Assessment and Concept Map Structure:
The Interaction Between Subscores and Well-Formed Mental Models[1]

Debbie Denise Reese
Center for Educational Technologies®
Wheeling Jesuit University
Wheeling, WV 26003
debbie@cet.edu (304) 243-4327

Abstract

The concept map is a visual tool, a nontraditional assessment tool designed to specify individual mental models through graphic representation. This study investigated relative sensitivity of concept map subscales to well or poorly defined mental models of a targeted domain (learner achievement). Fifty-eight preservice teachers completed a computer-based lesson introducing the educational philosophy concept of functionalism within a traditional society. Analysis of concept maps constructed from learners who completed responses to four essay questions indicated a significant interaction between Subscore and Achievement, $F(3,51)=33.42$, $p< .01$, $\eta^2= .66$. The Cross-links subscale, measuring a learner's domain integration, is more sensitive than Branches (domain differentiation), Propositions (domain knowledge), or Levels (domain differentiation). Implications for research and practice are discussed.

As expert advisory bodies like the United States National Research Council urge policy makers, scientists, and educators to recognize the limits of traditional assessment instruments (2001), concept maps provide an attractive alternative. A concept map is a graphic representation of perceptions of a domain's relational structure. Each concept map is constructed to visually characterize understanding of the relationships between a targeted domain's concepts and its subconcepts. Thus, the concept map is inherently applicable to the measurement of higher-order conceptual understanding that characterizes meaningful learning (Ausubel, 1962; Novak & Gowin, 1984; Wiggins & McTighe, 2001). In its broadest application, alternative assessment initiates a process through which instructional goals are explicitly and holistically reinforced within a learning community (Hernandez & Reese, 2004; Wiggins & McTighe, 2001). This requires ongoing feedback and learner growth within an environment of publicly advertised standards and criteria (Hernandez & Reese, 2004; Wiggins & McTighe, 2001). Because a concept map is an explicit representation, it provides an object that can be objectively shared, digested, and discussed. As concrete objects of shared discourse, concept maps can be used as formative assessment tools to shape a learning community's progress toward instructional goals that require higher-order understanding. Since Joseph D. Novak and his colleagues (Novak, 1990, 1992; Novak & Gowin, 1984; Novak & Musonda, 1991) developed, pioneered, and validated the use of the concept map, educators and researchers have added it to the more traditional array of pedagogical and assessment tools. Assessment and evaluation of a concept map have involved quantification of three characteristics: (a) amount of knowledge, (b) domain differentiation, and (c) domain integration. These parameters are normally measured through subscale scores. Unfortunately, evaluators typically develop idiosyncratic sets of subscales and subscale weightings for each concept map application (Edmondson, 2000; Markham & Mintzes, 1994; Novak, 1990; Novak & Gowin, 1984; Wallace & Mintzes, 1990). This practice curtails generalization across domains and populations, limiting the contribution of concept mapping results within curricular and research efforts (Edmondson, 2000; Reese, 2003a; Shavelson & Ruiz-Primo, 2000). In addition to the generalization issues, ill-conceived scoring rubrics may unintentionally reinforce superficial rather than deeply structured knowledge organization.

The National Research Council (2001) has suggested that three key elements should underlie any assessment: (a) cognition: how students learn and organize a targeted domain, (b) observation: how evaluators collect evidence of student learning in a domain, and (c) interpretation: how evaluators conduct an analysis and explanation of those observations. One implication of this "assessment triangle" is that an expert-novice paradigm comparing a learner's concept map to the expert's map of a targeted domain might more closely align with the triangle's cognition vertex than a summation of subscales measuring a concept map's open-ended richness. Comparisons between expert and novice concept maps are a relatively new area of research, requiring an algorithm and computer program to quantify the maps (Fisher, 2000; Reese, October, 2003; Clariana, Koul, & Salehi, in press; Poindexter & Clariana, in press). More commonly, researchers and educators follow Novak's paradigm and score concept maps by the sum of

[1] Paper presented at the 2004 meeting of the American Educational Research Association, San Diego.

weighted subscale scores (Markham & Mintzes, 1994; Novak & Gowin, 1984; Novak & Musonda, 1991; Wallace & Mintzes, 1990). Therefore, a pragmatic consideration of implementation practice, pedagogical consequences, and methodological concerns suggests a research agenda that investigates the structure of the concept map through identification of subscale sensitivity toward differences in academic achievement. In other words, are some subscales more sensitive than others? Which subscales, if any, are the most sensitive indicators of learner achievement when learner achievement is defined as the richness of a learner's mental model of a targeted domain (e.g., the sum of weighted subscores)?

Theoretical Roots of Concept Map Structure

The concept map tool derived from David P. Ausubel's theory of reception learning (Ausubel, 1962, 1963, 1968, 2000). That theory is now referred to as assimilation theory (Edmondson, 2000; Novak, 1992). Whether modern scholars cite Ausubel's theory or not, they often construct representations of structure knowledge systems such that superordinate concepts subsume subordinate concepts (e.g., Anderson & Lebiere, 1998; Gentner, 1983, 1989; Gentner & Gentner, 1983; Gentner & Markman, 1997; Lakoff, 1987; Lakoff & Johnson, 1980, 1999; Mayer, 1979). In addition to the hierarchical structure that develops through subsumption, assimilation theory also identifies the processes of (a) progressive differentiation (differentiation)– when a learner separates subordinate concepts into conceptual strands and (b) integrative reconciliation (integration)– when a learner makes connections between disparate conceptual strands (Ausubel, 1962, 1963, 2000). Research supports the theory, demonstrating that individuals holding well-formed mental models of targeted domains structure those domains through subsumption, differentiation, and integration. Studies using the novice/expert paradigm have found that experts organize domain knowledge "hierarchically, under fundamental concepts" (Edmondson, 2000, p. 17), and that successful science learners as well as professional scientists develop domain-specific knowledge structures (mental models) that differentiate and integrate subsumed concepts (Edmondson, 2000). Concept maps were designed to allow one to model domain knowledge according to hierarchical structure, differentiation, and integration (Novak & Gowin, 1984). Concept maps are composed of nodes (concepts represented by labels enclosed within ovals) and the arcs that connect them (relations represented by directional lines, often labeled). A proposition is a pair of concepts connected by a relationship. Evaluators have summed the number of concepts within a concept map and/or summed the number of propositions within a concept map to indicate amount of domain knowledge. Evaluators typically devise a scoring rubric to score a concept map. Although scoring rubrics are idiosyncratic, they are usually based upon the sum of weighted subscores– some combination of domain knowledge (weighted number of Concepts and/or Propositions), hierarchy (weighted number of Levels within a map), differentiation (weighted number of conceptual Branches), and integration (weighted number of instances in which two branches merge at a concept, creating a Cross-link).

Meaningful Learning and Well-Formed Mental Models

Ausubel (1962) was careful to distinguish meaningful learning from rote learning. According to Ausubel, "rote learning is arbitrary and verbatim. Meaningful learning is nonarbitrary and nonverbatim" (Reese, 2003a, p. 67). Meaningful learning is "human understanding of a domain" (Gentner & Stevens, 1983, p. 1). Meaningful learning occurs when an individual assimilates or accommodates new information within existing cognitive structure. While the structure and interpretation of an individual's prior knowledge may be shaped by sociocultural history, expectations, and interactions and while sociocultural aspects may be shared within a community of practice (Reese & Hergert, 2004; Wenger, 1998), the unit of analysis here is the individual.

Domain experts have organized their knowledge of that domain. Cognitively, they have categorized and chunked (Miller, 1956) concepts and procedures, facilitating retrieval, application, and problemsolving. An expert's collection of concepts, procedures, rules, and principles and the relations that interconnect them is that expert's mental model of the domain. An expert's model is wellformed because it is a richly differentiated and integrated relational network "characterized by the relational structure of the domain that affords inferences and elaborations" (Reese, 2003a, p. 65). By contrast, the mental models built by novices are sparsely differentiated and integrated. Where connections do exist between subconcepts, they are often based upon superficial attributes rather than trans-concept relations (Gentner & Markman, 1997). They are often contain nonnormative or naïve understandings that inhibit learning because they contribute to misunderstandings and/or misapplication.

Map Construction

When concept maps are used as assessment instruments, map construction can be characterized by three continuums (Shavelson & Ruiz-Primo, 2000): (a) The constructor of the concept map may be the learner whose knowledge is being mapped or a trained rater who interprets the learner's protocol responses, (b) Learners may be provided with a set of domain concepts or asked to generate their own set, and (c) Learners may construct the map from scratch or complete blanks within a pre-fabricated structure. When domain novices construct their own maps, graders can find nonnormative understandings revealed as inappropriate relational links between concept dyads. These develop into inappropriately specified levels, branches, and cross-links. When trained graders construct the maps from participant narratives (either oral or written), transcription procedures can either map all relations as specified or ignore the nonnormative dyads.

Evaluating the First Concept Maps

Novak and his colleagues (Novak, 1990; Novak & Gowin, 1984; Novak & Musonda, 1991) developed concept mapping during a 12-year, longitudinal study, refining it as an empirical method used to track changes in learners' mental models of science domains. Learners were first tested in first or second grade, when the treatment group received aural instruction in specified science concepts. Over the progression of the students' maturation from elementary school to high school graduation, trained research assistants conducted Piagetian-like interviews with each learner. The researchers taped and transcribed the sessions. Then researchers charted each learner's domain knowledge, using concept maps. Each map represented the consensus of two or three researchers' interpretations of the structure of one learner's domain knowledge at the time of the interview. The researchers organized a learner's concepts hierarchically, subsuming concepts according to the relevant relationships indicated by a learner's comments. Integration of learner's domain knowledge was measured by (a) the number of levels of the hierarchy (How inclusive is the most subsumptive node?), (b) progressive differentiation (What are the relationships between one concept and another?), and (c) integrative reconciliation (What are the cross-relationships between one branch of the hierarchy and another?).

While Novak (Novak & Gowin, 1984) proposed a framework for a concept map scoring scale, he admitted that any scoring scale, including any one particular scale developed for concept maps, must confess "a certain degree of subjectivity" (p. 105) and/or arbitrariness. However, Novak argued, the arbitrariness or subjectivity within a particular experiment or educational environment would not bias the results unless that research should indicate a bias toward a particular set of learner characteristics.

Novak (Novak & Gowin, 1984) proposed four scoring components:

1. Score all valid relationships (connections from one node to another). Rationale: We can get an index of the level of a person's knowledge in a domain by counting the number of valid propositions (Novak & Musonda, 1991, p. 129).
2. Count the number of valid hierarchical levels. Multiply them by a predetermined weighting factor derived from the number of concepts available within the domain. The number is determined by the learner's level of development and the instructional content. Rationale: "More levels in an individual's concept may signal greater differentiation of concept meaning–and more sophistication" (Novak & Musonda, 1991, p. 129).
3. Score valid cross-links at 3 times the value of the hierarchical level. Rationale: Interlinkages (when correct) among key concepts in different segments of a concept map are taken as evidence of some degree of integrative reconciliation and hence scored higher than the inclusion of more examples of simple superordinate to subordinate concept relations achieved through simpler subsumptive assimilation processes (Novak & Musonda, 1991, p. 127).
4. Score each valid example.

In accord with Novak's (Novak & Gowin, 1984) own advice, concept map scoring systems have remained flexible. Indeed, they may be developed ad hoc for each instructional or experimental context. Novak himself has published alternative scoring rubrics (1990, p. 128). Wallace and Mintzes (1990) added branching (the number of vertical conceptual threads) to their scoring categories. Markham, Mintzes, and Jones (Markham, Mintzes, & Jones, 1994) also used branching as a scoring category, because "branching in a map represents progressive differentiation of domain knowledge" (p. 94). Subscale weightings were idiosyncratic in each case.

Subscale Discrimination Sensitivity

Novak's (Novak & Gowin, 1984) original weighting system identified the cross-link as the scale the most sensitive to discrimination of a well-formed mental model, weighting it 3 times the weighting of the hierarchical level weighting, which was already weighted more strongly than the number of propositional dyads. Is there evidence to support this weighting? Does analysis of concept map scorings support this weighting system or any other? Is there an interaction between concept map subscales and richness of a learner's mental model? If a sample of learners were trained in a targeted domain and their mental models of that domain were represented as concept maps, would the cross-link subscale be the most sensitive discriminator of those learners with the richest mental models? In other words, will the half of a sample with the richest normative models demonstrate significantly greater reconciliative integration (number of Cross-links) than those with naïve or undeveloped mental models? Moreover, can there be a significant difference between the groups' Cross-links and no significant difference in progressive differentiation (number of hierarchical Levels and number of Branches) and domain knowledge (number of Propositional dyads)?

Method

This study was part of a larger investigation of the effect of a metaphorically enhanced interface on the learner's mental model of a targeted domain (Reese, 2003a, 2003b, October, 2003). Fifty-eight study participants self-selected from three classes of fourth and fifth year pre-service teachers. Participation in the study was optional. Students participated either to fulfill a course research requirement or as an enrichment activity. All participants completed the session during a three-hour block conducted in place of a regularly scheduled class meeting. All but two of the participants were females. Ages ran from 20 to 21 years. One 25-year-old had participated in part of the study but elected not to complete the assessment activities.

Participants were identified by only a participant number. As part of the larger study, participants were randomly assigned to work through computer-mediated instruction presented using a metaphor-enhanced interface (treatment) or a non-metaphor-enhanced interface (control). The study was a double blind, as both participants and raters were unaware of assignment condition. Other than the interface distinctions, the instruction was identical. The targeted domain was functionalism, a component of educational philosophy curriculum requirements for students in many preservice and graduate education programs. The sequencing (navigation) and content of instruction had been determined by an expert model derived from the discipline of educational philosophy. The domain expert model was specified as a concept map (see Figure 1). The expert model derived from Walter Feinberg's introductory educational philosophy textbook, *School and Society* (1998), and the concept map was reviewed and approved by Feinberg and two additional domain experts (Dr. Megan Boler and Dr. James Garrison). Participants proceeded through the computer-mediated instruction individually and at their own pace. The instruction required participants to (a) navigate through the title, (b) read information, and (c) write either a definition paraphrase (the control group) or the details of a metaphoric mapping (treatment group). This learning task was designed to encourage learners to actively engage with each domain concept as it is presented. The task was modeled after techniques developed by Gentner and her colleagues (Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 1999; Thompson, Gentner, & Loewenstein, 2000).

After they completed the multimedia instructional activities, participants completed an assessment activity and a cognitive abilities test.

Instruments and Protocols: Assessment

Each participant was provided with (a) a test booklet referenced to the participant's treatment assignment and (b) a list of all domain concepts presented by the instruction. The computer software presented participants with four probe questions, one at a time. While a probe was displayed on the computer screen, a learner had five minutes to write a narrative in response to the probe questions. Participants wrote their responses by hand in the test booklets provided. Probes were presented consecutively.

Two types of long-answer, constructed-response probe protocols were developed for the assessing learners' perception of the targeted domain concepts. Both asked a learner to apply knowledge of the targeted concept within a new context:

1. *Conflicting Model*[2]*:* The targeted domain concept map is key to constructing a conflicting model probe. The researcher examines the concept map, selects one to a few concepts from the map, and creates a hypothetical situation in which those concepts or the relationships between them are changed. For example, to test a learner's mental model of a functionalism-traditional society domain, a probe might present a hypothetical civilization in which a traditional society employs *formal* education (as opposed to the *informal* education concept that is an important relationally integrated concept within the targeted domain). The probe would describe the hypothetical society and ask the learner to detail the ways in which the hypothetical society agreed or disagreed with what the learner had learned about the functionalist's perspective of a traditional society. This type of probe is presenting a parallel but conflicting model. The learner's job is to align that model with the one presented during instruction and discriminate those differences.
2. *Position Analysis:* Position analysis begins with an expert's position statement (about a paragraph in length) containing components that align and/or disagree with the targeted concept. Once again, the designer selects critical components from the targeted domain's concept map. The task is similar to that used for the conflicting model

[2] The conflicting model format derived directly from Dedre Gentner's suggestions during a conversation at the 2000 annual meeting of the American Educational Research Association.

probe. The learner must identify and explain how components within the position statement align or disagree with the targeted domain. For example, a statement by John Dewey expressing an opinion about the relationship between school and society could be used as a position analysis probe to assess learners' mental models of a traditional society.

Again, students were allowed five minutes to respond to each probe question. After the time had elapsed, the computer software replaced the first question probe by the next probe for five minutes, and then the next, until the four probes had been delivered. In order to tap semantic memory and domain integration as opposed to rote recall of labels, I had provided participants with a list of the 13 targeted domain concepts. Both session proctor directions and computer software directions prompted participants to use this list as they prepared their responses to the four assessment prompts.

Instruments and Protocols: Cognitive Abilities Instrument

During data analysis participants would be identified as low or high achievers based on their final concept map scores across the four assessment probes. It was possible that individuals scoring in the upper half of the sample might possess cognitive abilities that differentiated them from individuals scoring at the lower half of the sample. The private company, PsychTests.com, offers academic and private sector business clients online administration and scoring of instruments measuring participants' cognitive abilities and psychological traits. The cognitive abilities test (Jerabek, 2000) measures seven subscales (Pattern Recognition, Classification, Analogies, Arithmetic, General Knowledge, and Logic; see Table 1 for description of subscales as well as psychometrics). Jerabek and her associates calculated instrument psychometrics using a randomly selected sample of 351,595 persons from the pool of nearly 840,000 participants who have completed the online inventory. Jerabek reports a split-half reliability correlation between forms of 0.7387. It is 0.8497 using Spearman-Brown formula and 0.8407 using Guttman's formula. Inter-item consistency, measured using Cronbach's Coefficient Alpha, is 0.8756.

Within the experimental session of my study, all participants completed the PsychTests.com cognitive abilities instrument after completing the assessment probes. Each participant's responses to the test items were submitted online. PsychTests.com scored responses and immediately reported results on each subscale and the composite IQ score to participants. Upon my request, PsychTests.com e-mailed me a database containing all participant subscores and final scores. Again, participants were identified by participant ID numbers.

Scoring Protocols

Two raters completed eight hours of training to prepare them with adequate knowledge of the domain and scoring procedures. At the conclusion of training, there was 100 percent interrater reliability in scoring a participant protocol. Raters' scoring was also in 100 percent agreement with the scoring key. The rater dyad met for an additional session to review map consolidation procedures for merging the four participant concept maps into one overall map (Reese, 2003a) and the procedure for entering the rater dyad's consensus concept map into the concept mapping software program (Inspiration, available through http://www.inspiration.com).

Although the raters were familiar with the study's premise, they were blind to participants' random assignment and identity. Each rater had access to a copy of the study's rater notebook. The notebook contained reference materials pertinent to the study (e.g., domain concepts, domain expert concept map, the rater dyad's designated list of terms equivalent to targeted concepts). Raters consulted notebooks for clarification if needed. Raters independently identified, highlighted, and labeled (using concept numbers) target concepts present in each protocol response within a participant's response set and created a concept map for each participant's responses. Raters parsed protocols for only the 13 targeted domain concepts (see Figure 1, concepts are enclosed in rectangles). Raters also coded terms and phrases as targeted concepts if those terms and phrases were semantically equivalent to a targeted concept. Because the larger study investigated the effectiveness of an interface metaphor (a concrete analog to the targeted abstract domain), metaphorical concept analogs presented within the instruction were also coded as equivalent to their abstract correlates. Raters ignored (a) incorrectly represented concepts, (b) concepts that were simply listed without meaningful connections, and (c) all other content, except to use it to identify a relational connection among identified concepts. The rater dyads came to consensus in construction of the final concept map for each participant, the one that merged all of the relationships specified within each participant's four individual maps. Finally, the raters entered that final concept map into a concept mapping software program.

1. A predetermined scoring rubric, based upon previous concept mapping rubrics in the literature (Markham & Mintzes, 1994; Novak, 1990; Novak & Gowin, 1984; Wallace & Mintzes, 1990) and the structure of the

TABLE 1

Classical IQ Test – Second Revision, Subscores and Psychometrics

Description
A 60-item test assessing the Classical Intelligence. Low scores indicate low general IQ; high scores indicate high general IQ. The test yields one main score (overall IQ score adjusted for age and gender) and six subscores. The general IQ score is normalized and adjusted for age and gender. The mean of the general score is 100; the standard deviation is 15. The subscores are standardized and converted to a scale from 0 to 100.

The subscores are:

1. Pattern Recognition: measures the ability to make out patterns in a series of images, numbers, word, or ideas. The higher the pattern recognition score, the higher the classical intelligence.

2. Classification: measures the ability to discover the commonalities among collections of words, pictures, objects, etc., and the ability to organize them accordingly. The higher the classification score, the higher the classical intelligence.

3. Analogies: measures the ability to find the relationships between elements of things (words, numbers, images, etc.) otherwise unlike. The higher the analogies score, the higher the classical intelligence.

4. Arithmetic: measures a branch of mathematics that generally deals with the nonnegative real numbers and with the application of the operations of addition, subtraction, multiplication, and division to them. The higher the arithmetic score, the higher the classical intelligence.

5. General Knowledge: measures awareness of things that are commonly wellknown. The higher the general knowledge score, the higher the classical intelligence.

6. Logic: measures the ability to make deductions that lead rationally to a certain probability or conclusion

Sample

Sample size: 351,595
Sample description: The sample used in this study was randomly selected from a pool of nearly 840,000 participants. It includes men and women, aged 10 to 80.

Reliability and Internal Consistency

Split-Half Reliability: Correlation between forms (0.7387), Spearman-Brown formula (0.8497), Guttman's formula (0.8407)
Inter-Item Consistency: Cronbach's Coefficient Alpha (0.8756)

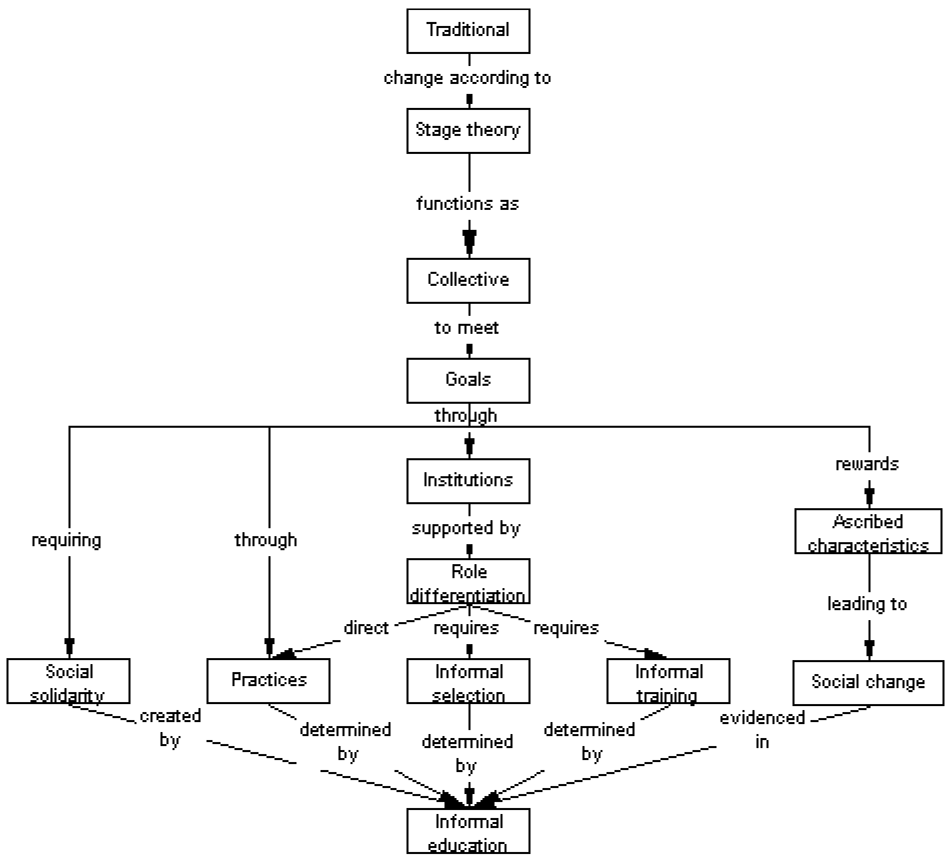

FIGURE 1 Expert Model of Targeted Domain, Specified as a Concept Map.

source domain concepts was developed to score the concept maps prepared by the raters. Initially, the concept map components were weighted and then summed:

1. Concept node = 1 point each
2. Hierarchy levels = 1 point times the number of levels
3. Branches = 2 points times the number of branches at the horizontal level of the map at which the number of branches is the greatest
4. Cross-links = 4 points times the number of cross-links

Because participants developed responses to four probes, and because the final map was a composite of four individual maps, each concept-to-concept relationship could be represented up to four times on the composite concept map. Multiple iterations of a concept-to-concept relationship were drawn with one dotted connector line labeled with the number of connections represented by the relationship. Initially, multiple iterations were counted as one unit when Branches were scored, but counted as multiple links when Cross-links were scored.

In addition to the weighted sum, an unweighted, unbiased sum was totaled. For the purposes of the unweighted sum, (a) all Cross-links between dyads were scored only once, even if multiple iterations of a Cross-link were specified within a participant's final concept map and (b) all subscores were weighted by 1. A Proposition is defined as two concept nodes connected by a relation. A fifth subscale was also totaled, the unweighted Proposition score.

<u>The Subscale-to-Node Ratio</u>

During the assessment task participants had been prompted to consult, if needed, a list 13 targeted domain concepts as they prepared their responses to the four prompts. Therefore, it was possible for test-savvy participants to prepare an essay-type response by simply listing concepts, whether or not that participant had integrated the concepts within a mental model of the targeted domain. This motivates an analysis that measures the ratio between the number of concept nodes present within a final concept map and scores for each of the other subscales. During data processing the raters isolated and labeled all concepts and then prepared concept maps according to the connections present within a response set. This procedure makes it possible to isolate a relationship between the mean number of Concept nodes present within a final concept map and the mean number of Levels, Branches, or Cross-links. This is the subscale-to-node ratio. For each subscale, it is composed of an individual's subscale score in the numerator and number of Concept nodes in the denominator. This is similar to Fisher's (2000, pp. 208-209) instant/concept ratio, which she uses to measure the interconnectivity of a network. Fisher's instant, like the Proposition defined for this study, is two concepts connected by a relation. Within Fisher's system, a high instant/concept score indicates a robust, highly interconnected expert network. The subscore-to-node ratio assumes that a larger fraction indicates greater domain integration.

Results

<u>Analysis 1</u>

Means and standard deviations are reported in Appendix A. I divided the 57 participants into high (N=28) and low (N=29) achievement according at the median score of 27 on the unweighted concept map sum (node + link + branch + cross-link), and I conducted a 2 X 2 X 4 multivariate mixed design analysis with Achievement (low and high) and Interface Condition (Control vs. Metaphor) as the between-subjects factors and the four subscale-to-node ratios derived from the Level, Branch, Proposition, and Cross-link (unique, unbiased version) subscores as the within subjects factors. I used Achievement specifically to test its interaction with Subscore and ignored any main effect for Achievement.

This design also allowed me to investigate for Subscore redundancy by testing pairwise comparisons. I used the Bonferroni procedure to correct for possible Type I errors.

A nonsignificant Box's Test of Equality of Covariance Matrix, $F(30, 6787)=1.27$, $p=.148$, suggested that the data meet the assumption of equal variance-covariance matrices across the cells formed by the between-subjects effects (SPSS Inc., 1999). I report the Pillai's Trace statistics, as they are considered robust. The main effect for Subscores was significant, $F(3,51)=249.64$, $p<.01$, $\eta^2=.94$; however, interpretation of the main effect and mean differences is influenced by the significant interaction between Subscore and Achievement, $F(3,51)=33.42$, $p<.01$, $\eta^2=.66$. The remaining interactions were not significant: $F_{interfaceXsubsore}(3,51)=.281$, $p>.05$; $F_{subscoreXachievementXinterface}(3,51)=.081$, $p>.05$.

Levene's Test of Equality of Error Variances produced nonsignificant F statistics for each of the four subscales, evidence that the data met the assumption of equal variance across the cells (SPSS Inc., 1999). I used the Bonferroni correction procedure; pairwise comparisons were significant for all pairings except for Branch and Cross-link.

However, examination of the interaction between Subscore and Achievement, (see Figure 2) reveals that the characteristics of the two distributions are quite different. On average the low achievers' cross-link-to-node ratio was their lowest ratio. The proposition-to-node ratio was their highest. That is, participants developed far fewer Cross-links per Node than they did Branches, Levels, or Propositions. In contrast, on the average, the high achievers' cross-link-to-node ratio was second only to their proposition-to-node ratio. Figure 3 illustrates the relative mean differences between high and low achievers' subscore-to-node ratios. The Cross-link mean difference is the largest, followed by the Proposition mean difference. The level-to-node mean difference is the smallest of the four ratios. The figure also illustrates the ratio of the two mean differences, D_r, which I formed by dividing a subscore dyad's (subscore high and low means) mean difference by the mean of the dyad means (see equations 1– 3) to correct for any floor effects:

$$D = \overline{X}_{High} - \overline{X}_{Low} \tag{1}$$

where D = mean difference, $\overline{X}$ = subscore mean

$$M\overline{X} = \frac{\overline{X}_{High} + \overline{X}_{Low}}{2} \tag{2}$$

where $M\overline{X}$ = mean of the subscore dyad means

$$D_r = \frac{D}{M\overline{X}} \tag{3}$$

where D_r = ratio of the two mean differences

The patterns within the mean differences are heightened by the ratio of mean differences version. Cross-links are twice as sensitive to differences between high and low achievement as Propositions, 3 times as sensitive as Branches, and 15 times more sensitive than Levels.

Levels measure hierarchical differentiation, such as causal relationships (e.g., one hierarchical pattern found in multiple participants' protocols was social collective → informal education → role differentiation → static change). Branches measure a learner's ability to differentiate domain concepts into conceptual strands. It appears that learners can state hierarchical differentiations whether or not they hold richly connected domain models. Ability to sort relationships into strands requires a richer domain understanding than Levels; it is about 4 times more sensitive to differences in low and high achievement than Levels. Propositions are dyads, relational connections between two concepts (e.g., *goals* are met through *social solidarity*; see domain concept map in Figure 1). The greater the quantity of relational connections, the richer the mental model. High achievers record more Propositions. Cross-links are evidence of domain integration, what Ausubel (1963, 2000) defined as integrative reconciliation. Within the expert model of the domain used for this study, Cross-links occurred for every branch because all Branches converge at informal education (see Figure 1). Whenever a learner makes connections between disparate Branches, the learner constructs a Cross-link. The ability to identify Cross-links, to make relational connections between concept strands, was the primary difference between high and low achievers.

The subscale-to-node ratio provides its own illustration of the average structural relationships between the concept map nodes and the subscales for this domain and this population (see Figures 4, 5, 6, and 7). For example, Figures 6 and 7 illustrate hypothesized concept maps constructed according to the treatment high and low achiever subscore-to-node ratios. Figure 6 illustrates a map constructed according to the treatment low achiever mean ratio for Cross-links. It contains seven Concept Nodes and two Cross-links. Figure 7 illustrates a map constructed according to the treatment high achiever mean ratio for Cross-links. It contains seven Concept Nodes and seven Cross-links. The high achiever map exemplifies greater integration than does the low achiever map.

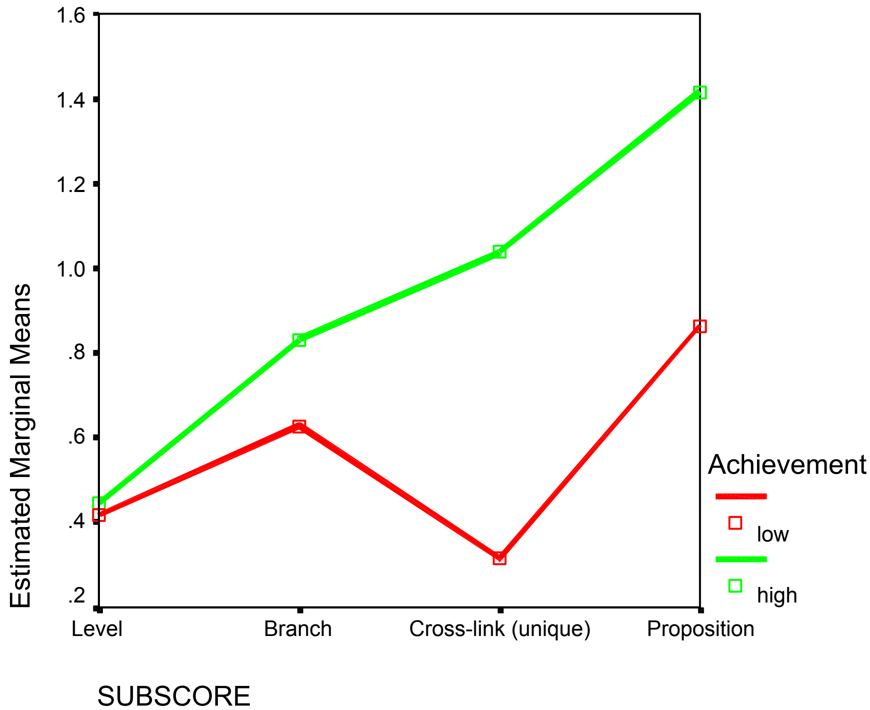

FIGURE 2 The Interaction Between Achievement and Subscore-to-Node Ratio. The median score for the unweighted sum of the Node, Level, Branch, and Cross-link subscores was 27. Low achievers earned 27 or less, while high achievers earned 28 or better. I calculated the subscore-to-node ratio by dividing a participant's subscore total by that participant's number of Concept Nodes. Only unique cross-link relationships counted toward this subscore total.

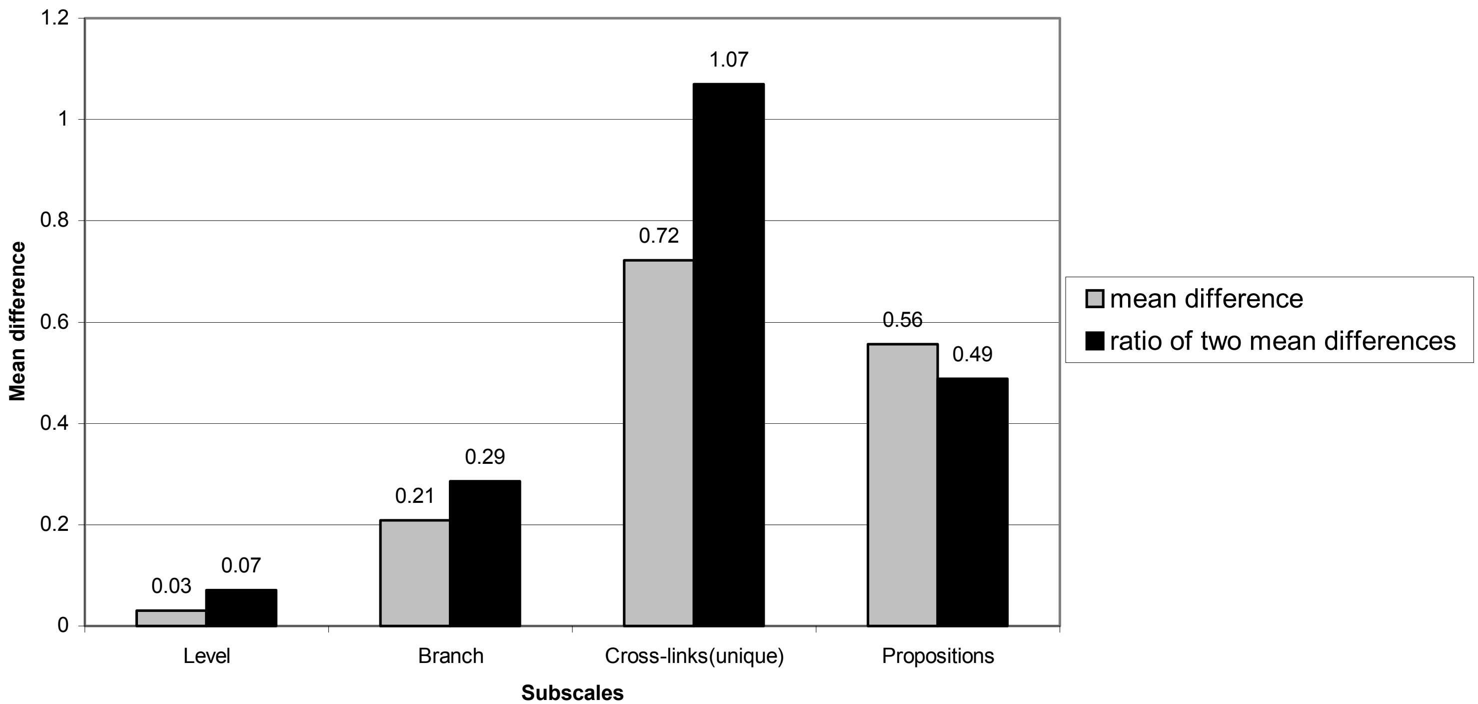

FIGURE 3 Mean Differences and Ratio of the Two Mean Differences Between High and Low Achievers on the Subscore-to-Node Ratio, Unbiased and Unweighted. Gray bars represent the mean difference between low and high achievers for one of the four subscale-to-node ratios. The Cross-link mean difference is the largest, followed by the Proposition mean difference. The level-to-node mean difference is the smallest of the four ratios. The pattern is the same for the ratio of the two mean differences (mean difference/mean of the high/low dyad means), represented by the black bars.

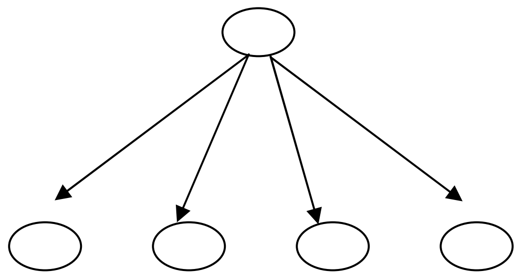

FIGURE 4 Hypothetical Concept Map Showing the Average Proportion of Levels to Concept Nodes. There are two levels and five nodes.

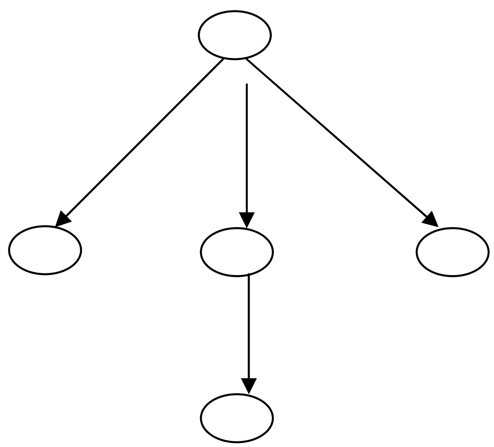

FIGURE 5 Hypothetical Low Achiever (Treatment) Concept Map Showing the average Proportion of Branches to Concept Nodes. There are three Branches and five Nodes.

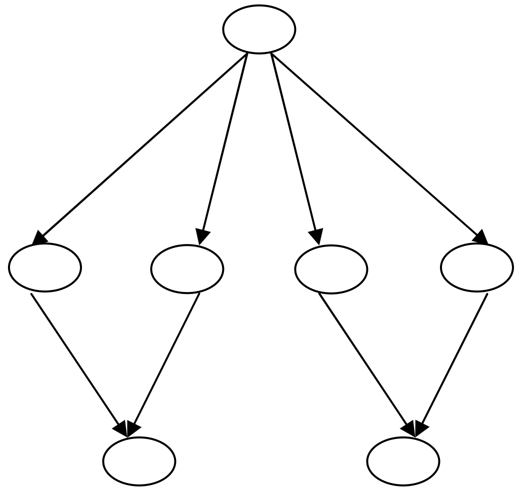

FIGURE 6 Hypothetical Low Achiever (Treatment) Concept Map Showing the Average Proportion of Cross-links to Concept Nodes. There are two Cross-links and seven Nodes.

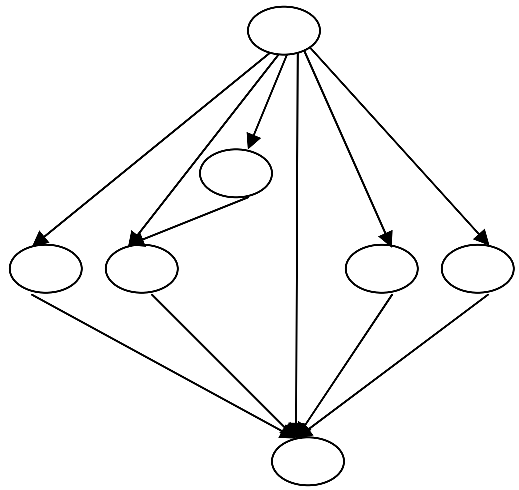

FIGURE 7 Hypothetical High Achiever (Treatment) Concept Map Showing the Average Proportion of Cross-links to Concept Nodes. There are seven Cross-links and seven Nodes.

Analysis 2

Although I had used the Achievement factor to test the interaction and not its main effect on Subscores, I wanted to run a parallel analysis in which the criteria for Achievement did not depend directly upon the subscores that comprised the Subscore factor. Therefore, I also ran a parallel analysis, Analysis 2, using the median (10) of the Proposition subscore to divide participants into high (N=27) and low (N=30) Achievement. I did not include the Proposition subscore as one of the within-subjects factors. Though main effect for Subscore remained significant, the *F* statistic decreased, and the interaction effect within the second analysis was a bit stronger than the Subscore main effect, F_{Subscore} (2,52) = 40.61, $p < .01$, $\eta^2 = .61$; $F_{\text{SubscoreXAchievement}}$ (2,52) = 41.42, $p < .01$, $\eta^2 = .61$. The other main effects and interactions remained the same as when level of achievement had been determined by the unweighted sum of Concept Node, Level, Branch, and Cross-link subscores (Analysis 1). The mean differences between high and low Achievement on Subscores for Analysis 2 is almost identical to that produced within Analysis 1 (see Figure 7 and Table 2). According to the subscore-to-node ratio, Cross-links remain a more sensitive indicator of the difference between high and low achievement than the other two subscales (Levels or Branches). These findings duplicate the results explained within Analysis 1.

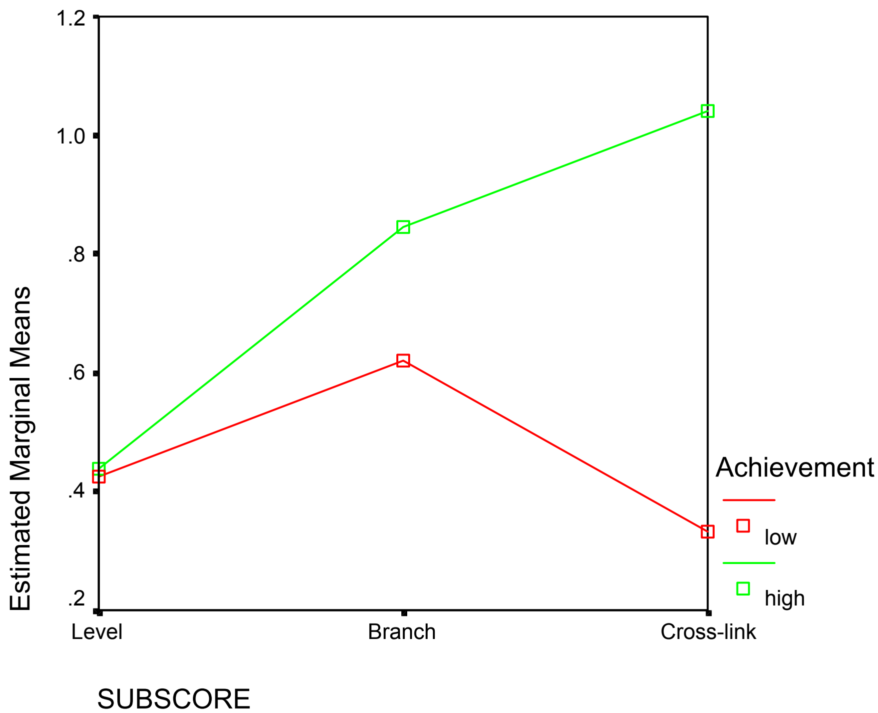

FIGURE 8 Subscore-to-Node ratio, Achievement Groupings Determined by Proposition Median.

TABLE 2

Comparison of Mean Difference Between Analysis 1* and Analysis 2**

Subscore	Subscore Mean**		Subscore Mean Difference by Achievement	
	Low	High	Analysis 1*	Analysis 2**
Level	.42	.44	.03	.02
Branch	.62	.85	.20	.23
Cross-link	.33	1.04	.72	.71

Note: * Analysis 1 divided participants into low (*N*=29) and high (*N*=28) achievers based upon the median of the sum of the unweighted subscores (Concept Nodes, Levels, Branches, Cross-links)

**Analysis 2 divided participants into low (*N*=30) and high (*N*=27) achievers based upon the median of the Proposition subscore.

Analysis 3: Learner Profile

To tease out cognitive characteristics that might coincide with participants' high or low achievement of rich, normative mental, as indicated by final concept map scores, all participants were asked to complete the PsychTests Cognitive Abilities instrument. Whether participants were divided (a) at the unweighted concept map subscore sum median as in Analysis 1 or (b) at the proposition subscore median as in Analysis 2, there was no significant difference between high and low achievers' scores on the six subscales (Analogical Reasoning, Arithmetic, Classification, General Knowledge, Logic, and Pattern Recognition) or on the composite IQ score. Appendix B reports results for high and low achievement divided at the unweighted concept map subscore sum median. IQ scores ranged from 94 to 131 for the low achievers and 97 – 135 for the high achievers.

Discussion

Scoring Concept Maps: Application Within Research and the Classroom

Novak and others (Markham & Mintzes, 1994; Novak & Gowin, 1984; Novak & Musonda, 1991; Wallace & Mintzes, 1990) have described concept map scoring procedures that involve differential weighting of the concept map attributes (subscores). Analysis of the subscale-to-node ratios within this study indicated that Propositions and Cross-links are much more sensitive to differences between high and low achievement than Levels or Branches. The number of Concept Nodes within a concept map is an ambiguous indicator. Both low and high achievers can produce protocols that mention a great quantity of concepts. The greatest difference between high and low achievement is the amount of integration represented within the map. Nodes appear to serve as indicators of richness and integration but only when used as a ratio in conjunction with the other subscales. If these findings can be validated across domains, they provide empirical support, rationale, and direction for subscale weightings. While consideration of relative subscale sensitivity may impact subsequent research, a pragmatic impact of the finding concerns direct application within classrooms whenever concept maps are used for assessment.

The cognitive abilities test did not identify significant difference between high and low achievers on the IQ or the Pattern Recognition, Classification, Analogies, Arithmetic, General Knowledge, or Logic subscale test scores. This suggests that differences between high and low achievement for these preservice teachers was not predicted by individual differences within this set of cognitive abilities. The cause of the point spread on final maps scores between high and low achievers remains unidentified. However, to the extent that one can generalize from preservice teachers at the upper undergraduate level and the graduate level, subscale sensitivity also provides some pedagogical insights. Given instruction, it appears to be fairly easy for learners to link key concepts hierarchically. Concept map translations of high and low achieving participants' protocols evidenced very little difference in achievement for hierarchical discrimination. Learners seem to be pretty capable of organizing domain concepts into conceptual strands (branches). However, those individuals with rich mental models of the domain (high achievers) constructed more relational connections between pairs of concepts (propositions). Therefore, these results suggest that instruction should be designed to help students to make relational connections between concepts.

Finally, these results indicate that the greatest dissimilarity between high and low achievement is the ability to integrate disparate conceptual strands (Cross-links). Within an earlier paper (Reese, 2003a) I discussed the results of a simplex analysis of these data. Those results preclude explanations that suggest learners progress through a stage of differentiation on their journey to domain integration (i.e., the subscales are not hierarchical)—at least in the case of this domain, these learners, and these data. High achievers might be predisposed toward integrative reconciliation. Or, both the concept map and metaphor versions of the functionalism instruction might have interacted with their individual characteristics, helping high achievers to form Cross-links. The high achievers' achievement was not explained by superior abilities in Pattern Recognition, Classification, Analogies, Arithmetic, General Knowledge, Logic, or overall IQ. Using only the data reported here, any explanation of high achievers' Cross-link accomplishment is pure speculation, a subject for further investigation. However, results do suggest that the low-to-average achievers found it 4 to 15 times more difficult to integrate domain knowledge than to differentiate.

Ausubel had suggested that "the principle of integrative reconciliation in programming instructional material can be best described as antithetical in spirit and approach to the ubiquitous practice among textbook writers of compartmentalizing and segregating particular ideas or topics within their respective chapters or subchapters" (1963, p. 80). Thus, integrated instructional materials should help learners to integrate content. In the case of this study, every aspect of the instruction (e.g., the computer interface, the text narratives, and the images) reinforced the central fact that every strand converged to the same concept—informal education (see Figure 1). This domain was highly integrated, as all concepts converged to one node. The interface was highly integrated.

Even within an instructional environment like this one, in which integration is ubiquitous and highly implied, low-to-average achievers' mental models evidenced lower subscore-to-node ratios for Cross-links than for the any of the other three subscores. While metaphor and concept map advance organizers, coupled with the text narratives and the interface navigation, enabled high achievers to form Cross-links, low-to-average learners, even at the graduate level, seem to need help in making those connections. These results suggest that, for some students, implicit integration might not be good enough. Low-to-average students might need to have those connections explicitly stated through didactic exposition and reinforced through practice and feedback.

Implications: Future Directions

Through analysis of the subscore-to-node ratio, I found a dissociation between concept map subscores for high and low achievers. Compared to Levels and Branches, Cross-links and Propositions appear to be more sensitive to domain richness and integration. This result suggests direction for researchers and educators who use concept maps as an assessment tool. When the instructional goal is richness of the learner's mental model, the subscale-to-node ration could help educators and researchers to assess the relative sensitivity of concept map subscales. It may also help educators to help their students to self-regulate. If valid Propositions and Cross-links were rewarded (i.e., they yield higher grades), engaged learners might seek to create these connections among domain concepts and to verbalize and visualize the connections within concept maps and prose narratives. Pedagogically, the dissociation suggests that learners with characteristics similar to those of the low-to-average participants require explicit instruction in order to achieve the level of domain integration average-to-high achievers can achieve through implicitly integrated instructional materials. Using the PsychTests.com instrument, this study was not able to identify the specific characteristics that predict a learner's ability to integrate domain knowledge. This study was limited to one sample from one population who studied an introduction to one domain. Future research addressing replication (literal replication, operational replication, replication with extension, and constructive replication) may substantiate generalization across learner populations and domains.

These results also pose the question of identification of those learner characteristics, if any, which predispose an individual for reconciliative integration. Are these characteristics domain general or domain specific?

Policymakers, educators, researchers, and theorists have called for instructional environments that advance learners' understanding of core domain understandings. At the same time, policymakers, educators, researchers, and theorists call for aligned assessment methodologies, assessment instruments, and techniques that help teachers to teach and learners to learn by informing both about the state of each learner's understanding of each targeted domain (Hernandez & Reese, 2004; Wiggins & McTighe, 2001). These assessment instruments should become integrated within the culture of a classroom's practice, part of the ongoing and ubiquitous language of discourse. Concept maps are visual tools (Hyerle, 1996), designed to make a learner's mental model explicit and concrete so that it can be analyzed and discussed. By design the concept map is a tool positioned to align with education reform efforts. The concept map has an empirically supported theoretic foundation. By design the concept map models cognitive structure. Experts and high-achieving learners constructed richly integrated models of targeted domains. Educational researchers can systematically authenticate whether the concept map subscales and scoring systems, by design, reveal and quantify expertise.

Appendix A

Descriptives for Subscores and Final Map Score

Dependent variable	$\overline{X}_{control}$	$\overline{X}_{treatment}$	$\overline{X}$	*SD*	Mean difference*
Time	1 hr 47 min.	1 hr 53min.	1 hr 50 min.	19 min.	6 min.
Node	9	9	9	2	.07
Level	4	4	4	1	-.15
Branch	7	6	7	3	-.53
Cross-link	10	9	9	6	-1.3
Final Map Score	67	61	64	33	-6.4

*Mean difference = (treatment – control)

Appendix B

Descriptives on IQ Test and Subscales **for High and Low Achievement Groups Divided at the Median of the Unweighted Sum of Subscores**[1]

	Mean	Std. Deviation	Std. Error Mean
IQ[2]			
high	116.3	9.91	1.9
low	112.4	10.2	1.9
Pattern Recognition[3]			
high	79.2	13.7	2.6
low	74.5	15.8	3.0
Classification[3]			
high	62.2	8.6	1.6
low	62.7	10.5	2.0
Analogy[3]			
high	89.6	8.4	1.6
low	84.8	12.4	2.3
Arithmetic[3]			
high	42.8	17.0	3.2
low	38.5	17.3	3.3
General Knowledge[3]			
high	61.7	17.3	3.7
low	61.7	18.5	3.5
Logic[3]			
high	71.7	19.3	3.7
low	65.0	14.2	2.7

[1] N=56, n=28 for each subgroup.
[2] Scale standardized to a mean of 115.
[3] Scale runs from 0 to 100 with a mean of 50.

t-tests for Equality of Means on IQ Test and Subscales Using High and Low Achievement Groups Divided at the Median of the Unweighted Sum of Subscores

	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
IQ	1.447	54	.15	3.9	2.7	2.0	9.3
Pattern Recognition	1.192	54	.24	4.7	4.0	-3.2	12.6
Classification	-.195	54	.85	-.50	2.6	-5.6	4.6
Analogy*	1.720	47.6	.09	4.9	2.8	-.8	10.5
Arithmetic	.941	54	.35	4.3	4.6	-4.9	13.5
General Knowledge	.000	54	1.0	.00	4.8	-9.6	9.6
Logic*	1.473	49.6	.15	6.7	4.5	-2.4	15.8

* degrees of freedom corrected for unequal variances

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ausubel, D. P. (1962). A subsumptive theory of meaningful verbal learning and retention. *Journal of General Psychology, 66*, 213-224.

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning*. New York: Grune & Stratton.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, and Winston.

Ausubel, D. P. (2000). *The acquisition and retention of knowledge: A cognitive view*. Boston: Kluwer Academic Publishers.

Clariana, R. B., Koul, R., & Salehi, R. (in press). The criterion related validity of a computer-based approach for scoring concept maps. *International Journal of Instructional Media, 33*(3).

Edmondson, K. M. (2000). Assessing science understanding through concept maps. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 15-40). New York: Academic Press.

Feinberg, W., & Soltis, J. F. (1998). *School and society* (3rd ed.). New York: Teachers College Press.

Fisher, K. M. (2000). SemNet software as an assessment tool. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 197-221). New York: Academic Press.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199-241). New York: Cambridge University Press. (Reprinted in *Knowledge acquisition and learning*, 1993, 673-694.).

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99-129). Hillsdale: New Jersey: Lawrence Erlbaum Associates.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45—56.

Gentner, D., & Stevens, A. L. (1983). Introduction. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hernandez, V., & Reese, D. D. (2004). Integrating strategies and technology in education practice: Conducting case studies— Guidelines for classroom observations. Wheeling, WV: Center for Educational Technologies, Wheeling Jesuit University.

Hyerle, D. (1996). *Visual tools for constructing knowledge*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jerabek, I. (2000). Classical IQ test - 2nd revision. Retrieved 2003, from QueenDom.com, PsychTests.com

Kurtz, K. J., Miao, C.-H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of Learning Sciences*, 10(4), 417-446.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: The University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*(4), 586-597.

Markham, K. M., Mintzes, J. J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching, 31*(1), 91-101.

Mayer, R. E. (1979). Can advance organizers influence meaningful learning? *Review of Educational Research, 49*(2), 371-383.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

National Research Council. (2001). *Knowing what students know*. Washington, DC: National Academy Press.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching, 27*(10), 937-949.

Novak, J. D. (1992). *A view on the current status of Ausubel's Assimilation Theory of Learning.* Paper presented at the American Educational Research Association, San Francisco.

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

Novak, J. D., & Musonda, D. (1991). A twelve-year longitudinal study of science concept learning. *American Educational Research Journal, 28*(1), 117-153.

Poindexter, M. T., & Clariana, R. B. (in press). The influence of relational and proposition-specific processing on structural knowledge and traditional learning outcomes. International Journal of Instructional Media, 33(2).

Reese, D. D. (2003a). *Metaphor and content: An embodied paradigm for learning*: Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University.

Reese, D. D. (2003b). Trees of knowledge: Changing mental models through metaphorical episodes and concept maps. In R. E. Griffin, V. S. Williams, & J. Lee (Eds.), *Turning trees: Selected readings*. International Visual Literacy Association.

Reese, D. D. (October, 2003). PFNET translations: A tool for concept map quantification. In M. Simonson (Ed.), *Proceedings of Selected Research and Development Papers Presented at the National Convention of the Association for Educational Communications and Technology* (Vol. 1). Anaheim, CA.

Reese, D. D., & Hergert, T. R. (2004). Informal dissemination of visual representations: A function of nexus characteristics. In R. E. Griffin, V. S. Williams & J. Lee (Eds.), *Changing tides: IVLA book of selected readings*. Newport, RI: International Visual Literacy Association.

Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 304-341). New York: Academic Press.

SPSS Inc. (1999). *SPSS advanced models 10.0*. Chicago: SPSS Inc.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes, 82*(1), 60-75.

Wallace, J. D., & Mintzes, J. J. (1990). The concept map as a research tool: Exploring conceptual change in biology. *Journal of Research in Science Teaching, 27*(10), 923-936.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wiggins, G., & McTighe, J. (2001). *Understanding by design*. Upper Saddle River: NJ: Prentice-Hall, Inc.