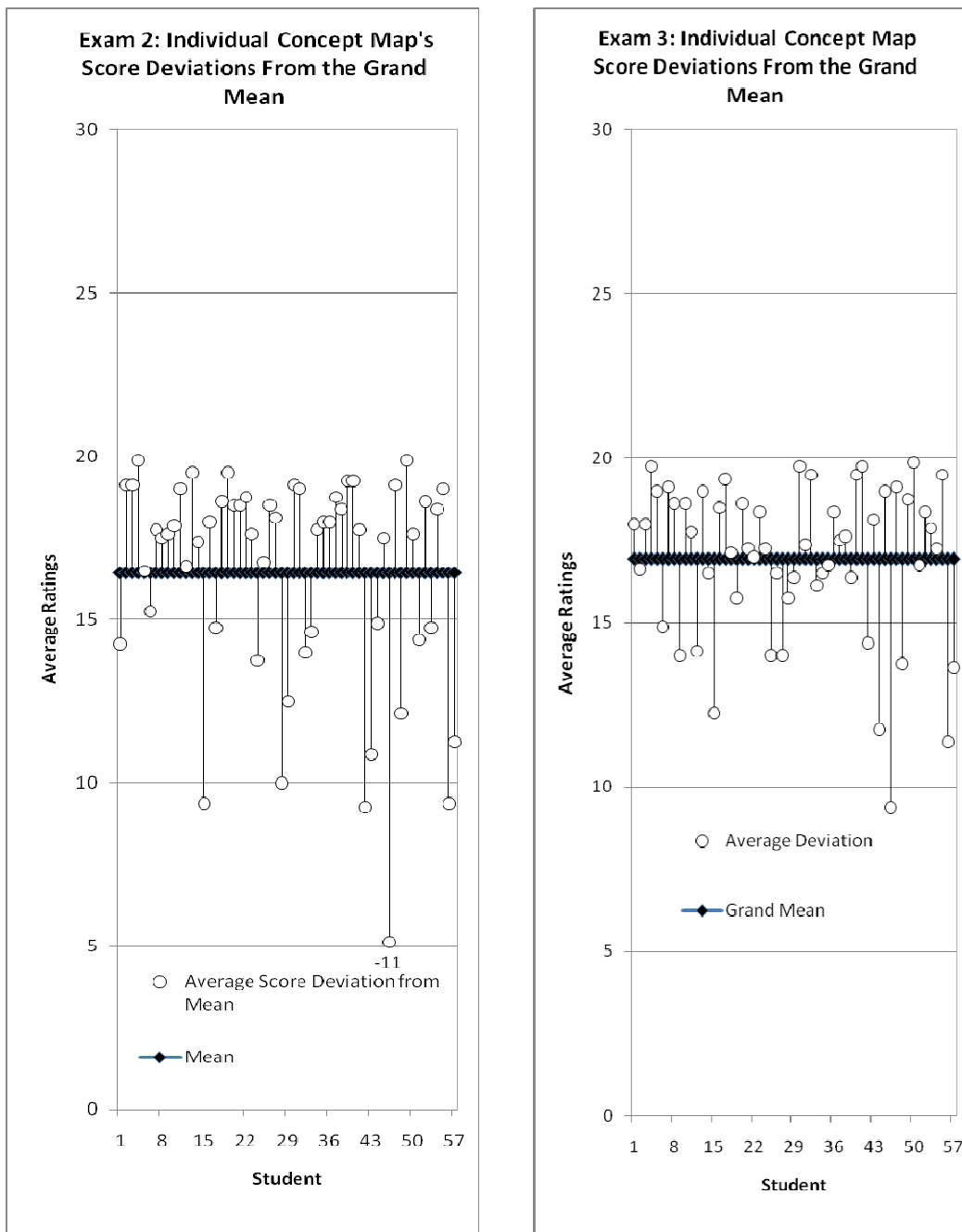
Figure 4. Variability of student mean ratings about the grand mean for Exams 2 and 3.

One reason for the differences in the ratings from the two exams may be due to differences in the subject-matter content of the two exams. But the variance component estimates reported in Table 1 indicate the ratings obtained from Exam 3 were subject to greater inconsistencies between the four raters and to an interaction between persons (students) and raters. We will explain these observed differences in greater detail later in this section.

When interpreting the data presented in Table 1, readers should keep in mind that the variance component estimates produced by a G-study are for single observations only (Brennam, 1983, 2001; Shavelson & Webb, 1991; Thompson, 2002) even though the ratings for each student analyzed in this study were collected from four raters who rated each midterm exam on two different occasions.

Inconsistencies in the Ratings

Research Question 2 includes four sub-questions that each focus on a different kind of inconsistency in the ratings due to different sources of measurement error.

Differences between raters. Research question 2a focuses on the inconsistencies between raters, lack of inter-rater reliability. The mean ratings (averaged across all 57 persons and both rating occasions) for each of the four raters is shown in Table 2 for each of the two exams. The variance components for raters describe the variability of the four raters' means about the grand mean for that exam. As shown is Table 1, the estimated value of the variance component for raters is 5% for Exam 2 and 25% for Exam 3.

The difference in the size of these two variance components is reflected by the spread of the rater means displayed in each column of Table 2. The mean ratings assigned by Raters 1 and 2 are consistently less (more severe) than the overall mean rating. Conversely, Rater 3 was

consistently more lenient than any other raters. The mean ratings at the bottom of Table 2

summarize the mean rating averaged across all four raters.

Table 2

Mean Ratings for Exam 2 and 3 Averaged Across Raters

	Midterm Exam	
Rater	2	3
1	15.82	16.46
2	15.67	16.30
3	17.68	19.28
4	16.39	15.31
Grand Mean	16.39	16.84

Differences across rating occasions. Research question 2b focuses on the

inconsistencies between raters across rating occasion. Table 1 shows results of the rater-by-

occasion interaction and we can see that the percent of the total variability due to rating occasion

is practically zero. Table 3 shows findings regarding differences in the mean ratings averaged

across the two rating occasions for each exam. The results indicate that these differences were a

negligible source of error and do not undermine the dependability of the concept map ratings.

This finding is consistent for both Exam 2 and 3. Therefore, a single rating occasion for each

exam would have been sufficient to assess the students' ability to connect concepts into a

meaningful whole.

Interactions. Research question 2c focuses on the three 2-way interactions.

One particular rater might be lenient while another might be much more stringent. The variance

components for the person-by-rater interaction is relatively small for Exam 2 (8%) and

somewhat higher for Exam 3 (13%) indicating that the relative ordering of the students were different from rater to rater within each exam.

Table 3

Mean Ratings for Exam 2 and 3 Averaged Across Raters and Rating Occasions

Rater	Exam 2			Exam 3		
	Occasion 1	Occasion 2	Mean	Occasion 1	Occasion 2	Mean
1	15.94	15.71	15.82	16.65	16.27	16.46
2	15.75	15.59	15.67	16.53	16.08	16.30
3	17.51	17.86	17.68	19.22	19.34	19.28
4	16.41	16.37	16.39	15.52	15.09	15.31

The person-by-occasion variance component indicates the degree to which raters consistently rate the persons across occasions. The percent of variability for the interaction between person and occasion in this study is practically zero (0%). These results indicate that the relative standing of students' ratings does not differ from one rating occasion to another. Figure 5 shows that the variance component for the rater-by-occasion interaction was zero (0%) for both Exam 2 and Exam 3.
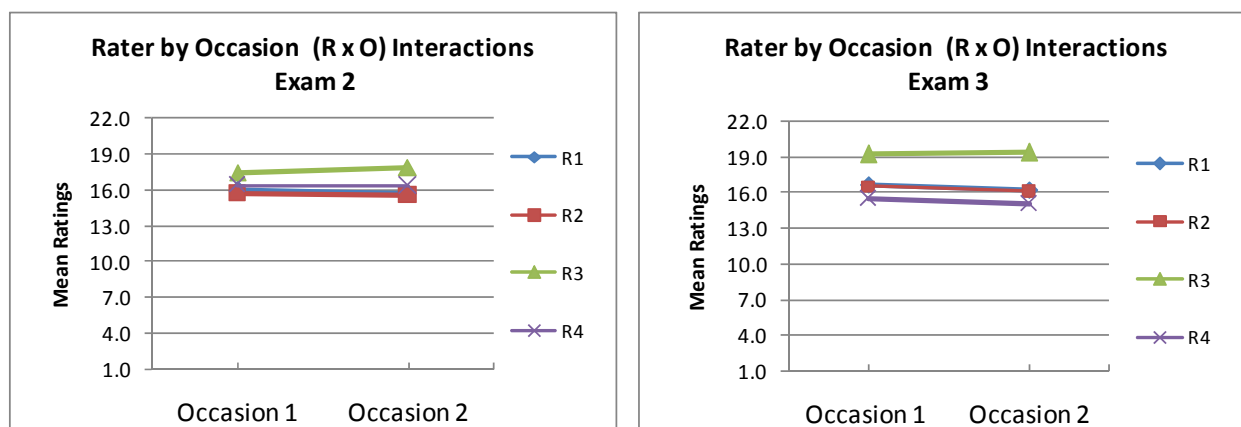

Figure 5. Rater-by-occasion interaction for Exams 2 and 3.

Residual error. Research question 2d focuses on the unexplained residual error that cannot be attributed to any of the identified sources of variability. The last row in Table 1 shows the residual error which accounts for roughly 14% of the variability in the ratings for Exam 2 and 19% of the variability for Exam 3. This means that more than 86% of the variability in Exam 2 and more than 81% of the variability in Exam 3 has been explained by the factors included in this study and their interactions.

Reliability Estimates

Research Question 3 focuses on the reliability of the mean ratings (averaged across the four raters and the two rating occasions) for Exam 2 and Exam 3. The results are summarized in the four graphs shown in Figures 6 and 7. The reliability of the mean ratings for Exam 2 was .95 for relative decisions and .93 for absolute decisions. The reliability of the mean ratings for Exam 3 was .88 for relative decisions and .78 for absolute decisions, respectively. These reliabilities can be seen on the fourth line starting from the bottom (squared marked line) of Figures 6 and 7.

Projected Effect of Changing the Number of Raters and Rating Occasions

Research question 4 focuses on estimating the projected effect of changing the number of raters and rating occasions. Using the variance components seen in Table 1, a D-study was conducted to estimate how changing the number of raters and rating occasions would increase or decrease the error variances and the generalizability coefficients.

The graphs in Figures 6 and 7 show how the reliability of the mean ratings varies as a function of the number of raters and rating occasions used to compute the mean for each examinee. Both increasing the number of raters and increasing the number of rating occasions will increase the reliability.
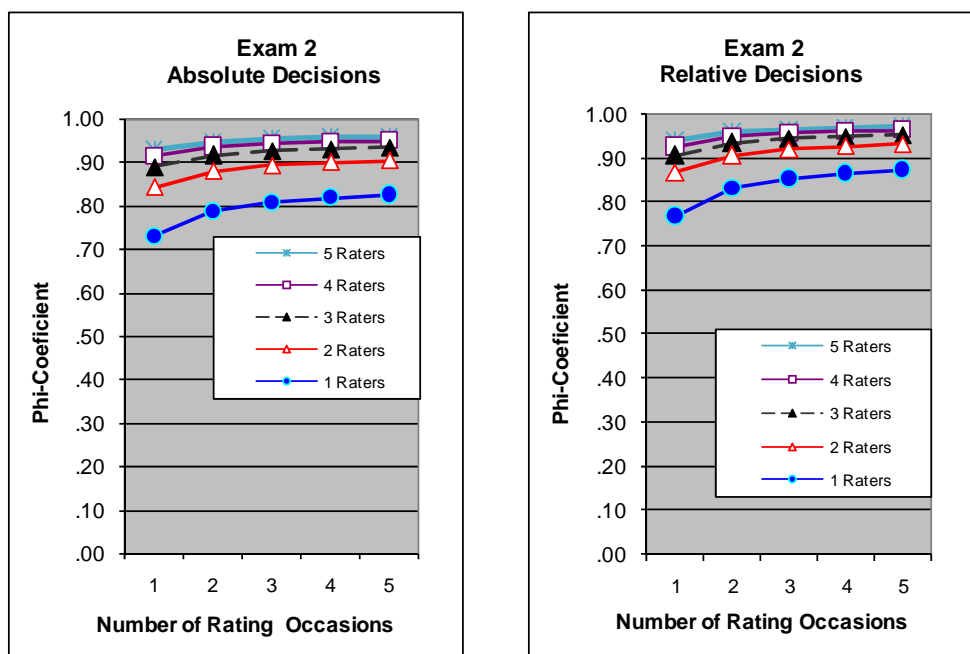
Figure 6. Reliability of relative and absolute decisions for Exam 2 as a function of the number of persons, raters, and rating occasions.

Each of the curvilinear lines in Figures 6 and 7 tend to become flatter as the number of raters and rating occasions increases. The effect of both of these changes has a rate of diminishing returns.

The two graphs in Figure 6 summarize the results of the reliability analysis for Exam 2. The two graphs in Figure 7 present the reliability results for Exam 3. The graphs on the left side of Figures 6 and 7 show the estimated reliabilities for relative decisions, and the graphs on the right hand side of Figures 6 and 7 show the estimated reliabilities for making absolute decisions.

Figure 6 shows the results for Exam 2 in which we can see that increasing the number of raters will increase the reliability more than increasing the number of rating occasions. The reliability for Exam 3 is generally lower than for Exam 2. However, the relationship between

increasing reliability patterns hold the same for both exams. In other words, increasing the

number of raters will increase the reliability of the mean ratings more than increasing the number
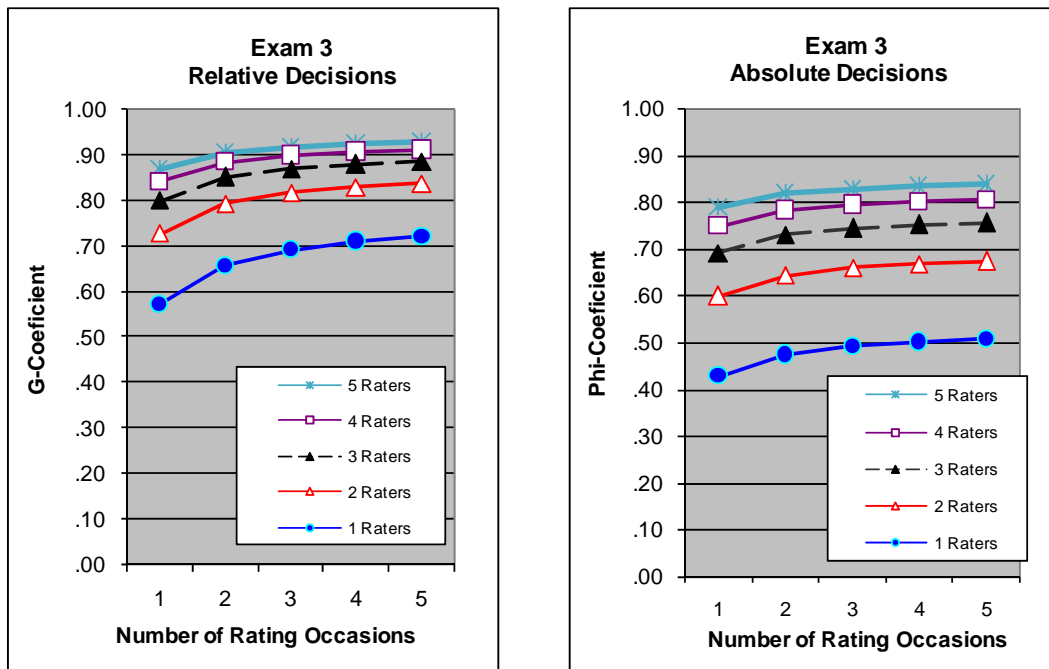
of rating occasions.

Figure 7. Reliability of relative and absolute decisions for Exam 3 as a function of the number

of person, raters, and rating occasions.

Discussion

The rationale for using concept maps for assessment purposes is that they are superior to other means of assessing students' understanding of how well they understand the relationships among the component concepts in a conceptual domain. The maps have to be evaluated or judged in order to translate them into assessments and this translation effort is necessarily a rater-mediated process. The idea of using concept maps as an assessment device is based on the assumption that the maps can be reliably rated. This assumption is tantamount to asserting that the variability in the ratings is due mainly to differences in the nature of the students' understanding rather than to differences between the raters or differences in individual raters from one rating occasion to another.

Summary

The twofold purpose of this study was (a) to determine the relative contribution of various sources of variability in the ratings of concept maps obtained from students in an introductory, college-level, biology course and (b) to identify how the reliability of the ratings is likely to vary as a function to the number of raters and rating occasions utilized. We used the C-mapping approach to delimit the scope of the task and the pencil-and-paper draw-a-map mode to further define the task. In addition, we focused on rating the adequacy of the propositions supplied by the students to describe the relationships between the component concepts in the domain. The scoring rubric focused on three aspects of the propositions: (a) importance, (b) completeness, and (c) accuracy.

Two separate generalizability studies were conducted as part of this study. The first G-study was conducted to estimate various sources of variability in the ratings of the concept map administered in association with Exam 2. The second G-study was conducted to estimate the

various sources of measurement error in the concept maps obtained from Exam 3. The researcher assessed the degree to which the ratings were subject to inconsistencies between raters (inter-rater variability), inconsistencies with individual raters from one rating occasion to the next (intra-rater variability), and variability in the ratings from the three 2-way interactions.

Because of these multiple sources of potential error that can be simultaneously operating in the ratings, the reliability of the ratings were expected to take different values depending on the number of raters and the number of rating occasions used. Therefore, the size of the reliability coefficient for a set of ratings is best conceptualized as a mathematical function rather than as a single value that holds under all measurement conditions.

For Exam 2, the results of this study indicated that the largest sources of measurement error in the ratings were (a) inconsistencies between the mean ratings obtained from the four raters (5% of the total variability), (b) differences in the relative ordering of the person means by the four raters (8% of total variability), (c) and the residual error not otherwise accounted for by the design (14% of the total variability).

The percent of the total variability for each of these three sources was larger for Exam 3 than for Exam 2: (a) inconsistencies between the mean ratings obtained from the raters accounted for 25% of the total variability, (b) differences in the relative ordering of the person means by the four raters (13% of the total variability), and (c) the residual error not otherwise accounted for by the design (19% of the total variability).

The reliability of the mean ratings varies as a function of the number of raters and rating occasions used to compute the mean for each examinee for Exam 2. Increasing the number of raters and increasing the number of rating occasions will both increase the reliability. Both of

these effects depict a trend of diminishing returns resulting from increasing the number of raters and rating occasions, respectively.

Conclusions

The purpose of this study was to estimate the reliability of the ratings resulting from using the C-mapping technique as a means of assessing students panoramic understanding of a conceptual domain and the interrelationships of the concepts within that domain. The resulting ratings demonstrated high degree of reliability. The C-mapping approach coupled with the pencil-and-paper draw-a-map mode, plus a scoring rubric that focuses on three aspects of the propositions: (a) importance, (b) completeness, and (c) accuracy can produce dependable measures of Biology students' connected understanding. Plummer, (2008) pointed out that each proposition on the map should be rated separately using a scale for each rating element. He found out that a scoring method of this vein would make the ratings less cognitive taxing for raters.

The two largest sources of error variance in the ratings of both examinations included (a) the raters, and (b) the person-by-rater interaction. Hence, averaging each students' ratings across multiple raters has a greater effect in increases the reliability than averaging across multiple rating occasions.

The use of well trained raters in using the scoring criteria can contribute to obtain high reliability coefficients.

Implications of Using Concept Maps for Assessment Purposes

When using a concept map for classroom assessment the following recommendation should be considered. These recommendations are based on the results of this study.

1. Before concept maps are used for assessment purposes students should be taught how to create concept maps. They should also be taught the criteria that would be used to rate the maps as a means to judge their own understanding, by identifying the importance, accuracy, and completeness of the proposition included in the maps.

2. Raters should also be taught how to construct maps. But rater training should focus on the content of the concept maps rather than the form. They should be trained to locate the focal concepts, supporting concepts, and all the propositions included in the map regardless of the physical layout of the map components. They must also be taught the meaning of the criteria and how to consistently apply those criteria

3. At least two raters and at least two rating occasions should be used initially. If the variance component for rating occasions is small compared to the other sources of variation, then the number of rating occasions can be reduced to one.

4.   The raters selected should have a thorough knowledge of the course content. This trait could facilitate the construction of a quality master map. More than two raters should be used during master map construction to take advantage of their different levels of knowledge and points of view.

5. The instructor should participate directly in designing and constructing the master maps by delimiting the domain, choosing the concepts to be included, and more importantly identifying the focal concept to which all the other concepts should be related

Recommendations for Future Research

The following issues are unresolved and need to be investigated in future research.

1. If concept maps are used as the primary means of assessment in a course, how will this practice affect the way the students study in that course?

a. How will this practice affect students' performance in subsequent courses in which they enroll in the same discipline?

b. How will it affect their long term memory of what they have learned in the course?

c. What other advantages, if any, accrue to students who invest the time and effort to develop an integrated, panoramic ("big picture") view of a conceptual area?

2. To what degree do concept map ratings lead to valid conclusions about students' understanding of the targeted conceptual domain? The present study focuses only on the reliability of ratings of students' conceptual understanding. Reliability is typically viewed as a necessary, but not sufficient condition for validity. The validity of the concept map ratings needs further research.

3. What are the relative advantages and disadvantages of using concept maps compared to extended-response essay questions?

a. To what extent is there evidence to support the conclusion that concept maps are more economical than essay questions because the maps can be used to assess students' understanding of a broader conceptual domain in approximately same amount of time required for students to respond to an essay question on the average? In my review of the research literature I was unable to locate any research on this issue.

b. How does the amount of time and effort required to reliably score responses to concept maps compare with the amount of time required to reliably score extended response essay questions?

Contribution to the Field

Previous researchers have investigated the degree to which concept-map assessments generate reliable scores (Ruiz-Primo et al., 2001; Ruiz-Primo & Shavelson, 1996., McClure, 1999., Ruiz-Primo et al. (2001) claimed that the C-mapping technique was a practical useful way to obtain reliable ratings of concept maps, but several rating challenges have been raised with regard to the use of C-mapping (Yin et al., 2005). These rating challenges have been manifested in the difficulty of accounting for proposition choice. Proposition choice is referred to in this study as proposition importance. The rubric used in this study was paired with the C-mapping technique. The rubric was designed to account for three attributes of a concept map proposition: proposition importance, completeness, and accuracy. Completeness and accuracy attributes were added in hopes of obtaining a more thorough rating of each proposition in its context.

References

Acton, W., Peter, J., & Goldsmith, T. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86,* 303–311.

Anderson, T. & Huang, S. (1989). *On using concept maps to assess the comprehension effects of reading expository text* (Tech. Rep. No. 483). Center for the Studying of Reading, University of Illinois at Urbana–Champaign. (ERIC Document Reproduction Service No. ED310368).

Ausubel, D. (1963). *The psychology of meaningful verbal learning.* New York: Holt, Rinehart, and Winston.

Baker, E., Niemi, D., Novak, J., & Herl, H. (1991). *Hypertext as a strategy for teaching and assessing knowledge representation.* Paper presented at NATO Advanced Research Workshop on Instructional Design Models for Computer-Based Learning Environments, Enschede, The Netherlands.

Barron, R. F. (1969). The use of vocabulary as an advance organizer. In H. L. Herber & P. L. Sanders (Eds.). *Research in reading in the content areas*: *First year report* (pp. 81-87). Syracuse NY: Syracuse University Reading and Language Arts Center.

Bransford, J. D., Brown, A.L., & Cocking, R.R. (Eds.) (2000). *How people learn: Brain, mind, experience and school.* Washington, DC: National Academy Press.

Brennan, R. (1992). *Elements of generalizability theory.* Iowa City, IA: American College Testing.

Cohen, G. (1983). *The psychology of cognition* (2nd ed.). London: Academic Press, Inc.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability of scores and profiles.* New York: Holt, Rinehart, & Winston.

Fisher, K. (1990). Semantic networking: The new kid on the block. *Journal of Research in Science Teaching*, *27*, 1001–1018.

Fisher, K. M. (2000). *Mapping biology knowledge*. Dordrecht, Holland: Kluwer

Francisco, J. S., Nakhjleh, M. B., Nurrenbern, S. C., & Miller, M. L. (2002). Assessing student understanding of general chemistry with concept mapping. *Journal of Chemical Education, 79,* 248–257.

Jacobs-Lawson, J. M. & Hershey, D. A. (2001). Concept maps as an assessment tool in psychology courses. *Teaching of Psychology, 29,* 25–29.

Kinchin, I. M. (2000). Using concept maps to reveal understanding: A two-tier analysis. *School Science Review*, *81*, 41–46.

Liu, X. (2002). Using concept mapping for assessing and promoting relational conceptual change in science. *Science Education, 88,* 373–396.

Liu, X. & Hinchey, M. (1996). The internal consistency of a concept mapping scoring scheme and its effect on prediction validity. *International Journal of Science Education, 15,* 921–937.

Mawson, COS. (1975). *Dictionary of foreign terms (2nd* ed.), (Revised and updated by C. Berlitz). New York, NY: Thomas Y. Crowell.

McClure, J., Sonak, B., & Suen, H. (1999). Concept-map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36,* 475–492.

Merrill, M.D. & Tennyson, R.D. (1977). *Teaching concepts: An instructional design guide*. Englewood Cliffs, NJ: Educational Technology Publications.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (1998). *Teaching science for understanding: A human constructivist view*. San Diego, CA: Academic.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nakhleh, M. B. & Krajcik, J. S. (1991). The effect of level of information as presented by different technology on students' understanding of acid, base, and pH concepts. Paper presented at the annual meeting of the National Association for the Research in Science Teaching, Lake Geneva, WI.

Nicoll, G., Francisco, J., & Nakhleh, M. (2001). An investigation of the value of using concept maps in general chemistry. *Journal of Chemical Education, 78,* 1111– 1117.

Novak, J. & Gowin, B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Osborne, R. & Wittrock, M. (1983). Learning science: A generative process. *Science Education, 67*, 489–508.

Palinscar, A. S. (1998). Social constructivist perspectives on teaching and learning. In J.T. Spence, J. M. Darley, & D. J. Foss (Eds.), *Annual Review of Psychology* (pp. 345–375), Palo Alto, CA: Annual Reviews.

Phillips, D. (1997). How, why, what, when, and where: Perspectives on constructivism and education. *Issues in Education: Contributions from Educational Psychology, 3*, 151–194.

Pope, M. & Gilbert, J. (1983). Personal experience and the construction of knowledge in science. *Science Education, 67*, 193–203.

Plummer K. (2008). Analysis of the psychometric properties of two different concept-map assessment tasks. Unpublished doctoral dissertation Brigham Young University, Provo UT.

Rice, D. C., Ryan, J. M. & Samson, S. M. (1998). Using concept maps to assess student learning in the science classroom: Must different method compete? *Journal of Research in Science Teaching, 35*, 1103–1127.

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*, 569–600.

Ruiz-Primo, M. A., Li, M., Yin, Y., Shavelson, R. J., Vanides, J., Schultz, S., & Ayala, C. (2004). *Concept maps as an assessment tool: A framework for examining their cognitive validity.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ruiz-Primo, M. A., Schultz, S., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques*. Journal of Research in Science Teaching*, *8*, 260–278.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7* (2), 99–141.

Ruiz-Primo, M. A., Shavelson, R. J., & Schultz, S. E. (1997). *On the validity* of concept map based assessment interpretations: An experiment testing the assumption of hierarchical concept-maps in science. Paper presented at the American Educational Research Association, Chicago, IL.

Rye, J. & Rubba, P. (2002). Scoring concept maps: An expert map-based scheme weighted for relationships. *School Science and Mathematics, 102*, 33–46.

Schau, C. & Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *The American Statistician*, *51*, 171–175

Shavelson, R., Lang, H., & Lewin, B. (1993). *On concept maps as potential "authentic" assessments in science.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED367691).

Shavelson, R. & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*, 115–163.

Smith, E. & Medin, D. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Southerland, S. A., Smith, M. U., & Cummins C.L. (1998). "What do we mean by that?": Using structured interviews to assess science understanding. In J. J. Mintzes., J. H. Wandersee., & J. D. Novak, (Eds.). *Teaching science for understanding: A human constructivist view* (pp. 72-92). San Diego, CA: Academic.

Wallace, J. & Mintzes, J. (1990). The concept map as a research tool: Exploring conceptual change in biology. *Journal of Research in Science Teaching, 27,* 1033–1052.

Thompson, B. (Ed.) (2003). *Score reliability*. Thousand Oaks, CA: Sage.

West, D. C., Park, K. P., Pomeroy, J. R., & Sandoval, J. (2002). Concept-mapping assessment in medical education: A comparison of two scoring systems. *Medical Education*, *36*, 820–826.

Wandersee, J.H. (1990). Concept mapping and the cartography of cognition. *Journal of Research in Science Teaching, 27*, 923–936.

Yin, Y., Vanides, J., Ruiz-Primo, M., Ayala, C., & Shavelson, R. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching, 42,* 166–184.

Appendix A

Informed Consent: Consent to be a Research Subject

Introduction. Richard R. Sudweeks, Ph.D. and Laura Jimenez M.Ed. are conducting a study, in an effort to analyze the effectiveness of concept maps as a measure of how students interconnect or interrelate concepts. You were selected to participate because you are currently taking Biology 100 with Dr. Booth.

Procedures. You have been randomly selected to participate in a ten to fifteen minute interview. In this interview you will be asked a series of questions regarding your understanding of how a list of concepts from the last midterm exam are interrelated. This exercise will not be graded.

Furthermore, after you have received a grade for your concept mapping and essay exam questions by Dr. Booth's TAs, your responses to these test questions will be rescored using a specialized scoring method. <u>The rescoring of your exams will in no way whatsoever affect your grade</u>.

Minimal risks/discomforts. There will be no minimal risks of discomforts with the concept map intervention in class.

Benefits. There are no foreseeable benefits to students that would result from the interviews.

Confidentiality. All information provided will remain confidential and will only be reported as group data with no identifying information, and only those directly involved with the research will have access to it. The resulting scores will be seen only by the researchers specified above and one of Dr. Booth's research assistants, Julie Low. Your concept maps, essays, and interview transcripts will be assigned a number in an effort to maintain your

anonymity. As previously explained, audio tapes will be destroyed, and only the researchers will have access to the transcripts which will have no identifying information.

Compensation. If you consent to participate in the interview described above, as compensation for your time, you will be given a $10.00 gift certificate to be used at the BYU bookstore for each interview in which you participate.

Participation. Participation in this research study is voluntary. You have the right to withdraw at anytime or refuse to participate entirely without jeopardy to your class status, grade, or standing with the university.

Questions about the research. If you have questions regarding this study, you may contact Dr. Gary Booth at 422-2458, gary_booth@byu.edu; Dr. Richard R. Sudweeks at 422-7078, richard_sudweeks@byu.edu; or Laura Jimenez at 422-4975, ljimenezron@gmail.com

Questions about your rights as research participants. If you have questions you do not feel comfortable asking the researcher, you may contact Dr. Renea Beckstrand, IRB Chair, 422-3873, 422 SWKT, renea_beckstrand@byu.edu. I have read, understood, and received a copy of the above consent and desire of my own free will and volition to participate in this study.


Signature:_____   Date:_____

Appendix B

Concept Map Exam 1 Test Item

The skeletal concept map below includes some blank nodes and blank links. The Exam's central concept is highlighted with a thicker line. Your task is to fill in the blank nodes and blank links from the list of concepts and linking phrases provided at the bottom of the page. After your concpet map is completed, it should reflect the relationship between concepts .
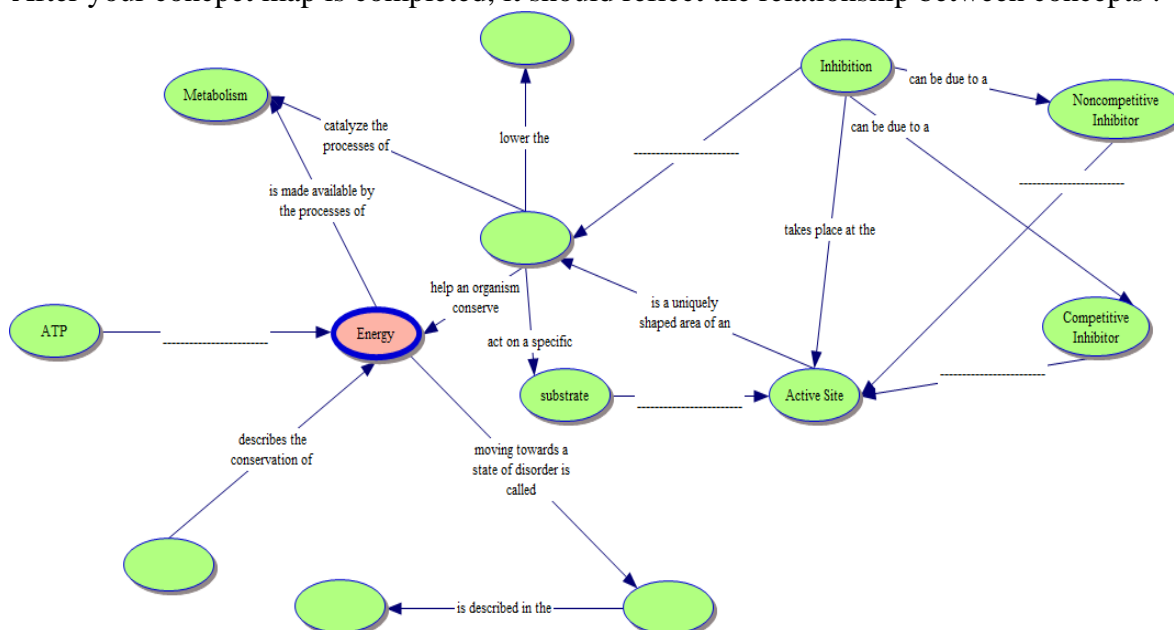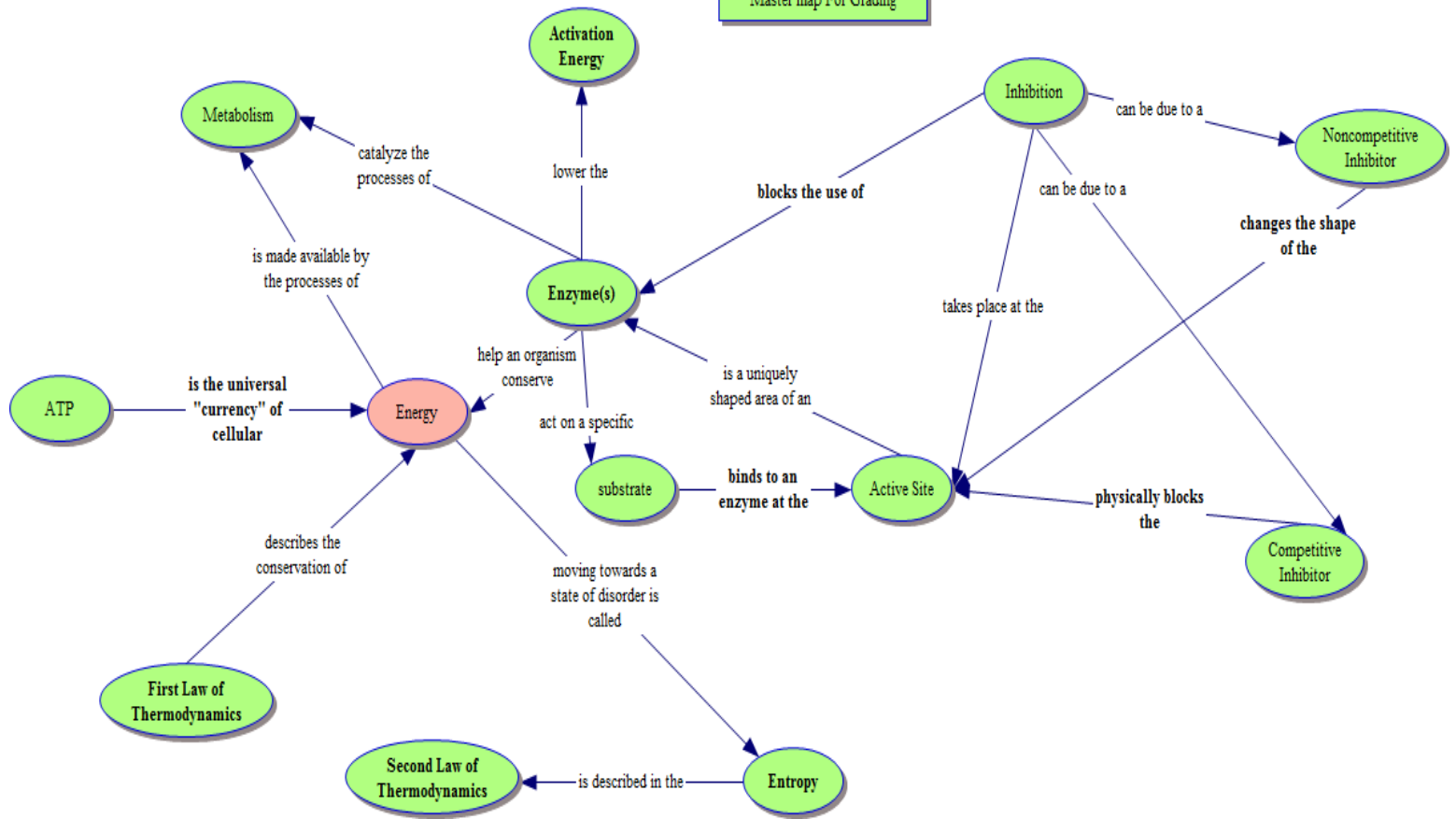

List of Linking Phrases
1. is the universal "currency" of cellular
2. binds to an enzyme at the
3. blocks the use of
4. physically blocks the
5. changes the shape of the

List of Concepts
1. Entropy
2. Activation energy
3. Enzyme(s)
4. Second Law of Thermodynamics
5. First Law of Thermodynamics

Exam 1: Master Map Lecture 5

Lecture 5
First Exam
Master map For Grading

Activation Energy

Metabolism

Inhibition

can be due to a

Noncompetitive Inhibitor

catalyze the processes of

lower the

blocks the use of

can be due to a

changes the shape of the

is made available by the processes of

Enzyme(s)

takes place at the

ATP

is the universal "currency" of cellular

help an organism conserve

is a uniquely shaped area of an

Energy

act on a specific

substrate

binds to an enzyme at the

Active Site

physically blocks the

Competitive Inhibitor

describes the conservation of

moving towards a state of disorder is called

First Law of Thermodynamics

Second Law of Thermodynamics

is described in the

Entropy

Exam 2: C-Mapping Task Test Item

Instructions: Construct a concept map on a separate sheet of paper showing how the concepts listed below are interrelated. Your map should be constructed around the central concept: **Cell**

---

1. Include all of the concepts listed below in your map.

2. Your map should include at least 24 propositions.

---

1.   **Cell**
2.   Double Helix
3.   hydrogen bonds
4.   Replication
5.   DNA
6.   Uracil
7.   mRNA
8.   Transcription
9.   Central Dogma
10.   Ribosome
11.   Translation
12.   Anti-codon
13.   Codon
14.   tRNA
15.   AUG
16.   Stop Codon
17.   Protein
18.   Amino Acid(s)
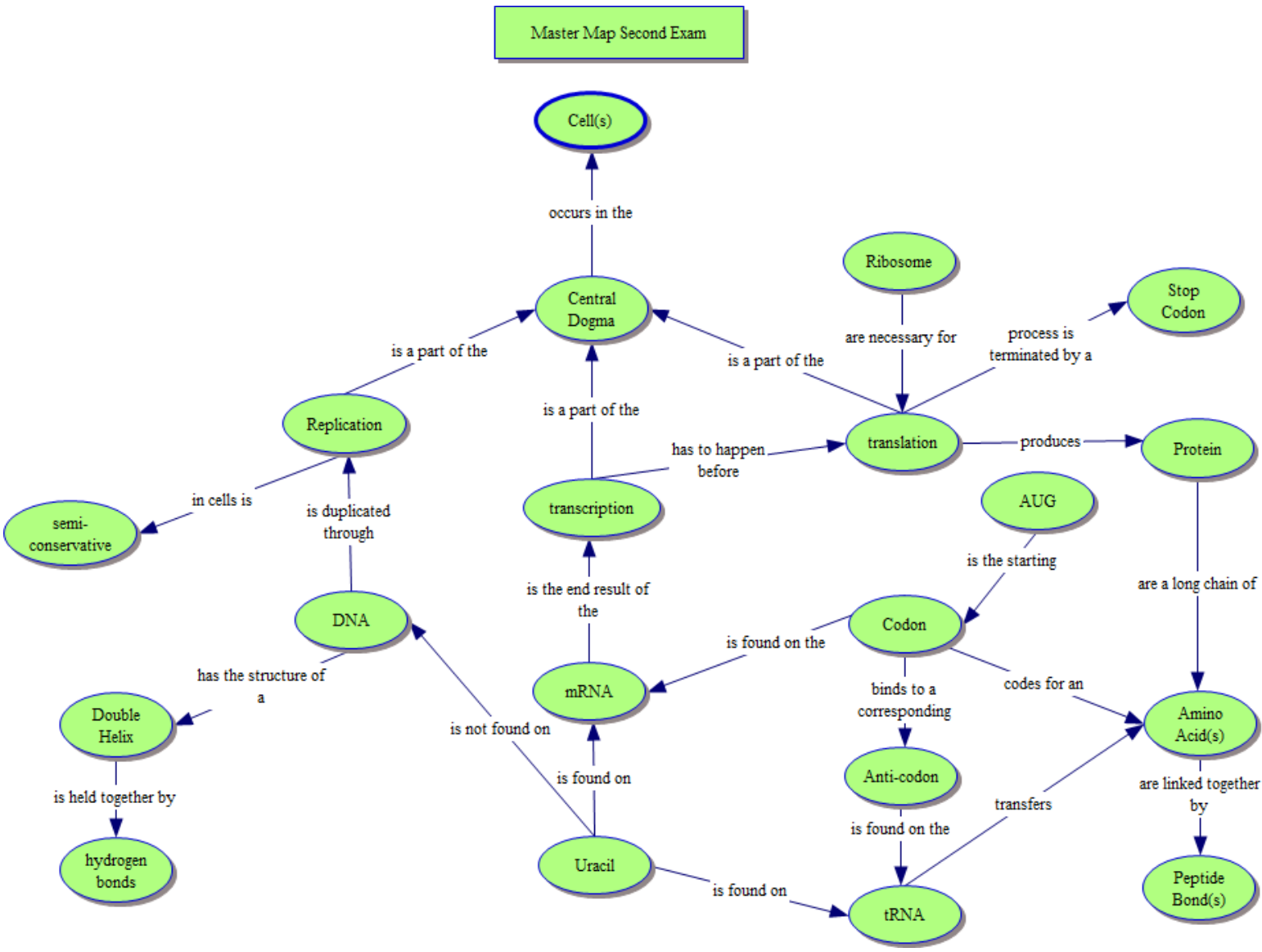19.   Peptide Bond(s)
20.   Semi-conservative

Each connection between two concepts will be rated based on

---

1.   The <u>importance</u> of the linked concepts.

2.   The <u>accuracy</u> of the proposition, and

3.   The <u>completeness</u> of the linking phrase expressing the relationship between concepts, in other words, the completeness of the proposition.

Each proposition in your map will be given -1 to 3 points based on its importance, accuracy and completeness. Your total score will be the sum of the points given for each proposition. One point will be deducted for each incorrect proposition.

*Note: A **<u>proposition</u>** is an element of a Concept Map (in other words <u>the sentence</u>) that results after connecting two concepts and a linking phrase.

Exam 2: Master map

Master Map Second Exam

Cell(s)

occurs in the

Central Dogma

Ribosome

Stop Codon

is a part of the

is a part of the

process is terminated by a

are necessary for

Replication

is a part of the

has to happen before

translation

produces

Protein

in cells is

is duplicated through

transcription

AUG

is the starting

semi-conservative

is the end result of the

is found on the

Codon

codes for an

are a long chain of

DNA

mRNA

binds to a corresponding

Amino Acid(s)

has the structure of a

is not found on

Anti-codon

transfers

Double Helix

is found on

is found on the

are linked together by

is held together by

Uracil

is found on

tRNA

Peptide Bond(s)

hydrogen bonds

Appendix C

Exam 3: C-Mapping Task Test Item

Instructions: Construct a concept map on a separate sheet of paper showing how the

concepts listed below are interrelated. Your map should be constructed around the central

concept: **<u>Evolution</u>**

1. Include all of the concepts listed below in your map.

2. Your map should include at least 13 propositions.

1.   **<u>Evolution</u>**
2.   Recessive
3.   Phenotype
4.   dominant
5.   Prophase I
6.   Allele(s)
7.   genotype
8.   genotype
9.   Crossing over
10.   Recombinants
11.   Genetic Variation
12.   Vestigial Features
13.   mutation
14.   chromosome(s)
15.   Linked genes
16.   Sex linkage

Each connection between two concepts will be rated based on

1.   The importance of the linked concepts.
2.   The accuracy of the proposition, and
3.   The completeness of the linking phrase expressing the relationship between concepts, in
other words, the completeness of the proposition.

Each proposition in your map will be given -1 to 3 points based on its importance,

accuracy and completeness. Your total score will be the sum of the points given for each

proposition. One point will be deducted for each incorrect proposition.

*Note: A **<u>proposition</u>** is an element of a Concept Map (in other words <u>the sentence</u>) that results after connecting two concepts and a linking phrase.

Exam 3: Master Map

Exam 3
Master map

Vestigial
Features

are an evidence for

Evolution

contributes to

Phenotype

genotype

can cause

mutation

physical expression
of the

genetic composition
of

can alter the

Recessive

can be

Allele(s)

can be

dominant

is the ultimate
source of

can alter the

is the process of
exchanging

Genetic
Variation

is a source of

Prophase I

occurs in

Crossing
over

occurs between
homologous

chromosome(s)

results in

are both found on
the same

refers to genes that
are only found in
the sex-

Linked
genes

Recombinants

Sex linkage

Appendix D

Biology 100 Concept Map Training Packet—*Fall 2008 Dr. Booth's Biology 100*

Class "Student Concept Map Training Packet"

1. Purpose of Concept Mapping

2. What is a Concept Map – Example of a Concept Map

3. Elements of a Concept Map

4. Fill in the Blank Concept Mapping Assignment

5. Construct a map from a list of concepts (C-map) Assignment

Purpose of Concept Mapping

In order to "really" understand a subject like Biology we must engage in the learning of

1. Biological Facts –Watson and Crick discovered the structure of a DNA.

2. Biological Concepts –*DNA or mRNA*

3. How Biological Concepts are Interrelate – ***mRNA*** *is a mobile transcription of **DNA***

Learning occurs when you bring order / structure to the information you are receiving. Concept mapping is one of many ways you can facilitate for yourself an organized understanding of a subject.

In Dr. Booth's class this Fall you will learn Facts, Concepts, and how these Concepts interrelate with one another. We will use Concept Mapping as a way to see how the concepts you learn this semester are interrelated.

What is a Concept Map? –Example of a Concept Map

A concept map is a graphic representation intended to reveal a student's understanding of how the concepts within a content domain are interrelated. An example of a concept map is shown in Figure 1.
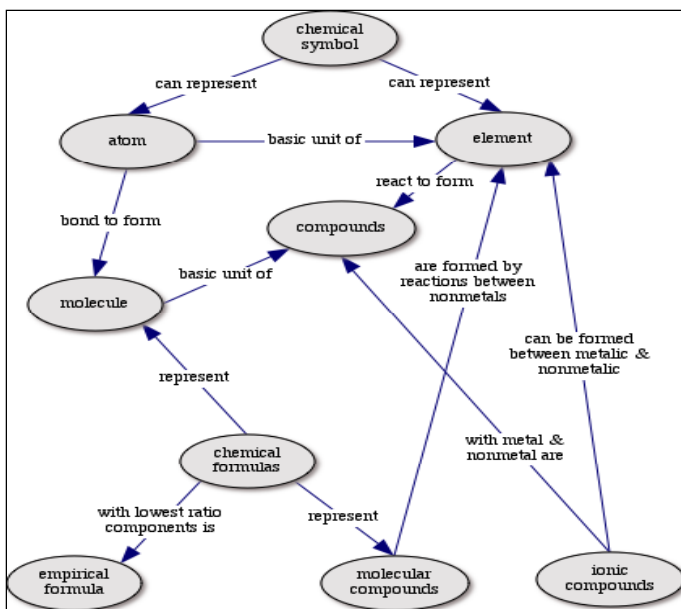
Figure 1. Example of a concept map.

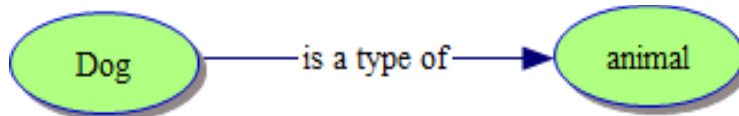**Elements of a Concept Map**

A concept map has

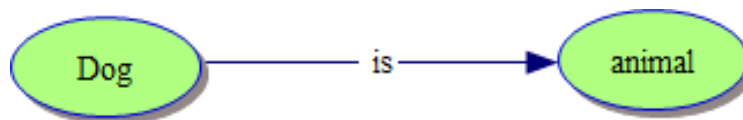1.   **words** representing **concepts** that you write in an ellipse

2. **linking phrases** which specify a relationship between two concepts. The concepts and linking phrase form a **proposition,** e.g. "Dog is a kind of animal". The proposition should communicate a complete thought.

Good example of a proposition:
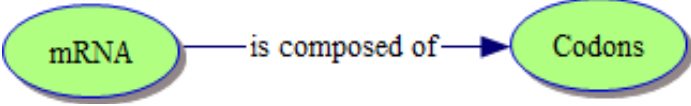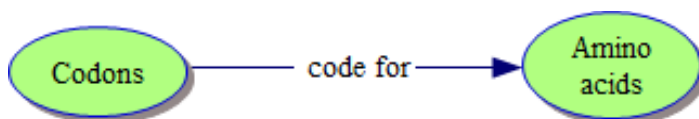

Bad example of a proposition**:**

Linking phrase training. The linking phrase should communicate a complete and accurate relationship between two concepts. Linking phrases can describe how two concepts are related by their involvement in an important process, their structure, and/or organization. You may add new concepts in the linking phrase only if absolutely necessary, and only if this helps make the relationship between concepts complete and accurate.
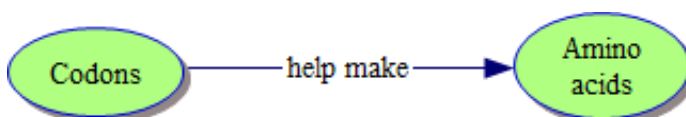
Example of structural relationship:

Example of relationship by involvement in a process:
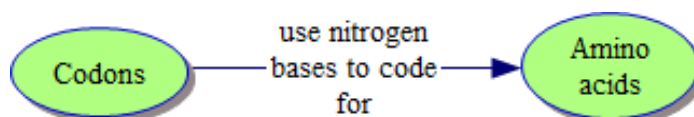
Even if we remove the "use nitrogen bases to" element of the linking phrase, the most important element that describes the process remains.


Example of a less effective relationship by involvement in a process

A linking phrase that serves to establish a relationship between concepts should form a proposition that is *important*, *accurate,* and *complete*. While necessary in some cases, students should avoid excessive use of redundant and less meaningful linking phrases such as "can be," "is a type of," or "follows" when more meaningful relationships can be expressed. Consider the following linked concepts:

Good example:
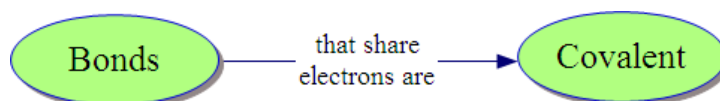
Notice that the concept "electrons" is in the linking phrase, because it is a critical part of the relationship between the concepts "bonds" and "covalent"

Bad example:
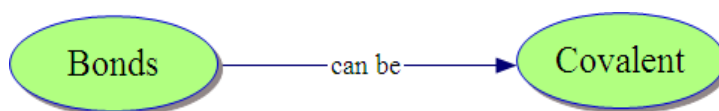
Although <u>important</u> and <u>accurate,</u> this linking phrase is <u>incomplete</u>. The student has not demonstrated adequate understanding of the relationship between the two concepts in a meaningful way.

Fill in the Blank Concept Map Assignments

For Lectures 3-7 you will be given assignments to complete skeletal concept maps by filling in blank nodes and linking phrases as shown in the following example (see next page).

Concept Map assignment for Lecture # 3

The skeletal concept map below includes some blank nodes and blank links. The lecture's central concept is highlighted with a thicker line. Your assignment is to fill in the blank nodes and blank links from the list of concepts and linking phrases provided at the bottom of the page. After your concept map is completed, it should reflect the content provided in Lecture 3. This assignment is worth ten (10) points. You will be given one point for each correctly labeled concept node and link.
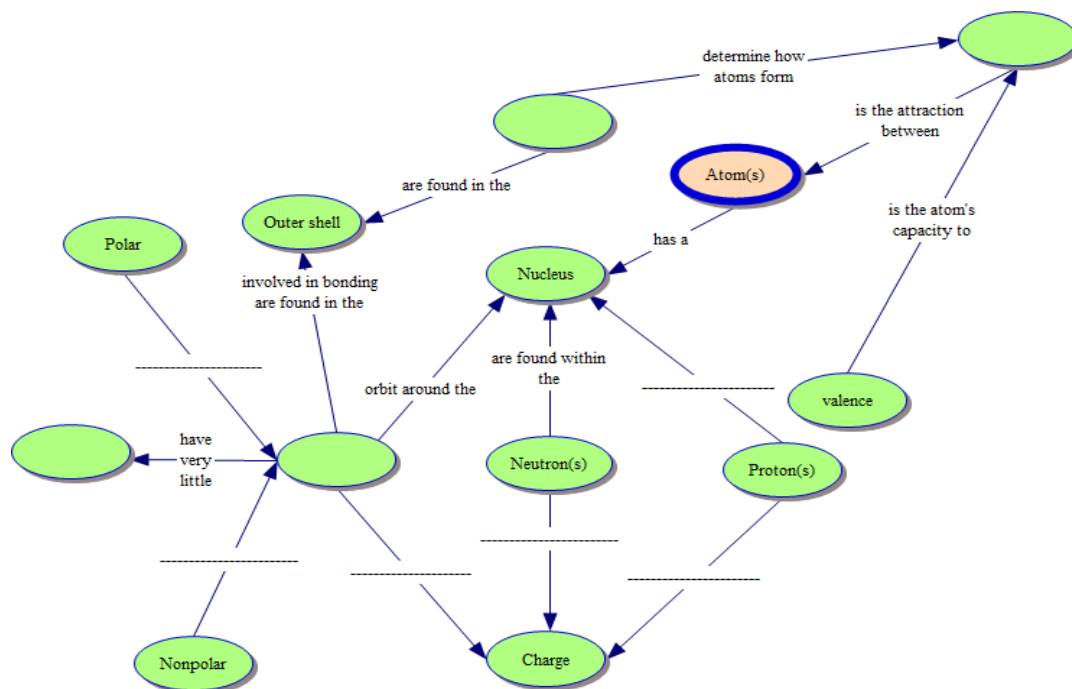
List of Concepts

1. Mass
2. Electron(s)
3. Valence electrons
4. Bond

List of Linking Phrases

1. is a molecule with an equal distribution of
2. have a negative
3. is a molecule with an unequal distribution of
4. have no
5. have a positive 6. are found within the

Construct a map from a list of concepts (C-map) Assignments

After the seventh lecture your assignments will consist in constructing your own concept map from a list of concepts. An example of this assignment with its corresponding master map can be seen below

Sample of C-mapping task assignment (Lecture 4). Instructions: Construct a concept map on a separate sheet of paper showing how the concepts listed below are interrelated. Your map should be constructed around the central concept: **Biological Macromolecules**

> 1. Include all of the concepts listed below in your map.
>
> 2. Your map should include at least 18 propositions.

1.   Biological Macromolecules
2.   DNA
3.   Starches
4.   Polysaccharides
5.   Monosaccharides
6.   Sugars
7.   Carbohydrates
8.   Fats
9.   Steroids
10.   Lipid(s)
11.   Phospholipids
12.   Proteins
13.   Amino Acids
14.   Peptide Bonds
15.   Nucleic Acids
16.   Condensation Synthesis
17.   Hydrolysis

Each connection between two concepts will be rated based on

1.   The importance of the linked concepts.

2.   The accuracy of the proposition, and

3.   The completeness of the linking phrase expressing the relationship between concepts; in

other words the completeness of the proposition.

> Each proposition in your map will be given -1 to 3 points based on its importance,
>
> accuracy and completeness. Your total score will be the sum of the points given for each
>
> proposition. One point will be deducted for each wrong proposition.

Sample of a Master Map (Lecture 4)
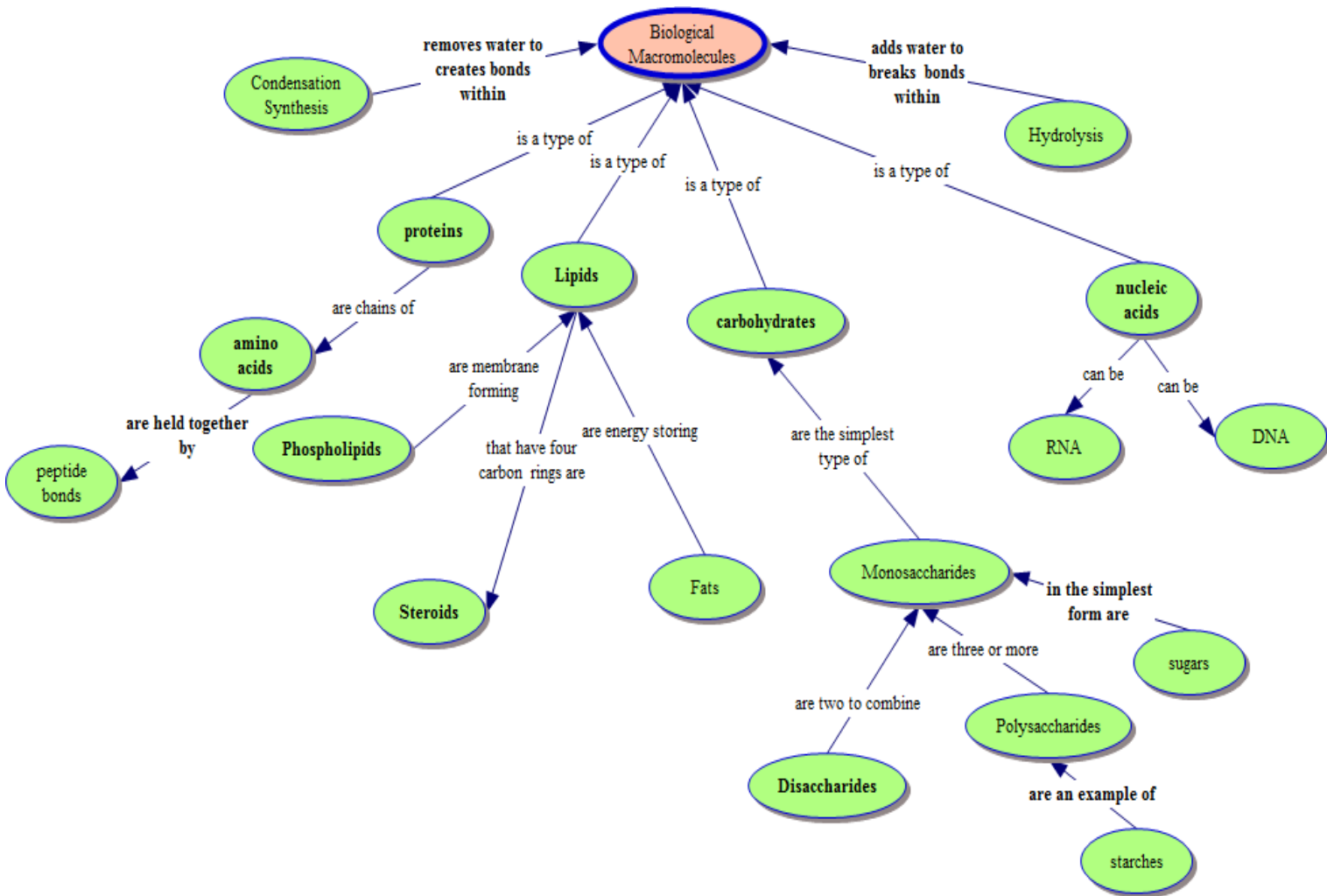
Concept Map (C-Mapping Task) Construction Guidelines

Below you will find a description of how the map's propositions should be constructed and how the propositions in the map will be scored.

1. Students' maps should be constructed around a specified **<u>central concept</u>** (in the map above, the ellipse line is in bold).

2. A **list of concepts** to be included in the map will be provided.

3. The maximum number of **propositions** that will be scored will be specified.

4. Each scored proposition will be given -1 to 3 points based on its importance, accuracy, and completeness. The central concept, list of concepts, and maximum number of propositions will be determined based on a master map approved by the course instructor.

5. The total score for a complete map is the sum of the points given for each proposition.

6. The total number of points possible for a given concept map will be three (3) times the number of key or important propositions identified by the instructor on the master map and specified in the item instructions. However, since each map does not have the same number of propositions, the final score possible for each map will be converted to a 1-10 scale so that each map will be worth ten points total. Thus if the master map includes 15 important propositions (and the item instructions specify that 15 propositions should be included on the map), the maximum total score would be 45 (3x15). If a student scores 40 out of 45 on his or her propositions, then the equalized score for the map would be: 40/45 = 0.8888 X 10 = 8.8 points out of 10.

7. Propositions included on a student map that are not identified on the master map will only receive points if the total number of propositions has not been reached.

The Scoring Rubric

Please Read the Section Below Carefully is VERY IMPORTANT for your

Constructed Concept Map Final Grade

The following scoring will be used to determine the number of points assigned to each proposition:

1.   An important, accurate, and complete proposition will receive three (3) points.

2.   An important, accurate, but **not** complete proposition will receive two (2) points,

3.   An accurate and complete but **not** important proposition will receive one (0.5) point,

4.   **A wrong** and **important** proposition will receive minus one (-1) point if it is one of the key propositions and is completely inaccurate.

5.   A wrong and **not** important proposition will be ignored.

Propositions will be considered important if they are included on the master map or are judged by the raters to be of equal importance to those included on the master map.

Map overall quality option. In addition to the total points possible based on individual propositions, an additional number (x) of points can be awarded based on the overall quality of the map. A good map should convey holistic understanding. This will be determined by evaluating the map's organization and logical flow between related concepts.

Figure 2 shows examples of how the scoring rubric is applied to specific propositions. The proposition shown in Figure 2a will receive three points because it is important (included on the instructor's master map), accurate, and complete. The proposition in Figure 2b will only receive two points since the proposition is important, accurate **but not** complete. The proposition shown in Figure 2c will only receive 0.5 point because it is an accurate and complete **but not** an important proposition. The proposition shown in Figure 2d will receive minus one (-

1) point because is a **<u>wrong</u>** and **<u>important</u>** proposition. The proposition shown in Figure 2e is a
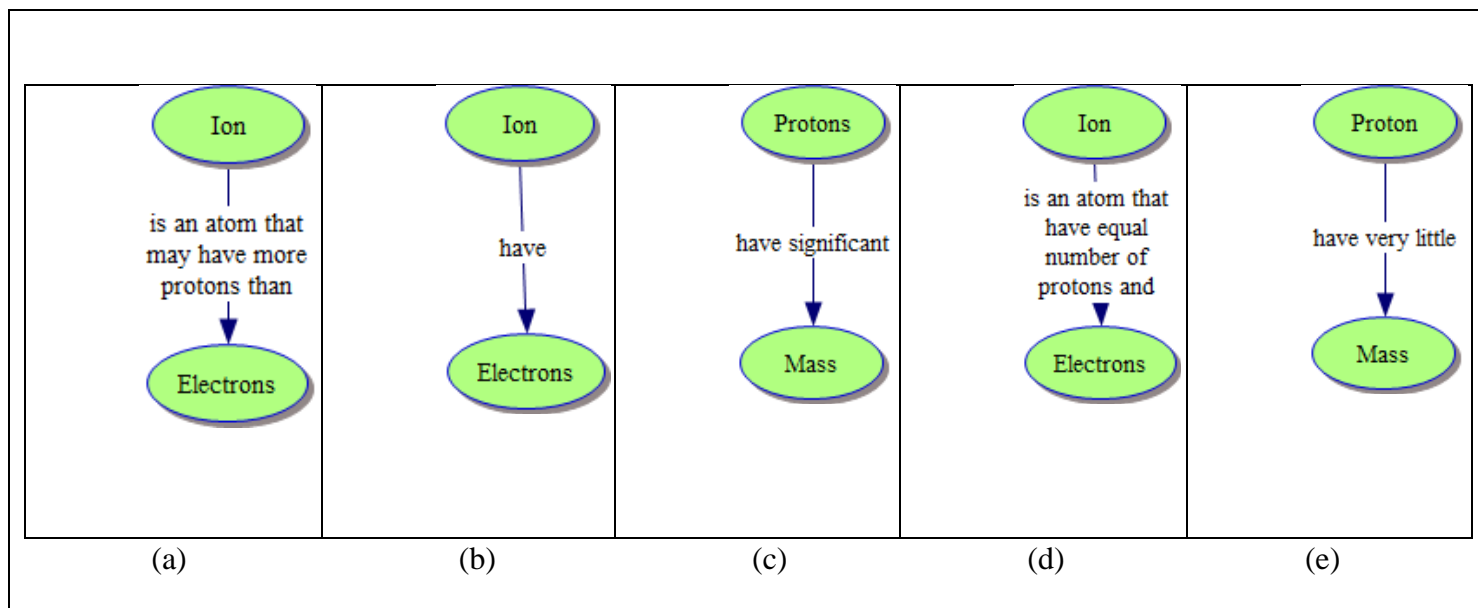
<u>wrong and **not** important proposition</u> therefore it will be ignored.

Figure 2: Applied examples of scoring rubric.