# STUDYING C-TOOLS: AUTOMATED GRADING FOR ONLINE CONCEPT MAPS

DOUGLAS B. LUCKIE[1], SCOTT H. HARRISON[2], JOSHUA L. WALLACE[1] AND DIANE EBERT-MAY[3]

*[1]Lyman Briggs College of Science and Department of Physiology, [2]Department of Microbiology and Molecular Genetics, [3]Department of Plant Biology, Michigan State University*

**Abstract**. The C-TOOLS project has developing a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring. The Concept Connector is designed to enable students in large introductory science classes at the university level to visualize their thinking online and receive immediate formative feedback. The Concept Connector's flexible scoring system, based on tested scoring schemes as well as instructor input, has enabled automatic and immediate online scoring of concept map homework. Criterion concept maps developed by instructors in the C-TOOLS project contain numerous expert-generated or "correct" propositions manually created by expert users connecting two object or subject phrases together with a linking phrase. A range of holistic algorithms as well as WordNet, an electronic lexical database, are being used to test automated methods of scoring. For this study 1298 students created concept maps (with 35404 propositions) were evaluated by automatic grading and for validity of computer generated WordNet® propositions. We studied how successful these approaches were at creating and/or evaluating additional linking words extrapolated from criterion maps generated by experts. By comparing manual assessments of derived propositions to manual assessments of original propositions, the persistence of correctness was evaluated.

Category: Poster Paper

## 1 Introduction

Expert-level thinking depends on a web of mental connections developed over a lifetime of education and experience (Bruner, 1960). Yet, in an attempt to turn college science students into experts, instructors often just focus on passive transmission of large amounts of "content" in a short time period and then test students to see if they "got it" (NRC, 1999). In response, students tend to focus on practical ways to succeed in their courses and thus often adopt strategies like memorization or rote learning (Ausubel, 1963; Novak & Gowin, 1984). Visual models such as concept maps may help instructors teach expert thinking as well as assess domains of student understanding. In our own learning as scientists, we frequently use visual models (Casti, 1990). The value of knowledge scaffolding tools such as concept maps is that they reveal student understanding about the direct relationships and organization among many concepts.

The use of paper and pencil seems to be the most natural way to create concept maps. Students can easily create shapes, words, lines etc and can add small illustrations. As students become more proficient or engaged in making a concept map, problems arise when they'd like to revise it. Erasing can become tedious and inhibit the process of revision. Using "Post-It" notes can allow easy revision, yet a record or copy of the map is not easily generated in the active classroom. An additional challenge is scoring maps. While grading a single concept map may be less time-consuming than grading a long essay or extended response, it is still more complex than grading multiple choice exams. Even if a chemistry instructor would like to use concept maps in their large introductory course of 500 students, they will point out that grading 500 maps is not practical for them. Computer software is an avenue to address these challenges. In fact, a number of projects, like the Inspiration™ commercial software and the freely downloadable, community-oriented IHMC CmapTools software, present excellent replacements for paper-and-pencil drawing environments and may help engage the resistant student.

Although computer-based tools for concept mapping are available to university faculty, few are web-based and none have embedded assessment components for automated scoring and feedback. The C-TOOLS project is to develop and validate a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring and feedback functionality. The Concept Connector is designed to help "make transparent" when students do not understand concepts and motivate students to address these deficiencies. Web-based concept mapping can enable students to save, revisit, reflect upon, share and explore complex problems in a seamless, fluid manner from any internet terminal (Pea et al., 1999). Automated grading and feedback features can allow instructors to use concept mapping on a larger scale.

Automatic grading features associated with the C-TOOLS project's Concept Connector began in 2003 to amplify instructor-designed grading matrices with synonyms from WordNet®[1] (Fellbaum, Ed., 1998). At present,

[1] WordNet is a registered trademark of Princeton University.

for 35404 propositions, 9211 can be evaluated by an automated grading mechanism called Robograder$^{TM}$. WordNet-powered amplification enabled 971 of the 9211 propositions to be evaluated when the existing grading matrices would not otherwise make an assessment. Currently, Robograder$^{TM}$ indiscriminately accepts linking phrase synonyms independent of frequency and word sense.

Visual examinations of automatically graded maps indicated few false positives or false negatives despite Robograder's treatment of multiple synsets as interchangeably equivalent. In theory amplifying grading rubrics by using the superset of all available synonyms should introduce errors into rubrics since multiple and conflicting meanings often exist. One explanation for the observed success of indiscriminate acceptance of synonyms is that users may more likely choose words within a relevant set of synonyms (known in WordNet as a "synset"). Thus, when developing concept maps, users appear to be less inclined to take "stabs in the dark" than compared to a conceptual strategy which favors semantically plausible word choices.

## 2 Methods

### *2.1 Goals and Timeline*

With both the literature providing a solid theoretical basis for using concept maps and the field of computer science providing the proper software development tools and technology, the C-TOOLS project began in late 2002. A team of faculty from Michigan State University spent much of the first year of the project developing both the Java applet, called the Concept Connector (Figure 1), and the classroom problems sets with concept maps for science students. In parallel with software development was a study of how students use the tool. The Concept Connector was developed through a 'design' experiment (Suter & Frechtling, 2000) that involved testing the tool with undergraduate science-majors in biology, geology, physics and chemistry courses.

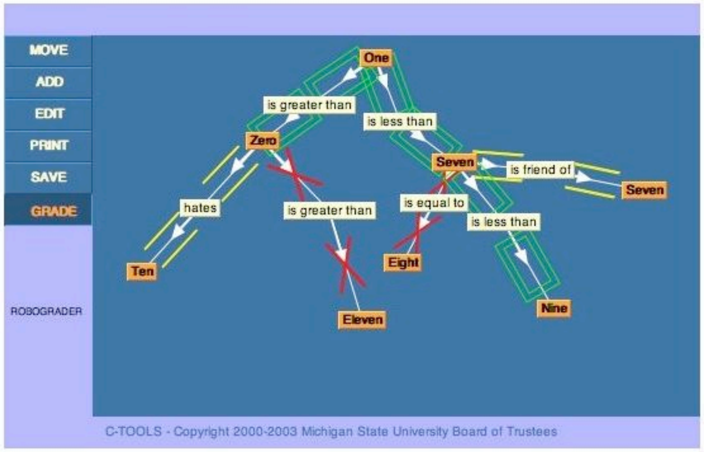

**Figure 1**: The Concept Connector Java applet graphic user interface (GUI). This particular screenshot shows the Java applet's GUI (blue colored areas), how the software draws a concept map, and how new colors (green and yellow rectangular halos or red X's) appear when the *Robograder* is asked to *GRADE* a concept map (http://ctools.msu.edu).

### *2.2 Faculty and Students: Concept Mapping in Large Introductory Courses*

For the C-TOOLS project, we recruited a cohort of over 1000 freshman and sophomore students enrolled in introductory science-major courses: Biology, Chemistry, and Physic, as well as non-major science courses: Introductory Biology and Geology. During class meetings, students learned how to use the web tools. Students completed concept maps as an integral part of the course. Online concept map-based homework assignments varied widely from analysis of scientific literature to answering a particular homework question. To complete an assignment students typically logged into a website and were presented with instructions and a map space seeded with approximately 10 concepts. The Concept Connector software allows students to move concept words around, organize hierarchy, and add linking words and lines. C-TOOLS exercises often challenged students to first construct a map individually, and submit it to the computer to receive visual feedback. They then could revise the map and resubmit. Finally, they often worked with a partner to complete a final collaborative concept map.

### *2.3 Holistic Scoring Approaches: Development and Data Analysis*

The Concept Connector™ has been created as the combination of an online Java applet that serves as a map drawing tool residing in an HTML page that communicates with server-side software on POSIX-conforming systems such as Mac OS X®, LINUX®, and FreeBSD®. The applet is small in size and is browser-compatible on every OS platform and presents a menu-driven, interactive GUI. In terms of architecture, as a technology, a C-TOOLS server incorporated freely available software tools and followed existing software conventions within the freeware community. By implementing and interacting with necessary software components such as cross-linking databases, resource-specific handlers, and servlets in this manner, the C-TOOLS project was careful to utilize open standards.

The online software allows students to seamlessly create their concept map on an "easel" page, save it in a private "gallery," restore, revise and submit it to receive automatic scoring feedback. In automated grading, our primary goal is to follow the scoring system developed by the Novak group (Novak & Gowin, 1984). "Robograder" gives visual feedback concerning the validity of the semantic relationship between linked words in a proposition. During the study period, automated scoring of student linking words graded 26% of the user-made propositions existing on Michigan State University's C-TOOLS server. In addition to the Novak scoring system, we are studying student maps for interesting trends and testing new "Gestalt" approaches for automated feedback that successfully mimic the human grader (Figure 2).

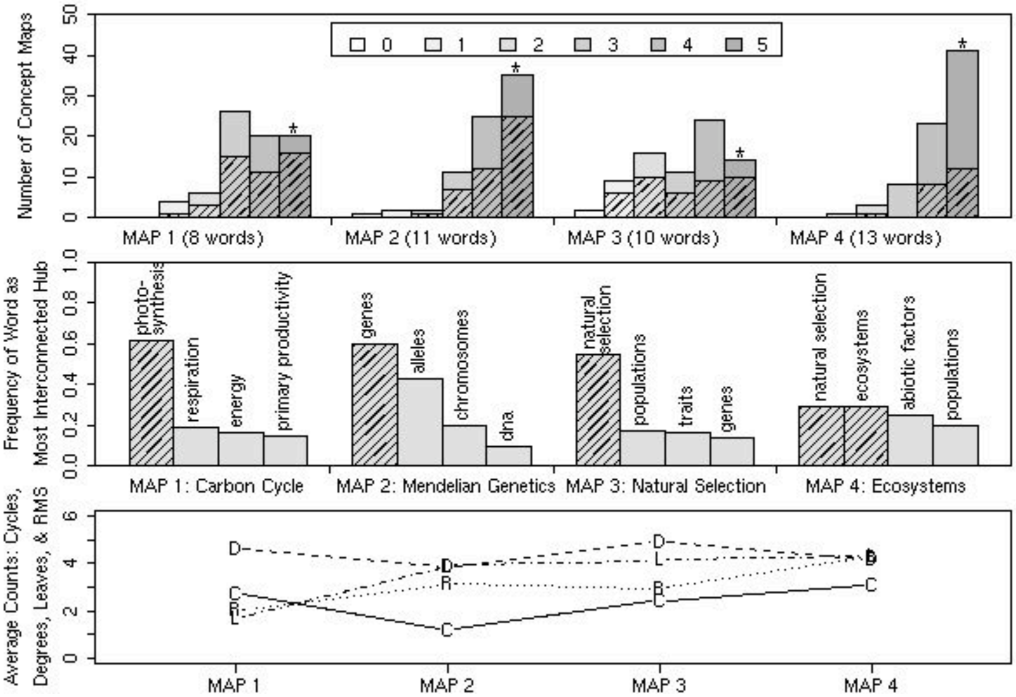

**Figure 2**. Human expert scoring of student maps from a non-majors biology course (top panel) and software analysis of trends in the map data (lower panels). Panel 1 (top) shows the distribution of scores (graded from 0 to 5) for each of 4 assignments given successively throughout a semester (n=76 students). The striped portions of the bars in panel 1 indicate the distribution of scores for maps that used the top "hub" concept word (for MAP1 this hub word was "photosynthesis," identified in Panel 2 (middle)). Panel 3 shows trends in Gestalt scoring approaches applied to the same maps. These are the average values of 4 network topology measurements for the maps that scored a "5" (*) from each of the 4 assignments. Cycles="C", the number of loops involving 3 or more concept words; Degrees="D", the number of propositions connecting to a given concept word; Leaves="L", the number of terminal ends in the concept map network; RMS, an indicator of non-branching chains within a concept map="R", the root of the mean sum of squared distances between all concept word pairs within a concept map.

We are studying more holistic domains such as frequency of word choice and links, network patterns and evaluative approaches based upon the structure of the map. Figure 2 presents an analysis of data from one C-TOOLS biology course. It aligns the distribution of grades (0-5) given by the expert faculty to student concept maps made during a semester (top panel) with an analysis of most common “hub” concept words found in the student maps (“hub”=concept with most links; middle panel) and “Gestalt” grading strategies where software attempts to evaluate the same student maps via content independent approaches (bottom panel).

C-TOOLS provides a well-curated data source with which to assess trends of classroom learning as shown in Figure 2. The instructor predicted the reduced student performance seen for MAP 3 based on complex interdependencies associated with the “Natural Selection” knowledge domain. The instructor also predicted that those students understanding certain critical concept words, as evidenced in MAP 1 by choosing words such as “photosynthesis” to be the most highly interconnected hub, would score the highest on their concept maps. The shift in grade distribution of maps (top panel, striped portions of bars) using the most popular “hub” word (identified in the middle panel) appears to support the prediction.

Automated “Gestalt” grading approaches currently being tested are based on the network structure of the student concept maps. Methodologies using map network patterns related to hierarchy (“Leaves” and “Degrees”), cross-linking (“Cycles” and “RMS”), as well as the use of software called WordNet® to amplify linking word databases are being studied (Harrison, Wallace, Ebert-May, & Luckie, 2004). In the bottom panel of Figure 2, four automated scoring strategies were tested on student concept maps that received a score of 5. Interestingly, topology measurements termed “RMS” and “Leaves” correlated best with the human grader. The capacity to analyze and verify these predictions will grow in power with the accumulation of additional data and classroom-to-classroom comparisons.

### *2.4 WordNet Scoring Approaches: Testing and Data Analysis*

At the time of this WordNet study there were 35404 propositions available from Michigan State University's C-TOOLS server. A random sample of 250 propositions was gathered and divided into 5 separate sets of 50 each. Each of these original propositions consisted of a starting concept phrase, a linking phrase, and a terminal concept phrase. Manual assessment of propositions was done by hand without aid of electronic references or algorithms. Manual assessments of the 5 sets were performed by the first two authors. Scorings for each proposition were: 1 (correct, e.g. “Photosynthesis - needs - Carbon dioxide”), X (incorrect, e.g. “DNA - transcribes – RNA”), 0 (ambiguous, e.g. “atom - is made of – neutron”), and S (structural violation, e.g. “ocans - evaporation – atmosphere”). Structural violations were for propositions with grammar problems such as spelling errors and linking phrases that do not contain a verb. Ambiguous scores were given to propositions that could only be scored as correct when viewed in a reasonably plausible context of surrounding propositions.

Version 2.0 of the software database WordNet was used to generate "proposition derivatives" by making linking phrase substitutions based on WordNet's thesaurus-like lexical capabilities. There were two criteria for the generation of proposition derivatives. First, derived propositions were made from linking phrases consisting of a single verb. Only 121 of the 250 original propositions met this single verb word criterion. Second, at minimum, the WordNet database had to have three available choices per lexical relationship. An original proposition’s linking verb could thus have up to nine derivatives (i.e. 3 antonyms, 3 troponyms, and 3 synonyms). Each triplet, as generated per lexical relationship (e.g. three antonyms), is called a trio. Trios provide an initial comparative range of data concerning the sense of usage and polysemy counts specific to both lexically similar and dissimilar derivations made from original propositions. With the single verb and triplet criteria, WordNet enabled us to construct 30 antonym derivatives, 243 troponym derivatives, and 234 synonym derivatives per proposition. Grading of proposition derivatives was delegated by the originating proposition sets. Graders A and B both graded derivative set 5. Grader A graded derivative sets 3 and 4. Grader B graded derivative sets 1 and 2.

The manual assessments of original and derivative propositions were scrutinized in order to both summarize and make insights into relationships that may concern automated strategies of assessment. Assessments of original and derived propositions were enumerated in order to show relative ratios of correctness, ambiguity, and grammatical errors. Trios were analyzed for fluctuations in correctness and incorrectness. Trends between correctness and polysemy were investigated.

Graders A and B had reproducible similarity to their scoring patterns as determined by the Kappa statistic (Cohen, 1960). The Kappa statistic (_ = 0.552) was calculated with $p_o = 0.720$ and $p_e = 0.374$ suggesting good reproducibility (0.4 ≤ _ ≤ 0.75). The level of significance for this degree of association is < 0.10. For the manual assessments of the 250 original propositions, 72% of the assessments between the two graders were identical (180 propositions). Opposite assessments of correctness (1 versus X) occurred 5.6% of the time. Remaining differences for the assessment of individual propositions were primarily attributable to issues unrelated to exacting qualifications of correctness. For example, 30 instances of disagreement involved only one grader assigning an S score and 26 instances of disagreement involved one grader cautiously assigning a 0 (ambiguous) score in contrast to 1 or X scorings. While our approach has statistically significant repeatability for scoring ratio properties and strong consistency for exacting qualifications of proposition correctness, further refinement would involve better synchronization between graders' approaches to assumptions of context and handling of grammatical logistics. When looking at the jointly graded WordNet derivative set (*n* = 144), the agreement between grader A and grader B was 70% ($p_o$ = 0.701). The degree of association is just marginally reproducible based on _ = 0.375 and this reduction may be attributable to fewer shared contextual assumptions between graders due to loss of the original word choice. Scoring dynamics appear to be conserved; joint scorings for derivatives rise in agreement when considering just 1 and X scores, and the _ value does not suggest complete insignificance (_ = 0.13).

Antonym derivatives were found to be always incorrect. Of 21 antonyms graded by grader A, 21 were graded as incorrect. Of the 18 antonyms graded by grader B, 18 were graded as incorrect. Assessments for original, synonym-derived, and troponym-derived propositions are shown in Table 1 and encompass a range of assessment across all four grading categories (1, 0, X, and S). When the range of assessment is limited to 1 and X, grader A found 25.6% of synonym-derived propositions to be correct and 16.8% of troponym-derived propositions to be correct. For 1 and X scorings, grader B found 43.5% of synonym-derived propositions to be correct and 31.7% of troponym-derived propositions to be correct.

| Score | Original propositions | | Synonym-derived propositions | | Troponym-derived propositions | |
|---|---|---|---|---|---|---|
| | Grader A | Grader B | Grader A | Grader B | Grader A | Grader B |
| Correct | 141 | 123 | 32 | 68 | 21 | 53 |
| Incorrect | 12 | 32 | 93 | 88 | 104 | 114 |
| Ambiguous | 33 | 13 | 16 | 0 | 22 | 0 |
| Structural violation | 64 | 82 | 0 | 0 | 0 | 1 |

**Table 1**: Summary of manual assessment scores for original, synonym-derived and troponym-derived propositions.

The construction of trios involves random sampling from each WordNet-generated set of antonyms, synonyms, and troponyms. If conflicting meanings inside each set cause a general variation of proposition correctness, then clustering of correct or incorrect assessments within trios should not differ from a distribution of correct assessments that is random with respect to triplet structure. For the 57 synonym-derived trios assessed by grader A and the 81 synonym-derived trios assessed by grader B, the distributions showed no significant difference (_$^2$ = 1.59, *p* = 0.66 and _$^2$ = 2.07, *p* = 0.56 respectively). For the 54 troponym-derived trios assessed by grader A and the 71 troponym-derived trios assessed by grader B, the distributions also showed no significant difference (_$^2$ = 2.64, *p* = 0.45 and _$^2$ = 5.57, *p* = 0.13 respectively).

The general variability of correctness occurring within trios was investigated further by measuring how assessment score changes relate to similarities in meaning for derived proposition linking verbs. The WordNet database organizes lexical sets into subsets (termed "synsets") grouped together by similar meaning. Pairs of propositions occurring within trios were analyzed for having dissimilar correctness scores 1 and X, and for whether each proposition's linking verb was a member of the same synset. Shared synset membership for troponym derivatives occurred for 67% (grader A) and 49% (grader B) of all trio pairings that had an assessment score transition from 1 to X. Scoring transitions from 1 to X were next contrasted to within-trio proposition pairs where both propositions were assessed with a score of 1. Shared synset membership for troponym derivative pairs occurred for 100% (grader A) and 81% (grader B) of all such trio pairings that had a common assessment score of 1. Synonym derivatives were analyzed in similar fashion. Shared synset membership for synonym derivatives occurred for 15% (grader A) and 14% (grader B) of all trio pairings that had an assessment transition from 1 to X. Shared synset membership for synonym derivative pairs occurred for 27% (grader A) and 31% (grader B) of all such trio pairings that had a common assessment score of 1. Thus, for both troponyms and synonyms, membership of two verbs in the same synset implicates retained assessments of correctness.

Correctness was then analyzed for its impact on polysemy count distributions. For original propositions, polysemy distribution values were _ = 3.43, _ = 7.72 and _ = 4.08, _ = 6.46 for incorrect and correct propositions respectively. For derived propositions, polysemy distribution values were _ = 6.01, _ = 7.51 and _ = 8.90, _ = 11.40 for incorrect and correct propositions respectively. The increase of polysemy counts with correctness was attributed both to moderately high polysemy count ranges (>20) corresponding to the correctness of propositions by a factor of 2.4 and to a distinct trend for low polysemy count ranges (<5) corresponding to a 10% rise in the incorrectness of propositions.

## 3 Summary

The C-TOOLS project stems from the combined activities of an interdisciplinary team of faculty from Michigan State University. This National Science Foundation (NSF)-funded project developed a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring and feedback functionality. The Concept Connector tool is designed to enable students in large introductory science classes to visualize their thinking online and receive immediate formative feedback. Further details concerning the C-TOOLS project have been previously published (Luckie, Batzli, Harrison & Ebert-May, 2003).

In this study of Holistic and WordNet automated scoring approaches of concept maps, an approach that amplifies correctness across multiple synsets appeared to work on concept maps made by real users. Such an indiscriminating approach was faulty when applied to randomly generated sets of synonyms and troponyms. Thus, the data supports that synonyms and troponyms can be used as sets for further identifying both correct and incorrect propositions. Although it may appear that there is only a 10% gain by using synonyms for automatic grading, this is only from the standpoint of automating the assessment at the proposition level. At the larger concept map level, there are highly interconnected concept words that follow a pattern of classroom consensus and also correspond to student performance (Luckie, Harrison, & Ebert-May, 2004). Better understanding of the linking words around major hubs would aid us to analyze the formative dynamics of how users in a classroom interconnect concepts and, potentially, knowledge domains. Analysis and further improvements to Robograder$^{TM}$ cannot just be limited to synset hierarchies of each individual linking word since there are content-dependent dynamics of semantic overlap that influence how words can sensibly connect to other words (Banerjee & Pedersen, 2003).

## References

Ausubel, D. (1963). *The Psychology of Meaningful Verbal Learning*. Grune/Stratton. New York, NY.

Banerjee, S., & Pedersen, T. (2003) *Extended Gloss Overlaps as a Measure of Semantic Relatedness*. Paper presented at IJCAI 2003 – 18th International Joint Conference on Artificial Intelligence.

Bruner, J. (1960). *The Process of Education*. Harvard University Press. Cambridge, MA.

Cañas, A. J., Valerio, A., Lalinde-Pulido, J., Carvalho, M., & Arguedas, M. (2003). *Using WordNet for Word Sense Disambiguation to Support Concept Map Construction*. Paper presented at SPIRE 2003 – 10th International Symposium on String Processing and Information Retrieval.

Casti, J. L. (1990). *Searching for certainty: what scientists can know*. New York, W. Morrow, 496 p.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Collins, A., Joseph, D. & Bielaczyc, K. (2004) Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13*(1), 15–42.

Fellbaum, C. Ed. (1998). *WordNet – An Electronic Lexical Database,* MA: MIT Press.

Fisher, K. M. (2000). *SemNet software as an assessment tool*. In J.J. Mintzes, et al (eds.), Assessing science understanding: A human constructivist view. Academic Press. San Diego, CA.

Harrison, S. H., Wallace, J. L., Ebert-May, D., & Luckie, D. B. (2004). C-TOOLS automated grading for online concept maps works well with a little help from "WordNet" Paper presented at CMC 2004 – 1st International Conference on Concept Mapping.

Ihaka, R., & Gentleman R. (1996). R: A Language for Data Analysis and Graphics*, Journal of Computational and Graphical Statistics, 5*, 299-314.

Luckie, D. B., Batzli, J. M., Harrison, S., & Ebert-May, D. (2003). C-TOOLS: Concept-Connector Tools for Online Learning in Science. *International Journal of Learning 10*: 332-338.

Luckie, D., Harrison, S., & Ebert-May, D. (2004). *Introduction to C-TOOLS: Concept Mapping Tools for Online Learning*. Paper in review for CMC 2004 – 1st International Conference on Concept Mapping.

National Research Council. (1999). *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology*. National Academy Press. Washington, DC

Novak, J. (1990). Concept Mapping: A Useful Tool for Science Education. *Journal of Research in Science Teaching, 27*(10), 937-949.

Novak, J. D., & Gowin., D. D. (1984). *Learning How to Learn*. Cambridge Press. New York, NY.

Pea, R., Tinker, R., Linn, M., Means, B., Bransford, J., Roschelle, J., His, S., Brophy, S., & Songer, N. (1999). Toward a learning technologies knowledge network. *ETR&D*. *47*(2): 19-38.

Suter, L., & Frechtling, J. (2000). Guiding principles for mathematics and science education research methods. *NSF Report* 00-113.

Wittrock, M. C. (1992). Generative Learning Processes of the Brain. *Educational Psychologist, 27*(4), 531-541.