DOCUMENT RESUME

ED 320 953	TM 015 237

AUTHOR	Parker, Richard; Tindal, Gerald
TITLE	The Reliability, Sensitivity and Criterion-Related Validity of Concept Comparisons and Concept Maps for Assessing Reading Comprehension.
PUB DATE	89
NOTE	56p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Criterion Referenced Tests; Educational Assessment; *Learning Disabilities; *Multidimensional Scaling; *Reading Comprehension; Reading Difficulties; Reading Tests; Secondary Education; *Secondary School Students; Secondary School Teachers; *Special Education; Test Reliability; Test Validity
IDENTIFIERS	*Concept Comparisons; Concept Maps; *Criterion Related Validity; Experts

ABSTRACT

The reliability, sensitivity, and criterion-related validity of concept comparison (CC) ratings and computer-derived multidimensional scaling (MDS) maps were studied as ways of assessing reading comprehension. Fifteen experts (2 district coordinators and 13 reading specialists/special education teachers) were included in this study. Reliability was studied by comparing CC ratings and maps produced independently by 5 teachers reading eight 250-word passages from elementary level science and social studies texts. For the 3 of the 8 passages that had sufficient interrater reliability, instrument validity was assessed with 104 reading disabled junior and senior high school students. A randomized control group design compared CC tasks completed before and after students read related or unrelated passages. Students reading related passages produced post-reading CC scores significantly more closely related to teacher scores than did readers of unrelated passages. Student and expert CC scores were then correlated with student scores on: (1) vocabulary and reading comprehension sections from norm-referenced reading tests; and (2) maze tests, multiple-choice questions, and oral reading fluency performance, all based on the passages. These three measures were substantially related to post-reading CC scores, but not to pre-reading scores. Standardized test scores were not significantly related to pre-reading or post-reading CC scores. Reliability and validity results support further research on the use of CC tasks and MDS maps for assessing reading comprehension for older disabled readers. Nine data tables, six figures, and three appendices providing study results are included. A 132-item list of references is provided. (Author/SLD)

The Reliability, Sensitivity and Criterion-Related Validity of Concept Comparisons and Concept Maps for Assessing Reading Comprehension

Richard Parker

Gerald Tindal

University of Oregon

Running Head: CONCEPT MAPS

Abstract

Most current reading assessment methods do not reflect the reading comprehension construct which has emerged from information processing research. Current methods rarely account for differences in relevant background knowledge or schema held by students prior to reading, and are insensitive to the structural nature of text information and student knowledge. This study investigated the reliability, sensitivity, and criterion-related validity of concept comparison (CC) ratings and computer-derived multidimensional scaling (MDS) maps for reading comprehension assessment. Reliability was assayed by comparing CC ratings and maps produced independently by five teachers while they read eight 250-word passages from science and social studies texts. For three of the eight passages, sufficient interrater reliability was obtained. For the three reliable passages only, two methods were applied to assay instrument validity with 104 reading disabled Junior and Senior High School students. First, a randomized control group design was used to compare CC tasks completed before and after students had read related or unrelated text passages. Students reading the related passages produced post-reading CC scores significantly more closely related to expert teacher scores than did readers of unrelated passages. Second, student and expert CC score similarities were correlated with student scores on two classes of external measures: (a) extant vocabulary and reading comprehension scores from published, norm-referenced reading tests, and (b) maze tests, multiple choice questions, and oral reading fluency performance—all based on the reading passages. The three passage-based measures were substantially related to Post-reading CC scores, but not to Pre-reading CC scores. Standardized test scores were not significantly related to either Pre- or Post-reading CC scores. The reliability and validity results were interpreted as supporting further validation research on the use of concept comparison tasks and derived MDS maps for assessing reading comprehension with older disabled readers.

The Reliability, Sensitivity, and Criterion-Related Validity of Concept Comparisons and Concept Maps for Assessing Reading Comprehension

Over the past decade, cognitive processing research has played a major role in explicating the construct of reading comprehension (Trabasso, 1978; Freedle, 1979). Central to the current view of reading comprehension is the notion that a reader integrates information s/he reads from text into a pre-existing, organized network of concepts and information, or "schema" (Anderson, 1977; Spiro, 1977; Rumelhart & Ortony, 1977). However, most current reading assessment methods do not reflect the reading comprehension construct which has emerged from information processing research (Kirsch and Guthrie, 1980; Curtis & Glaser, 1983, Johnston, 1984). First, current methods rarely account for differences in relevant background knowledge or schema held by students prior to reading. Furthermore, the information processing demands placed on examinees by test tasks or items often are not closely related to the information demands placed on them by the text (Surber, 1984). There is a growing professional view that the lack of a sound psychological basis for reading comprehension tests has resulted in inappropriate types of test items being presented, and inappropriate types of responses being demanded of the student (Linn, 1982; Glaser, 1981; Messick, 1980; Johnston, 1984).

Because of the inadequate relationship between "knowledge structure of the examinee and that of the test", most standardized reading comprehension tests have been characterized as being "atheoretical" (Schwartz, 1984) or lacking construct validity (Kirsch & Guthrie, 1980, p. 81). This form of validity is rarely addressed by test producers (Johnston, 1984), yet there is a growing recognition of the primacy of construct validity over the traditional categories of criterion-related and content validity (Messick, 1981; AERA, APA, NCME, 1985). Guion (1978) states that the category of "content validity" should be dropped in favor of a set of content-oriented rules for test development. In the same vein, Anastasi (1986) concluded that "all validation procedures contribute to construct validation and can be subsumed under it" (p. 12). This encompassing notion of construct validity has encouraged theorists with a cognitive processing point of view to suggest fundamental improvements in reading comprehension tests. Whereas test developers traditionally have been concerned with adequately sampling behavioral (Aiken, 1979; Anastasi, 1976) or content (Brown, 1976; Thorndike & Hagan, 1977) domains, that focus is shifting. Partly in response to past difficulties in defining these behavioral and content domains, the focus has shifted to defining what cognitive processes and structures are represented both in test content and requisite student responses (Kirsch & Guthrie, 1980; Guion, 1978).

A second shift in psychometric thinking about validity is related to the uses of test scores. Test use in schools has social consequences for teachers and students. Terms such as "decision validity", "discriminant validity", and "treatment validity" are used increasingly to refer to test score interpretation and test score use in decision making (Hambleton, 1980; Messick, 1981, 1989). Valid test scores are socially valued and their use is consistent with schools' broader mission and goals (Messick, 1989). Johnston (1984) notes that the concept of test validity has been moved back to its instructional context, and suggests that future validation studies include instructional interventions.

From the cognitive perspective, reading comprehension tests must reflect both organization of prior knowledge (pre-reading schema), and selection and organization of key concepts from text (Johnston, 1984). Researchers have therefore sought a standard symbolic notation for displaying the content and structure of both the text and the reader's recall of text: "where the content and structure of both...can be specified, the two structures can be compared" (Meyer & Rice, 1984, p. 320). Kirsch and Guthrie (1980) also seek a method for matching "the knowledge structure of the examinee and that of the test" (p. 81). Reading comprehension could then be described in terms of structural and content differences between text and reader's cognition.

Several formal psycholinguistic models exist for information structure in text and/or knowledge (cognitive) structure in the learner. Most imply that information is stored as abstract, non-spatial, non-analogical semantic or propositional networks, with rule systems which can be made explicit (Anderson & Bower, 1973; Rumelhart, Lindsay & Norman, 1972). These models (as described by Meyer & Rice, 1984) are not well suited for assessment, because they: (a) are laborious to apply, and require a high level of expertise, (b) have untested reliability, (c) are often tied to text conventions, and cannot measure pre-reading knowledge schema, (d) focus on detailed micro-level analyses of very short passages.

Alternative models are offered by cognitive psychologists who contend that "...humans use frameworks similar to geometric spaces for organizing or perceiving many types of objects or concepts" (Fenker, 1975.

p. 39). Johnson-Laird (1983), argues that propositional models are too narrow to reflect the cognitive maps or "Mental Models" which we use to capture key perceptual and logical relationships. Similarly, van Dijk and Kintsch (1983) introduced "situation models", to respond to deficiencies noted in narrower propositional models (Brewer, 1987).

The assessment method investigated in the present study is consistent with Johnson-Laird's Mental Models in that relationships among concepts are depicted as spatial arrangements, and may be interpreted in either concrete/perceptual or abstract/propositional terms. The concept maps produced in this study are also similar to those vocabulary- and text-maps used by teachers to help teach content vocabulary and explain key concepts in text (Niles, 1965; Hauf, 1971; Heimlich & Pittelman, 1986). These devices include two-dimensional "webs" of key concepts, characters, or events (with connecting lines sometimes labeled) and hierarchical branching trees, with a broad topic as the trunk, and details or subordinate concepts at the ends of branches (Calfee & Drum, 1986; Reutzel, 1986; Holley & Dansereau, 1984). In a variety of forms, these semantic maps have demonstrated usefulness as learning tools (Guri-Rozenblit, 1989; Vaughan, 1984; Reutzel, 1986; Voss, Greene, Post, & Penner, 1983). However, assessment in this area has unfortunately lagged behind instruction. Researchers lack proven, replicable methods for (a) producing maps and hierarchical diagrams from text, and (b) using these same structurally sensitive methods to measure student learning (Surber, 1984).

The purpose of this study was to use structurally sensitive methods to assess reading comprehension, including measurement of pre-reading schema, text structure, and post-reading semantic knowledge. Our goal was to use a spatial measurement method, following Johnson-Laird's (1983) hypothesis that physical space may serve as an analogue for one or more dimensions of perceptual/conceptual reality. Two measurement methodologies (one primary and one supplemental) can potentially address this need; however, to date their application to reading comprehension has been very limited.

Multidimensional Scaling and Hierarchical Cluster Analysis

Both multidimensional scaling (MDS) and hierarchical cluster analysis (HCA) (Preece, 1976; Shavelson, 1974) yield graphic displays of key topics or concepts, where spatial proximity or "linkages" depict similarity or closeness of relationship. MDS yields a map of concepts represented as points in two (or more) dimensional space, while HCA yields a branching tree, with concepts at the ends of the branches connected to a common trunk. MDS begins with judgments on the closeness of relationsnip of pairs of important concepts or key vocabulary words. The MDS map distances may then be analyzed through HCA to produce a cluster tree.

MDS maps may serve three purposes: (a) improved comprehension and communication of complex relationships among concepts, (b) verification of hypothesized concept patterns (comparing obtained configurations with external criteria), and (c) interpretation of map dimensions (Davison, 1983). Only the first two applications are relevant to this study. Concept maps can help provide and communicate meaning through the identification and labeling of (a) concept clusters, (b) relationships among concepts and concept clusters, (c) hierarchical (subordinate) relationships among concepts and concept clusters. Interpretation of concept clusters and inter-concept relationships is demonstrated on MDS maps (Figure 1) from a concept comparison rating task (Figure 2), completed after reading a 250 word science passage, "The Heart" (Figure 3). A more detailed explanation of the concept comparison task will be provided later.

Insert Figures 1, 2, & 3 about here

In Figure 1, MDS procedures were used to plot eight important concepts from a science text passage. In the top map, meaningful concept clusters were objectively identified, then outlined and labeled. In the bottom map, map interpretation highlights relationships rather than clusters. The relationship labels are similar to those listed by Holley & Dansereau (1984), Frederiksen (1975), and de Beaugrande (1980). Although mainly objective procedures are used to produce MDS maps from pairwise ratings, subjective judgment is required for map interpretation, as well as knowledge of the content area and the particular text passage.

Interpretation of the HCA cluster tree includes (a) selecting the most defensible branching level(s), and (b) providing definitions or descriptions for the categories (clusters) at those level(s). Cluster analysis has proved valuable in this secondary analysis role (Coxon, 1982; Griffeth, Hom, DeNisi, & Kirchner, 1985). Figure 4 displays a HCA solution for the MDS map configuration. The scree plot (explained later) beneath the tree indicates that a three-cluster solution is most defensible; however, a five-cluster solution is also interpreted on the tree for demonstratien purposes.

Insert Figure 4 about here

The "trunk" of the tree is labeled with the passage title, "The Heart". Note that the branch labels for the three-cluster solution are those used for the clusters on the first MDS map in Figure 1. The cluster tree thus provides a third, complementary interpretation of an MDS map configuration.

Figures 1 and 4 demonstrated the first main purpose of MDS—improved understanding and communication of complex semantic relationships. The second main MDS purpose—verifying concept patterns—can use externally produced "expert" maps as standards for evaluating an individual student's concept map. The expert standard map and learner maps can be compared through analysis of similarity of concept cluster membership and/or similarity of inter-concept map distances. Qualitative comparisons between standard and learner maps also are possible by interpreting map configuration differences as more or less serious or trivial.

Figures 1 and 4 demonstrate the strength and limitation of the MDS mapping procedures. The spatial dimensions are not well suited to displaying syntactic or mechanical text-based structures or detailed networks of propositional relationships. The maps do, however, provide a very flexible "problem space" for demonstrating a range of semantic relationships, including both abstract and perceptual anologue relationships. In this way, they most closely approximate Johson-Laird's (1983) Mental Models construct. Although the mapped elements in Figure 1 are "micro-units" (individual vocabulary terms), the interpreted map depicts a "macro-level" structure of total content organization (Meyer & Rice, 1984). The semantic maps appear equally well suited to measuring pre- and post-reading knowledge structures, and semantic relationships in text.

Neither MDS nor cluster analysis is as well validated as the more common parametric multivariate techniques of factor analysis and discriminant function analysis (Davison, 1983; Aldenderfer & Blashfield, 1984). However, MDS is supported by a body of psychometric research, summarized in recent reviews

(Carrol & Arabie, 1980; Young, 1984), textbooks, (Davison, 1983; Schiffman, Reynolds, & Young, 1981) and dedicated journal issues (Applied Psychological Measurement, Vol. 7, No. 4, 1983; Psychometrika, Vol. 51, No. 1, 1986). While MDS lacks the statistical power associated with normal distribution assumptions and interval/ratio measurement scales, it offers distinct benefits. Foremost are that (a) MDS solutions are easily interpreted, (b) MDS provides valid results with ordinally-scaled data, and (c) the methodology is suitable for small sets of observations (Schiffman, Reynolds, & Young, 1981). In addition, MDS can usually fit an appropriate model to the original data in fewer dimensions than factor analysis (Wilkinson, 1989).

HCA, which is relegated in this study to supplementary analyses, is considered an "exploratory" technique—seldom recommended for primary analyses (Everitt, 1988, p. 604). Together, MDS and cluster analysis offer spatial maps and hierarchical trees which are similar to the more flexible spatial cognitive models (Johnson-Laird, 1980; Holley & Dansereau, 1980; van Dijk & Kintsch, 1983), as well as the maps traditionally constructed by teachers intuitively and by hand. Unfortunately, relatively few studies exist in which the methods have been applied to reading or other student learning.

Multidimensional Scaling and Student Learning

Multidimensional Scaling has been used to study changes in students' semantic structures following instruction in social studies (Stasz, Shavelson, Cox, & Moore, 1976), research design (Fenker, 1975), and psychology (Weiner & Kaye, 1974; Deikhoff, 1982). Fenker (1975) conducted two studies matching student MDS maps with those from subject matter experts—both before and after instruction. The closeness of relationship of pairs of "research design" concepts were judged by eight experts, and then by twenty students enrolled in the university course. The MDS maps produced by the experts were substantially similar. Student maps showed only slightly stronger agreement with expert maps from before to after instruction. In the second study, 27 new students were additionally directed to give special attention to learning the key concepts and their interrelationships. Post-instruction results demonstrated greater similarity between student and expert maps. In addition, a significant relationship was found between students' course grades and the similarity of their own maps with the experts'.

External criteria such as course grades and test results have also been used to help validate concept comparison (CC) scores and the derived MDS maps (Brown & Stanners, 1984; Diekhoff, 1983; Stanners, Brown, Price, & Holmes, 1983). Diekhoff (1983) compared multiple-choice, essay test, and CC test results by 120 undergraduate students enrolled in a psychology class. Correlations between the CC task and the other two test forms were .44 and .58, respectively, leading the author to conclude that "relationship judgment tests tap both definitional knowledge of the sort measured by the multiple-choice tests ... and structural knowledge of the sort measured by essay tests" (p. 230).

In two studies, Stanners, Brown, Price, and Holmes (1983) compared performance by 64 psychology students on a CC task with three types of short-answer essay questions on the same content: definition questions, applications, and questions requiring discussion of relationships. Following analysis by MDS, CC scores correlated .66 with a composite of the three essay question types. The authors stated that

> the concept comparison task would appear to be useful whenever the focus of interest is on a complete pattern of relationships among units of knowledge. The rating data are relatively easy to gather and, when analyzed by multidimensional scaling, allow both visual and quantitative forms of representation. The results ... provide evidence that such representations reflect actual knowledge of conceptual interrelationships (p. 863).

Multidimensional Scaling and Reading Comprehension.

More directly related to the present study are the few applications of MDS to expository and narrative reading passages (Bisanz, LaPorte, Vesonder, & Voss, 1978; LaPorte & Voss, 1979; Beaugrande, 1980; Stanners, Price, & Painton, 1982). These studies produced two-dimensional cognitive structure maps from student recall of story elements, and compared the student-produced maps with either pre-reading maps or "expert" maps. Stanners et al. (1982) had 60 college students rate all possible combinations of five fictional characters and three settings after reading an O. Henry short story. Most of the MDS generated maps contained two dimensions: time sequence and character-setting connections. A second finding was that mapped configurations of story elements were found to change as a function of pre-reading the text.

LaPorte and Voss (1979) explored changes in cognitive maps produced by college students before and after text reading. Students in a control group also completed the concept-comparison tasks, but did not read the two, 100-word descriptive passages from which the words were drawn. Students judged relationships between vocabulary pairs immediately after reading the passages and again, 48 hours later.

Changes in concept ratings between the pre- and post-reading assessments accurately reflected the subjects' increased understanding of the story. The authors also found that the ease of delayed passage recall was due to the similarity of the story structure and students' pre-reading schema or knowledge structure structure.

Two of the preceding studies (LaPorte & Voss, 1979; Stanners, Price, & Painton, 1982) have focussed on Davison's (1983) third type of MDS application: dimensional interpretation to summarize a map configuration. That use of MDS is parallel to factor analysis, where the researcher seeks a relatively few factors with efficient explanatory power. Within the present study, however, the focus is on concept configurations—clusters and relationships. Reducing data to two dimensions greatly reduces the method's diagnostic utility (Shepard, Kilpatrick & Cunningham, 1975).

The few studies applying MDS to student learning and reading in particular are encouraging. However, those reading studies have employed a very limited number and variety of passages, mainly from adult-level reading material and with able readers Map interpretations most often have been dimensional rather than configurational, reducing their potential for diagnosis and instruction. Although the validity of MDS procedures has begun to be addressed in the few studies just cited, reliability has not.

Purpose

This study investigates the use of concept comparisons and spatial maps for assessing comprehension of expository reading passages by Jr. and Sr. High School students with reading disabilities. The study was conducted in two phases, addressing instrument reliability (Phase 1), and instrument sensitivity and criterion-related validity (Phase 2). The central question of Phase 1 was: After reading 250-word science and social studies textbook passages, will teachers independently produce similar concept comparison (CC) ratings and MDS maps? The usefulness of MDS in assessing reading comprehension depends partly on the reliable identification of "expert" maps to compare with pre- and post-reading student maps. In Phase 2, pre- and post-reading CC scores and MDS maps from disabled readers in Junior and Senior High Schools were compared with the expert teacher maps and with four external criterion reading measures.

Phase 1 : Instrument Reliability

Method

Reading Passages

Eight 250-word passages were selected from elementary level social studies and science texts (Holt Science, Holt General Science, Heath Life Science, Heath Social Studies). The content of the selected passages, with their Fry readability levels, are: One-Celled Organisms (3.0), "Igneous Rocks" (5.8), "The Heart" (5.2), "The Seashore" (7.5), The History of "Texas" (4.5), Regions of the "Soviet Union" (6), The Skeletal and Muscular Systems (7), Limits on Animal Population Growth (7). Selected passages are included in Appendix A.

Passages were selected to be cohesive and self-contained within a 250-word limit, and typically contained at least one central idea and 8 to 12 key content-related vocabulary terms. Passages were minimally edited to delete "asides", references to charts, figures, and text located elsewhere, and sentences of only peripheral reference.

Eight key vocabulary terms were selected from each passage for pair-wise judgments within a concept comparison (CC) test. "Key vocabulary" were words with central importance to the passage, including both content words and non-content words with content-specific meanings within the text. Words selected included all those highlighted by text publishers through bold/italic type, underlining, or margin notes. "Key vocabulary" and "concepts" are used interchangeably in this paper.

Concept Comparison Tasks

For each selected passage, all pairwise combinations of the 8 key vocabulary terms were listed in a "Ross ordering" sequence to avoid contaminating order effects (Cohen & Davison, 1973; Davison, 1983). Although a minimum of 9 concepts are recommended for a 2-dimensional MDS map (Kruskal & Wish, 1978), that recommendation assumes that only one CC task is conducted, and can be "weakened somewhat" (Schiffman, Reynolds, & Young, 1981, p. 24) for multiple ratings as in this study, where ratings for each passage were obtained (and then aggregated) from five different teacher experts.

Beside each pair of concepts, respondents used a 4-point scale to judge how closely the two terms were related or connected in the passage—i. e., how much the terms "had to do with each other" or to what extent they "could be used to describe each other". The cues "close relation" and "little or no relation" were attached to the two extremes of the scale. The CC task yielded a set of 28, 1-4 ratings on each passage from each teacher (see Figure 2).

Respondents

The 15 "expert" respondents, all employed by a rural, Pacific Northwest school district, included two district coordinators and 13 reading specialists and Special Education teachers from six Junior (Gr. 6-8) and Senior (Gr. 9-12) High Schools. Of the Junior High School teachers, three taught in Special Education resource rooms (PL 94-142 categorical), and three in Chapter 1 (remedial compensatory) programs. Five of the High School teachers taught Special Education, and two Chapter I. For each of the eight passages, five teachers separately completed a CC rating task. No teacher rated the same passage twice.

Procedure

The "expert" raters first read the 250-word passages and then independently completed related CC tasks. While making concept comparison judgments, they were encouraged to look back at the passage and to change initial ratings if they wished. No time limit was set for the task; most respondents required 7 to 9 minutes to read and rate each passage. Each teacher completed three or four CC tasks during each of two sessions. Members of the research team introduced the task to the group, and were present through the sessions to proctor and answer questions.

Data Analyses

Interrater agreement was first calculated for teachers' concept comparison ratings using two indices: the intraclass correlation (Brennan, 1983; Cronbach, Gleser, Nanda, & Rajartnam, 1972) and Cohen's Kappa (Fleiss, 1981; Cohen, 1968). Next, for only the most reliable passages, HCA was conducted on map clusters, and agreement on cluster membership was assessed (Rand, 1971; Morey & Agresti, 1984).

Results

Concept Comparison Reliability

Concept comparison ratings from five teachers were analyzed for each of the eight passages, using the intraclass correlation coefficient (Brennan, 1983) and Cohen's Kappa (Fleiss, 1981). Two methods for improving the interpretability of Kappa are (a) calculating the ratio of obtained Kappa to the maximum

Kappa obtainable (Brennan & Prediger, 1981), and (b) differentially weighting scores by the degree of disagreement on the ordinal rating scale (Cohen, 1968). Table 1 presents these measures of agreement for five "expert" raters on the eight passages.

Insert Table 1 about here

Intraclass correlations are all moderate to high, while simple Kappas are more variable and lower, ranging from .27 to .51; values at .40 and above indicate "good" agreement beyond chance (Fleiss, 1981). Reconsidering Kappas in ratio to their maximum possible value (Kappa Max.) yielded substantially higher values (.35 - .78 range). Similarly, differential weighting degrees of disagreement increased Kappas by .10 - .15 points. From the tabled information, three CC tasks—"The Heart", "Igneous Rocks", and "The Skeletal System"—demonstrated sufficient reliability for use with students in the second phase of the study. For each of these passages, the concept comparison scores were averaged across raters in preparation for the second phase of the study.

Map Configuration Reliability

The preceding reliability indices were based on raw CC rating scores. Reliability of map clustering was next assessed, but for only the three most reliable passages. For these passages, an MDS map was produced for each of the five raters, using the stand-alone ALSCAL-4 statistical software (Young & Lewyckyj, 1973) with the classical non-metric (CMDS) algorithm. The goodness of fit of each map to the rating data was first assessed through Kruskal's Formula 1 Stress (Davison, 1983). All but one of the fifteen Stress values were below .02, representing a very good fit for two dimensions and at least nine concepts (Kruskal & Wish, 1978). However, the small number of mapped concepts may have been somewhat overfit to two dimensions, artificially lowering Stress values.

Agreement among the five MDS map configurations was assayed by (a) comparing inter-concept map distances through the intraclass correlation, and (b) comparing cluster composition through the Rand statistic. Euclidean map distances between all possible concept pairings (28 in all) are analogous to the original 28 CC ratings. The intraclass correlation reliability estimates for map distances were: "The Heart"

.86; "Igneous Rocks" .69; "The Skeletal System" .83 all significant at $p < .01$.

To assess agreement of cluster composition from the maps, the number and composition of clusters first had to be determined. Although concept clusters often can be discerned visually, a more systematic procedure was used: HCA accompanied by scree plots (Davison, Richards, & Rounds, 1986; Coxon, 1982). The Group Average clustering algorithm (Sneath & Sokal, 1973), was used, as it produced interpretable solutions for these data and performed well in Monte Carlo studies (Milligan, 1980, 1981).

On a cluster tree, each branching level is a different potential clustering solution. The optimal clustering level(s) are identified on a scree plot of "number of clusters" by "joining distances" (Mojena, 1977; Aldenderfer & Blashfield, 1984). As in factor analysis, a flattening of the scree line indicates the optimal partition. These procedures identified one or two optimal clustering solutions for each rater for each map. Following map cluster identification, agreement on cluster membership was assessed using Rand's statistic, which was devised for this very purpose (Rand, 1971). A chance-correction for the Rand, "omega" (Ω), was used, which is scaled from 0 (chance agreement) to 1 (perfect agreement) (Morey and Agresti, 1984). The Ω ranges (and medians) showed uniformly high cluster agreement: "The Heart" .73, (1.0), 1.0; "Igneous Rocks" .48, (.68), 1.0; "The Skeletal System" 1.0, (1.0), 1.0. In summary, reasonable interrater reliability was obtained for these three passages, based on CC scores, map distances, and map clustering.

In preparation for Phase 2 of the study, an average "expert map" was then created for each of these three reliable passages. First, the five teachers' CC ratings were "externally averaged" (Schiffman, Reynolds, & Young, 1981, p. 179) For each average data matrix an MDS map was then processed through ALSCAL-4's classical non-metric algorithm. The more complex Replicated algorithm (RMDS) produced nearly identical clusterings to the simpler CMDS solution. The main advantage of RMDS is its ability to describe "dimensional variation" among individual respondents, which does not address our goal of producing a valid average map (Schiffman, et al. 1981, p. 65). Therefore, only the CMDS procedure was used in this study.

Optimal cluster solutions on the average expert teacher maps were then identified through the HCA-plus-scree plot procedure described earlier. These three average maps, with optimal clusters outlined, are presented in Appendix B.

Phase 2: Instrument Sensitivity and Criterion-related Validity

The purpose of the second phase of the study was to investigate the sensitivity and criterion-related validity of student CC scores and related MDS maps for assessing reading comprehension. Two main comparisons were conducted. To gauge sensitivity, students completed CCs before and after reading, and their pre- and post-reading scores were correlated with the average "expert" CC scores. To determine validity, each student's degree of association with "expert" scores was compared with his/her performance on two classes of external measures: (a) extant vocabulary and reading comprehension scores from published, norm-referenced reading tests, and (b) maze tests, multiple choice questions, and oral reading fluency performance—all based on the reading passages.

Method

Respondents

This study was conducted in a west coast low-middle SES rural community with an economy dependent on the logging industry. At the Jr. and Sr. High levels the lowest achieving nine percent of each grade cohort (approximately 330 in all) were enrolled in Chapter 1 (compensatory) or Special Education (LD category) programs in reading/language arts. From this population were sampled 240 students—all those for whom current standardized achievement data were available. The high rate of absenteeism, school transfers, and incomplete test protocols reduced this sample to 104 by the end of the study. Yearly enrollment turnover was nearly 40% for the district, and exceeded 60% for students in special programs. All data presented are for the104 students, drawn from thirteen classrooms within four Jr. (Gr. 6-8) and two Sr. (Gr. 9-12) High schools.

Fifty-three of the 104 students were enrolled in Jr. High, and 51 in Sr. High. Forty-three attended Special Education resource rooms for language arts, and 61 received pull-out Chapter 1 assistance. Current standardized achievement test scores from the district-administered Metropolitan Achievement Tests ( ) were available for 81 of the students. For the remaining 23 students, current Woodcock Johnson (13), WRAT (5), Nelson Achievement Tests (2), and Iowa Achievement Tests (2) were available. Available scores included percentile ranks, grade equivalents, and normal curve equivalents. From technical manuals, all scores were converted to comparable normalized percentiles for the summary provided in

Table 2. Because percentile scales have unequal units, these scores were then converted to normalized standard scores prior to further analyses (Anastasi, 1988).

Insert Table 2 about here

ANOVA performed on the extant vocabulary and reading comprehension scores showed no significant differences among grades at either the Jr. or Sr. High School levels. Therefore, for Table 2 and all subsequent analyses, Grades 6-8 and Grades 9-12 were grouped together. Table 2 shows median scores around the 20th to 24th percentiles for all students but those enrolled in Sr. High special education.

Instrumentation

Students were assessed through four procedures, all based on the three most reliable passages: ("The Heart", "The Skeletal System", and "Igneous Rocks"): (a) concept comparison (CC) rating tasks, (b) Maze (multiple choice cloze) tasks, (c) sets of 10 multiple choice questions, and (d) oral reading fluency.

Concept comparison (CC) tasks. Three of the CC tasks completed by teachers were also completed by students. Each CC task consisted of twenty-eight ratings of concept pairs drawn from a passage. Ratings were performed on a four-point scale to indicate the perceived relatedness of each pair of concepts, how much the two concepts "had to do with each other" (see Figure 2).

Maze tests. Multiple choice cloze tests (Howell & Kaplan, 1980) were produced from each passage. Every sixth word was omitted from all but the first and last sentences of the text. The omitted words (approximately 35 per passage) formed the pool or universe from which distractors were selected, with replacement. Distractors were excluded if they were both syntactically and semantically sensible within the sentence. For each deletion in the text, students selected one of five options.

Multiple choice questions. A set of ten, four-option multiple choice questions was developed for each passage. One was a "main idea" question, and the other nine required recognition of important facts and relationships selected consensually from the text by two experienced reading teachers. With the exception of the main idea question, only text-explicit questions were included.

Oral reading fluency. Assigned students also orally read an entire passage while bε.ng audio-taped at the back of the classroom. Tapes were later scored for oral reading error counts and for lapsed time, in order to calculate oral reading fluency—rate of words reao correctly per minute.

Procedure

CC assessment was conducted in two stages, approximately one month apart. Both stages followed a pre-postest control group design, with random assignment of groups to treatment conditions. At the first stage, two CC tasks were assigned to Jr. and Sr. High schools, respectively: "The Heart" (Fry readability 3.8), and "Skeletal and Muscular System" (Fry readability 5.4). During the second stage students were reassigned to treatment and control groups, and students at both levels received the same passage, "Igneous Rocks". The treatment group was administered a maze te:' immediately after the pre-reading CC test, and completed multiple choice and oral reading fluency tests following the post-reading CC test. Design elements are summarized in Table 3.

Insert Table 3 about here

Stage 1. On day 1 of the first stage, during reading/language arts classes, teachers demonstrated the CC task, from scripted instructions. Students then were asked to complete the CC test for the passage at their grade level. Fiftee' .nutes were allowed for the test, though all but a few students finished before 10 minutes.

On day 2, each student was randomly presented with one of two text passages for silent reading--either related or unrelated to the concept comparisons completed the previous day. The two passages were handed out to students in alternating order, according to classroom seating. The unrelated passages were from the s me science texts, had not been previously studied or read, and were of similar readability levels as the test-related passages. There was no discussion or instruction of passage content either before or after the reading. Immediately after reading, each student returned the passage to the teacher, and then completed the post-reading CC test.

Stage 2. Approximately one month later, the research team returned to the school district for a replication and expansion of the Stage 1 design, conducted over a four-day period. This design entailed student reassignment to treatment (n = 43) and control (n = 49) groups (again by classroom seating). On day 1, all students completed pre-reading CCs based on the same passage, "Igneous Rocks". Immediately afterwards, students completed Maze tests within a 25 minute set limit. For both groups of students, the Maze test was constructed from the passage they would read on day 2. The Maze test was administered to the control group to control for possible Maze influence on the post-reading CCs.

On day 2 all students in the treatment group (n = 43) silently read the related passage, "Igneous Rocks", and control group students (n = 49) read an unrelated passage of similar readability from the same text. Immediately afterward, all students completed the post-reading CC test for "Igneous Rocks". Students in the treatment group then also completed a 10-item multiple-choice test on the passage. On days 3 - 5 each student in the treatment group also read the "Igneous Rocks" passage into a tape recorder at the back of the room. The uneven quality of audio recordings reduced the number of useable oral reading samples to 38.

Data Analysis

The first analysis consisted of a three-way ANOVA conducted for each of the three passages: "The Heart", "The Skeletal and Muscular System", and "Igneous Rocks". Two between-subject variables were included, each with two levels: Reading passage (Related, Unrelated), and Program (Special Education, Chapter 1). The within-subject variable was the repeated measure, Time of CC administration (Pre, Post). The dependent measure was the correlation coefficient between student and expert CC scores. In order to analyze Pearson *r*'s as test scores within ANOVA, they were first transformed to Fisher Z scores (Hays, 1981). A significant "Reading passage x Time" interaction was hypothesized, with smaller main effects for the two variables. No significant main effects or interactions were hypothesized for Program. As a secondary analysis, for only those students who read the related passage, pretest and post-test CC expert correlations were tested for significant differences with the Hotelling-Williams Test of correlation equality (Darlington & Carlson, 1987).

The second major analysis was the intercorrelation among scores from (a) pre- and post-reading concept comparisons (Fisher Z scores), (b) standardized Reading Comprehension and Vocabulary tests, (c) Maze tests, (d) Multiple choice tests, and (e) Oral reading fluency samples. It was hypothesized that Post-reading CC scores would be significantly correlated with the other measures, while pre-reading scores would not be. These analyses were conducted to support the validity of the CC scores and maps. In particular, spatial maps appear to hold the potential for diagnosing students' understandings and misinterpretations of text, and planning relevant remedial instruction. To reinforce this priority, qualitative analyses of students' maps are presented first, then quantitative results.

Results

Qualitative Interpretation of Students Maps

Maps of two students, Alice and Bob, with typical CC pre- reading (.12, -.09) and post-reading (.46, .39) correlations (with expert maps) are presented in Appendix C. Agreement with the expert map of "The Heart" was measured by interpoint map distances (Kendall Tau-B), and on clusterings (Omega transform of Rand's statistic). For Alice's pre-reading map, $\underline{\tau - b}$ = .08, and $\underline{\Omega}$ = .27. Her post-reading map showed $\underline{\tau - b}$ = .36, and $\underline{\Omega}$ = .61. For Bob's pre-reading map, , $\underline{\tau - b}$ = .12, and $\underline{\Omega}$ = .33. For his post-reading map, $\underline{\tau - b}$ = .40, and $\underline{\Omega}$ = .74.

Alice's and Bob's maps can be qualitatively interpreted by comparing (a) their pre- and post-reading maps, and (b) their maps with expert teacher maps. Interpretations can be based on either the map distances among individual concepts or membership of outlined clusters. Both the average teacher map for "The Heart" (Figure 1 or Appendix B) and Alice's pre-reading map (Appendix C) suggest a three-cluster interpretation. The expert map yields two clusters, interpretable as (a) "composition and basic movement", and (b) "main parts and connector", with an "external part" as an outlier. These clusters are higher-order or superordinate concepts. Alice's pre-reading map configuration does not include those higher-order concepts. Instead, one large cluster exists, which is difficult to interpret beyond "everything but cardiac and tissue". In Alice's pre-reading map, "cardiac" and "tissue" are outliers, although the first term is used to describe the second in the passage.

By attending to inter-concept distances rather than only cluster membership, we can conduct a more micro-level analysis of Alice's pre-reading map. Within Alice's large cluster, "artery" is on the cluster periphery; it is also isolated on the expert map. However the close proximity of "contracts" and "ventricle" is difficult to explain, and best attributed to student misunderstanding. It is possible that such an uninterpretable relationship was due to random CC task ratings. However, random ratings are not indicated by the systematic relationship described later between pre- and post-reading CC scores.

Alice's post-reading map more closely approximates the expert teacher map in that "cardiac" and "tissue" are clustered apart from other concepts. In addition, the post-reading cluster of "valve", "chamber", and "ventricle" approximates the expert teacher "Main Parts and Connector" cluster (minus "atrium"). From the pre- to post-reading map, "contracts" has shifted from a central, integrated position to an isolated position. Even in this isolated position, it is in the vicinity of the "cardiac", "tissue" cluster, however. Note that "cardiac", "tissue", and "contracts" make up the "Composition and Basic Movement" cluster on the expert map. In summary, student Alice's post-reading map shows greater differentiation of concepts toward interpretable, higher-order clusters.

While changes in Alice's map more closely approximate the expert map, two post-reading map features imply comprehension problems. First, the "artery"-"atrium" connection is not easily interpreted; "atrium" should be closely associated with "ventricle" and "chamber". Second, the proximity of "contracts" with the "artery"-"atrium" cluster is not easily interpreted. Both problems could be clarified and confirmed in a student interview. A diagnostic interview would be especially useful when the purpose of assessment is to diagnosis misunderstandings and/or plan remedial instruction.

The main similarity between Bob's pre-reading map (Appendix C) and the expert map for "The Heart" is that "cardiac" and "contracts" are clustered together and separated from the other concepts. The two other pre-reading map clusters are, however, difficult to explain; each has a member ("tissue" and "artery", respectively) which appears semantically less related to the other two cluster members.

Bob's post-reading map more closely approximates the expert teacher map in that the two ill-fitting cluster members ("tissue" and "artery") have drifted away, and the remaining four concepts have become realigned to form the "Main Parts & Connector" cluster. In drifting away, "artery", "cardiac", and "tissue" have formed a cluster which is difficult to interpret. However, "artery" is clearly the outlying member of that cluster. The main comprehension problem implied by the post-reading map is the isolation of "contracts"—the failure to recognize its close relationship to "cardiac" and "tissue". Again, an interview with the student over the map would help confirm the interpretations made on the basis of clustering and inter-concept distances.

In summary, the comprehension problems inferred from student Bob's post-reading map appear less severe than those of Alice. Alice's most fundamental misunderstanding appears to be a confused "artery"-"atrium" connection), while Bob's shows a less central definitional problem—a misunderstanding of the "cardiac"-"tissue" relationship. Indices of expert agreement for post-reading maps based on inter-concept distances (t - b) and cluster membership (W) show that Bob (t - b = .40, W = .74) slightly outperfomed Alice (t - b = .36, W =.61). These same indices indicate that both students made similar gains from their pre- to post-reading maps.

MDS maps are worth interpreting only if the maps are reasonable stable, and show systematic differences between good and poor reading comprehenders. These qualitative interpretations are therefore be supported by quantitative analyses from control-group designs, with representative sampling of teachers and students. Results from Stage 1 and 2 help answer the question of instrument sensitivity, while results from Stage 2 address the question of criterion-related validity.

Sensitivity of Concept Comparison Scores

Sensitivity of CC scores was defined as systematic changes from Pre- to Post-reading scores by disabled readers who received no preteaching or other assistance. The systematic changes hypothesized were toward closer agreement with the expert teacher CC scores. Results from three-way ANOVA are presented for Jr. high ("The Heart") in Table 4, for Sr. high ("The Skeletal and Muscular System") in Table 5, and for both levels together ("Igneous Rocks") in Table 6.

Insert Tables 4, 5, & 6 about here

Table 4 presents main effects and interactions for the three variables in accounting for Jr. High CC scores on "The Heart". Strength of relationship is indicated by the generalized correlation coefficient, η ("eta") (Hays, 1981). Two of the first order interactions were significant, accounting for 45% (Time x Read.) and 10% (Time x Prog.) of the total variance, respectively. Although interpretation of main effects can be deceptive in the presence of significant interactions, one comparison stands out. The main effect for Time is much larger (74% of the variance) than that for Read. (10% of the variance), although we would hypothesize only a medium-small effect for both. This difference can be explained by the tendency by *all* students to slightly improve in their CC scores at Post-testing (the Time variable), presumably due to practice affect (as will be noted in Table 7).

Tabled ANOVA Results for Sr. high on "The Skeletal and Muscular System" were similar to those for Jr. high, and consistent with hypotheses (see Table 5). At the Sr. High, only one of the three first-order interactions was significant—"Time x Read." (41% of the variance). Both Time and Read. main effects were again significant, with the much larger effect for Time (66% of the variance). The variable, Prog., did not contribute significantly.

The replication study in stage two, with Jr. and Sr. High students together ("Igneous Rocks"), produced results similar to the previous two analyses (see Table 6). The two Time-related interactions were significant, but only "Time x Read." produced a sizeable effect (37% of the total variance, compared to only 7% for "Time x Prog."). Again, Time and Read. produced significant main effects, although only the former was large (66% of the variance). Plots for the three most significant interactions ($p < .01$) are presented in Figure 6. For the plots, the Fisher Z scores used in ANOVA were re-converted to Pearson r's.

Insert Figure 5 about here

The three very similar interaction plots indicate that at both Jr. and Sr. High, students who read the related passage made significantly greater gains in CC scores than did those who read unrelated passages, regardless of the type of special program enrollment.

Table 7 presents CC means and SDs for the three passages. For students reading the related passages, Mean Pearson r's were .07 to .10 before reading, and .36 to .47 after reading.

Insert Table 7 about here

Although the ANOVA s discussed above provided pre- and post-reading CC score comparisons at the group level, they do not provide information at the individual student level. Individual-level results are essential if when individual diagnosis or placement decisions may follow from the test results. Therefore, for only those students who read the "related" passages, the null hypothesis of no significant difference between pretest and post-test CC correlations with the expert scores was tested. The Hotelling-Williams Test of the equality of dependent Pearson correlations ($r_{12} = r_{13}$) was used to compare pretest-expert and posttest-expert correlations (Darlington & Carlson, 1987).

For only 6 of the 97 treatment group students were pretest—expert correlations stronger than post-test—expert correlations, and none of these differences was statistically significant. In contrast, post-test—expert correlations were greater for 91 of the 97 students, and 36 of the Hotelling-Williams Z scores were statistically significant at $p < .05$. Out of 97 score comparisons a number of significant pairs would be expected by chance alone, so a Chi-square test was performed on the proportion of significant versus non-significant findings. The resulting coefficient was highly significant: c^2 (1, N = 97) = 84.75, $p < .0001$.

Criterion-Related Validity

The second major analysis was comparison of pre- and post-reading CC scores of Phase 1 I treatment group students (those who read the related passage) with external measures of reading comprehension. Table 8 contains descriptive information on the CC scores, published Standardized Tests, Maze tests, Multiple choice tests, and Oral reading fluency which were intercorrelated.

Insert Table 8 about here

Table 8 clearly demonstrates the degree of students' reading disabilities. They averaged only 50% correct on the Multiple choice test, and only 66% correct on the Maze (80-90% is an average score). It was hypothesized that post-reading CC scores would be substantially related with the other measures, unlike pre-reading scores. The correlation matrix in Figure 10 shows small, non-significant relationships between the pre-reading CC scores and external measures.

Insert Table 9 about here

Pre-reading CC scores are significantly correlated only with their post-reading counterparts. In contrast, post-reading CC scores show significant, moderate size relationships with the Maze (r = .61), Oral Reading Fluency (r = .57), and the Multiple Choice Test (r = .45)—all based on the same passage. Of the two standardized reading tests, only Vocabulary was significantly related to other measures—the Maze (r = .43) and Oral Reading Fluency (r = .45).

To identify clusters and outliers in the correlation matrix, Ward's hierarchical clustering algorithm was applied (Ward, 1963; Blashfield, 1980) (see Figure 6).

Insert Figure 6 about here

The cluster tree indicates the relative isolation of the pre-reading CC scores and the two standardized test scores. Post-reading CC scores cluster with oral reading fluency, and then with the other two passage-based measures, the Maze and multiple choice test.

Discussion

This study investigated the reliability, sensitivity, and criterion-related validity of concept comparison (CC) scores and spatial maps for assessing content-area reading comprehension of Junior and Senior high school students with reading disabilities. This method offers several advantages sought by reading researchers: (a) reading comprehension can be measured as change from pre-reading schema to post-reading semantic structures, (b) the same metric can be used for both the information structure of text and the knowledge structure of the reader, (c) the maps are diagnostic; they encourage interpretation of how the reader is organizing or misorganizing information, (d) the technique permits multiple correct answers from different teacher "experts", (e) rather than isolated factual recall, the network of relationships among concepts is emphasized, (f) the dimensional maps and hierarchical trees are similar to teaching aids in common classroom use.

First, this study demonstrated the interpretability of student pre- and post-reading maps, through use of expert teacher maps as a standard. Two approaches to map interpretation seemed helpful: interpreting concept clusters (and changes in cluster membership), and interpreting inter-concept distances (and shifts in relative positions). A combined approach seems natural. Minimal interpretation of alternative structural views was undertaken. As a consequence, those qualitative interpretations which were made were not forced. The interpretations earn credibility, however, only if the maps are stable and systematically related to other accepted measures.

Besides map interpretability, this study addressed three requisites of any assessment method—reliability, sensitivity, and validity: (a) reliability of expert teacher concept comparison (CC) scores and MDS maps, (b) the sensitivity of CC scores to response changes following relevant reading, and (c) concurrent criterion-related validity: the relationship between CC scores and other reading measures.

The reliability of only the teacher CC scores and maps was directly studied; reliability of student CC scores and maps was not, nor was the stability of teacher scores over time. It appears that CC tests are reactive; pre-testing appeared to systematically influence post-test results in the direction of greater similarity to the expert map.

The question of reliability of expert teacher CC scores and maps requires a qualified answer; six of the eight passages met the minimum .70 to .80 reliability range for "early stages of research on predictor tests", where the main concern is with group differences (Nunnally, 1978, p. 245). None of the CC tests met the .90 to .95 reliability "desirable standard" for individual-level decision making (Nunnally, 1978, p. 246). Three of the eight CC tests exceeded .80 reliability (.81, .81, .87), justifying their use in the second phase of the study.

Reliability indices of MDS map clusterings were weaker. Only two of the Kappa/Kappa Max. ratios were substantial (above .70). However, the implications of this reliability figures for decision making based on a mapping test are not known. Substantially higher CC and map reliabilities would have been obtained if two alternative expert maps had been allowed per passage. That move would have been supported by observations of teachers' disagreements on the main idea of a story. Two "cognitive structures" may be equally defensible, and the potential for accepting alternative expert maps is a strength of this assessment method. Within the constraints of an initial study, however, it was necessary to delete passages rather than allow two alternative expert maps.

To speak of reliability of the CC test and MDS mapping technique in general would be misleading, as reliability clearly depended upon the particular passage. The variation in reliabilities among the eight passages appeared to be largely a function of the key vocabulary words selected. There were no constraints to key word selection; words were not required to conform to one or a few relationships or dimensions, e.g. "physical connection" or "superordination". Absence of selection criteria permitted a greater range of concept relationship interpretations, and a greater variety of maps. In light of the fact that key vocabulary selection was free to vary, the degree of reliability obtained is substantial. The presumed importance of key vocabulary selection to CC test reliability could be empirically studied from the existing data base.

The second major purpose, assessment of treatment validity, can be answered affirmatively , at least at the group level. Students did significantly improve their match with expert CC scores and maps after reading related passages. At the individual level most students (94%) improved their expert agreement from pre- to post-reading CC, but only 37% of the score improvements reached significance. The Hotelling-Williams test of significance depends not only on the intercorrelations among the three CC results (pre-, post-, expert), but on number of ratings -- only 28 for this task. More concept comparisons would have greatly increased the number of significant individual "improvements".

These group and individual treatment validity results were obtained despite the fact that the all students were deficient readers, and none receive pre-teaching or other instruction in the content area passages. Given those facts, the initial evidence on measurement sensitivity for disabled readers who received no instruction is encouraging.

The third research question, assessment of concurrent, criterion-related validity also receives a tentative, affirmative response. As expected, the post-reading CC scores were most closely related with the other three passage-related criterion measures—the Maze, multiple choice test, and oral reading fluency (r = .61, .45, .57). Among these four passage-related measures, the multiple choice test and Maze were most tightly clustered, followed by oral reading fluency and the post-reading CC scores. The pre-reading CC scores, on the other hand, were not significantly related to any measure but their post-reading CC counterparts. Pre-reading CC scores were clear outliers in the clustering of the six reading measures.

The largest matrix correlations were of only low-moderate to moderate size. The moderate reliability of the CC scores may have imposed a ceiling on these validity relationships. Other possible reasons for medium-low validity scores may reside in the external measures, themselves: (a) lack of structural sensitivity (Maze, ORF, Mult. Choice, St. Tests), (b) inability to account for pre-reading knowledge differences (Maze, ORF, St. Tests) , (c) information processing demands appear to differ from reading (Maze, ORF, Mult. Choice, St. Tests), (d) questions unintentionally cuing responses (Maze, ORF, Mult. Choice, St. Tests). Comparing a new measure with deficient standard criterion measures will always result in less than satisfactory validity coefficients.

This study served its purpose as an initial investigation of the reliability and validity of a relatively unresearched assessment approach. However, it raised several questions which need to be addressed before these innovative techniques are used outside an experimental setting. One question is how many different types of relationships among concepts can be plotted on a two-dimensional space while still rendering an interpretable map. Interpretation of the MDS maps intentionally was not based on map dimensions or axes (as in factor analysis), but rather on clustering of, and Euclidean distances within plotted configurations. This approach is legitimized by experts in the MDS field, though not frequently encountered in the literature (Davison, 1983). However, a "problem space" of only two dimensions may tend to limit the variety of relationship-types among concepts and clusters. In that sense, map dimensionality may play a crucial, underlying role in map validity.

Increasing the number of map dimensions in order to less constrain the variety of interpretable relationships is not a practical solution. The small number of concepts plotted would be seriously "over fit" to the higher dimensionality, and solutions would lack stability. The question of a limit on the numbers of types of relationships among concepts has direct bearing on how key vocabulary are initially selected. Map reliability and interpretability need to be studied under different vocabulary selection guidelines.

The diagnostic and instructional utility of MDS maps will hinge in part on evidence that qualitative interpretations have reliability and validity. This study demonstrated qualitative interpretation of a few teacher and student maps without providing such evidence. A logical approach to validating a qualitative map interpretation would be to directly interview a student before and after reading, followed by an evaluation of the maps by the same respondent. The interviews should be open-ended at first; then students could react to their MDS maps.

A second qualitative validation approach might include student selection or free-hand construction of spatial maps. Both approaches could help establish whether the MDS methodology unduly restricts or biases interpretations of cognitive structures. Information from these approaches might also generate new approaches to MDS map interpretation.

Three types of map interpretation were considered, based on cluster membership, relationships among individual concepts, and hierarchical arrangement of concepts. It is not known which type of interpretation could be most readily understood and communicated by reading specialists and teachers. Neither is it known if one method is better suited than another for different types of organization of expository text. Other semantic structure models (e.g. Holley & Dansereau) provide alternative structures for text written with different types of concept organization. Further research is needed on these questions.

Both interpretations based on cluster membership (whether on the map or in a hierarchical tree) rely on secondary hierarchical cluster analysis. Cluster analysis has some notoriety for instability, and has been classified as little more than a heuristic (Aldenderfer & Blashfield, 1984). Considerable agreement was noted between cluster solutions based on Ward and Average linkage algorithms. Other algorithms did not match well, however. The instability of cluster solutions and the complexity of the analysis need to be weighed against the benefits. When cluster definitions are desired on the map (rather than tree diagrams), human judgments may suffice. The ability of teachers to directly interpret map clusters would reduce the time and technical skills required. Reliability studies are needed on this question.

The disagreements obtained among teacher raters raises the question of what constitutes an "expert". Perhaps subject matter experts are required, rather than teachers who are more familiar with the textbooks as teaching tools and with the information their students could reasonably gain from the texts. Content knowledge also plays an unknown role in the interpretation of map clusters and relationships. What level of content knowledge is sufficient?

This study used only eight key vocabulary words per map, whereas most passages yielded at least eight to twelve terms. Eight concepts is a marginal number for scaling in two dimensions; nine or ten would be preferable. The biggest problem in increasing the number of concepts is the geometric increase in the length of the concept comparison task (28 comparisons for 8 concepts, 36 comparisons for 9 concepts, etc.). Incomplete block sampling schemes for reducing the number of necessary comparisons have been researched in Monte Carlo studies (Davison, 1983). Their stability appears to depend heavily on the nature and content of the comparison task. No research was found on incomplete block designs with small numbers of concepts. That type of investigation is urgently needed to help determine the utility of MDS mapping under less controled text conditions.

Despite the many unanswered questions, this study supports the further investigation of spatial maps for assessing reading comprehension. With the technical underpinning of MDS, spatial maps can potentially address several of the deficiencies attributed to most existing reading assessment techniques by increasing numbers of professionals who have adopted a cognitive processing view of reading comprehension. At this point, MDS for reading assessment is suitable mainly as a research tool, requiring technological and statistical expertise. However, concept comparison tests can be efficiently produced and group administered. This fact should encourage serious consideration of the technique for selected reading assessment purposes if other studies further support its reliability, sensitivity, and validity.

References

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. Washington, D.C.: A.P.A.

American Psychological Association, AERA, NCME. (1985). Standards for educational and psychological testing. Washington, D.C.: APA..

Aiken, L.R. (1979). Psychological testing and assessment. Boston, MA: Allyn & Bacon.

Aldenderfer, M. S. & Blashfield, R. K. (1984). Cluster analysis. Beverley Hills, CA: Sage Publications.

Allington, R. L. (1982). The persistence of teacher beliefs in facets of the visual perceptual deficit hypothesis. Elementary School Journal, 82, 351-359.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Anastasi, A. (1976). Psychological testing (3rd ed.), New York: Macmillan.

Anderson, J. R. & Bower, G. H. (1973). Human associative memory. New York: Winston.;

Anderson, R.C. (1977) Schema-directed processes in language comprehension (Tech. Rep. No. 50). Urbana: U. of Illinois, Center for the Study of Reading, July, 1977.

Applied Psychological Measurement, Vol. 7, No. 4, (1983). M. L. Davison & L. E. Jones (Eds.)

Armbruster, B. B., & Anderson, T. H. (1984). Mapping: Representing informative text diagrammatically. In C. Holley & D. Dansereau (Eds.), Spatial Learning Strategies: Techniques, Applications, and related issues, New York: Academic Press, Inc.

Armbruster, B.B. & Anderson, T.H. (1980). The effect of mapping on the free recall of expository text (Tech. Rep. No. 160). Urbana: Univ. of Illinois, Center for the Study of Reading.

Anastasi, A. (1988). Psychological testing (3rd ed.). New York: Macmillan Publishing Company.

Barufaldi, J. P., Ladd, G. T., & Moses, A. J. (Eds.) (1981). Heath science. Lexington, MA: D. C. Heath.

Bayne, R. Beauchamp, J., Begovich, C., & Kane, V. (1980). Monte Carlo comparisons of selected clustering procedures. Pattern Recognition, 12, 51-62.

Beaugrande, R. (1980). Text, discourse, and process. Norwood, N.J.: Ablex.

Bisanz, G., LaPorte, R., Vesonder, G., & Voss, J. F. (1978). On the representation of prose in memory: A multidimensional approach. Journal of Verbal Learning and Verbal Behavior.

Blashfield, R. K. (1980). The growth of cluster analysis: Tryon, Ward, and Johnson. Multivariate Behavioral Research, 15, 439-458.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.;

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City: ACT Publications

Brennan, R. L., & Prediger, D. LJ. (1981). Coefficient Kappa: Some uses, misuses and alternatives. Educational and Psychological Measurement, 41, 687-699.

Brewer, W. F. (1987). Schemas versus mental models in human memory (187-197). in P. Morris (Ed.) Modeling Cognition. Ney York: John Wiley and Sons.

Bridge, C. & Tierney, R. (1981). The inferential operations of children across text with narrative and expository tendencies. Journal of Reading Behavior, 13(3), 201-214.

Brown, F.G. (1976). Principles of educational and psychological testing (2nd. ed.). NY: Holt, Rinehart & Winston.

Brown, L. T., & Stanners, R. F. (1984). The assessment and modification of concept interrelationships. Journal of Experimental Education-----

Calfee, R., & Drum, P. (1986). Research on teaching reading. In M.C. Wittrock (Ed.), Handbook of Research on Teaching (3rd. ed.). NY: MacMillan Publishing Co.

Carrol, J. & Arabie, P. (1980). Multidimensional scaling. Annual Review of Psychology, 31, 607-49.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213-220.

Cohen, H. & Davison, M. (1973). Jiffy-scale: A FORTRAN IV program for generating Ross-ordered pair comparisons. Behavioral Science, 18, 76.

Cormack, R. (1971). A review of classification. Journal of the Royal Statistical Society (Series A) 134, 321-367.

Coxon, A. (1982). The users guide to multidimensional scaling. Exeter, NH: Heineman Educational Books.

Cronbach, L.J., Gleser, C.C., Nanda, H., & Rajaratnam, H. (1972). The dependability of behavioral measures. NY: Wiley.

Curtis, M.E. & Glaser, R. (1983). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20(2), 133-147.

Dansereau, D. F., McDonald, B. A., Collins, K. W., Garland, J. C., Holley, C. D., Diekhoff, G. M., & Evans, S. H., (1979). Evaluation of a learning strategy system. In H. F. O'Neil, Jr., & C. D. Spielberger (Eds.), Cognitive and affective learning strategies. New York: Academic Press.

Darlington, R. B., & Carlson, P. M. (1987). Behavioral Statistics: Logic & Methods. NY: The Free Press.

Davison, M. L. (1983). Multidimensional Scaling. New York: John Wiley & Sons.

Davison, M., Richards, P. & Rounds, J. (1986). Multidimensional scaling in counseling research and practice. Journal of Counseling and Development, 65, 178-184.

de Beaugrande R. (1980). Text, discourse, and process. Norwood, N. J.: Ablex.

de Leeuw, J., & Stoop, I. (1984). Upper bounds for Kruskal's stress. Psychometrika, 49, 391-402.

Deikhoff, G. M. (1982). Cognitive maps as a way of presenting the dimensions of comparison within the history of psychology. Teaching of Psychology, 9, 115-116.

Diekhoff, G.M. (1983) Testing through relationship judgments. Journal of Educational Psychology, 75, 2, 227-233.

Everitt, B. S. (1988). Cluster analysis. in J. P. Keeves (Ed.), Educational research, methodology, and measurement: An international handbook. (pp. 247-253). New York: Pergamon Press.

Fenker, R.M. (1975). The organization of conceptual materials: A methodology for measuring ideal and actual cognitive structures. Instructional Science, 4, 33-57.

Fisher, L., & Van Ness, J. W. (1971). Admissable clustering procedures. Biometrika, 58, 91-104.

Fitpatrick, A.R. (1983). The meaning of content validity. Applied Psychological Measurement, 7(1), 3-13.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. New York: John Wiley & Sons.

Frederiksen, C. H. (1975), Acquisition of semantic information from discourse: Effects of repeated exposures. Journal of Verbal Learning and Verbal Behavior, 14, 158-169.

Frederiksen, C. H. (1977), Semantic processing units in understanding text. In R. O. Freedle (Ed.), Discourse production and comprehension. Norwood, N. J.: Ablex.

Frederiksen, C.H. (1979). Discourse comprehension and early reading. In L.B. Resnick & P. A. Weaver (Eds.), Theory and practice of early reading (Vol. 1). Hillsdale, N.J.: Erlbaum.

Freedle, R. O. (1979). Advances in discourse processes, Vol. 2, New directions in discourse processing. Norwood, N.J.: Ablex.

Geva, E. (1980). Meta textual notions and reading comprehension. Unpublished doctoral dissertation, U. of Toronto.

Geva, E. (1983). Facilitating reading comprehension through flowcharting. Reading Research Quarterly, 18(4), 385-406.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Graef, J., & Spence, I. (1979). Using distance information in the design of large multidimensional scaling experiments. Psychological Bulletin, 86, 60-66.

Griffeth, R., Hom, P., DeNisi, A., & Kirchner, W. (1985). A comparison of different methods of clustering countries on the basis of employee attitudes. Human Relations, 38, 813-340.

Guion, R.M. (1978). Scoring of content domain samples. Journal of Applied Psychology, 63, 499-506.

Hagus, G. P., Reque, B. R., & Wilson, R. H. (Eds.) (1985). Heath social studies. Lexington, MA: D. C. Heath.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press.

Hauf, M. B. (1971). Mapping: A technique for translating reading into thinking. Journal of Reading, 14, 225-230.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.

Heimlich, J. E. & Pittelman, S. D. (1986). Semantic applications: Classroom applications. Newark, Delaware: International Reading Association.

Holley, C. D. & Dansereau, D. F. (1984), The development of spatial learning strategies. In C. Holley and D. Dansereau (Eds.) Spatial learning strategies: Techniques, applications, and related issues, New York: Academic Press.

Howell, K. W. & Kaplan, J. S. (1980). Diagnosing basic skills. Columbus, OH: Charles E Merrill.

Johnson, S. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241-254.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. Cognitive Science, 4, 71-115.

Johnson-Laird, P. N. (1983). Mental models. Cambridge, MASS: University Press.

Johnson-Laird, P. N. & Wason, P. C. (1977) Thinking: Readings in Cognitive Science, New York: Cambridge University Press.

Johnston, P. H. (1984). Assessment in reading. In P. D. Pearson (Ed.) Handbook of Reading Research, New York: Longman.

Kavale, K. (1981). Functions of the Illinois Test of Psycholinguistic Abilities (ITPA): Are they trainable? Exceptional Children, 47, 496-513.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kirsch, I.S. & Guthrie, J.T. (1980). Construct validity of functional reading tests. Journal of Educational Measurement, 17(2), 81-93.

Kruskal, J. B. & Wish, M. (1978). Multidimensional scaling. Beverley Hills, CA: Sage Publications.

LaPorte, R.E. & Voss, J.F. (1979). Prose representation: A multidimensional scaling approach. Multivariate Behavioral Research, 14, 39-56.

Levy, P. (1987). Modelling cognition: Some current issues. (pp. 3-20) in P. Morris (Ed.) Modeling Cognition, Ney York: John Wiley and Sons.

Mandler, J.M. & Johnson, N.S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111-151.

Messick, S. (1980). Test validity and ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. Educational Researcher, 10, 9-20.

Messick, S. (1989). Validity (pp. 13-105). In R. L. Linn (Ed.), Educational Measurement (Third Edition), New York: American Cour il on Education & Macmillan Publishing Company.

Meyer, B.J.F. (1975). The organization of prose and its effects on memory. Amsterdam: North-Holland Publishing Co.

Meyer, B.J.F., & Rice, G.E. (1984). The structure of text (Chapt. 11). In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.) Handbook of Reading Research. NY: Longman.

Meyer, B. F., Brandt, P. M., & Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension in ninth-grade students. Reading Research Quarterly, 16, 72-103

Milligan, G. (1980). An examination of the affect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325-342.

Milligan, G. (1981). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 379-407.

Millward, R. B. (1985). Mind your mental models. Journal of Psycholinguistic Research,14 (5), 427-446.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules—an evaluation. Computer Journal, 20, 359-363.

Morey, L. C. & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. Educational and Psychological Measurement, 44, 33-37.

Niles, O.S. (1965). Organization perceived. In H.H. Herber (Ed.), Perspectives in reading: Developing study skills in secondary schools. Newark, Del.: International Reading Association.

Nunnally, J. C. (1978). Psychometric theory. NY: McGraw-Hill Book Company.

Preece, (1976). Mapping cognitive structure: A comparison of methods. Journal of Educational Psychology. 68(1), 1-8.

Psychometrika, Vol. 51, No. 1, (1986).

Ramsey, W. L., Gabriel, L. A., McGuirk, J. F., Phillips, C. R., & Watenpaugh, F. M. (Eds.) (1985). Holt [illegible]eral science. New York: Holt, Rinehart & Winston.

Ramsey, W. L., Gabriel, L. A., McGuirk, J. F., Phillips, C. R., & Watenpaugh, F. M. (Eds.) (1985). Holt life science. New York: Holt, Rinehart & Winston.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Raphael, T.E., Englert, C.S., & Kirschner, B.W. (1986). The impact of text structure instruction and social context on students' comprehension and production of expository text. (Research Series No. 177), East Lansing, Mich: The Institute for Research on Teaching, Michigan State Univ.

Reutzel, D.R. (1986). Investigating a synthesized comprehension instructional strategy: The cloze story map. Journal of Educational Research, 79(6), 343-349.

Richardson, J. T. E. (1983) Mental imagery in thinking and problem solving. (197-226) in J. Evans, Ed., Thinking and reasoning: Psychological approaches. Boston: Routledge & Kegan Paul.

Rumelhart, D.E. & Ortony, A. (1977). The representation of knowledge in memory. In R.C. Anderson, R.J. Spiro, & W.E. Montague (Eds.), pp. 99-135.

Rumelhart, D.E. (1975). Notes on a schema for stories. In D.G. Bobrow & A.M. Collins (Eds.), Representation and understanding. New York: Academic Press.

Rumelhart, D. E., Lindsay, P., & Norman, D. (1972). A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), Organization of memory. New York: Academic Press.

Schan[illegible], R.C. & Abelson, R.P. (1977). Scripts, plans, goals, and understanding. Hillsdale, N.J.: Erlbaum.

Schiffman, S., Reynolds, M. & Young, F. (1981). Introduction to multidimensional scaling: Theory, methods and applications. San Francisco, CA: Academic Press, Inc.

Schwartz, R.M. (1984). Measuring Reading Competence: A theoretical-prescriptive approach. NY: Plenum Press.

Shavelson, R.J. (1974). Methods for examining representations of a subject matter structure in a student's memory. Journal of Research in Science Teaching, 11, 231-249.

Shepard, R. N., Kilpatrick, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. Cognitive Psychology, 7, 82-138.

Sinatra, R.C., Stahl-Gemake, J. & Morgan, N.W. (1986). Using semantic mapping after reading to organize and write original discourse. Journal of Reading. 30(1), 2-13.

Sneath, P. & Sokal, R. (1973). Numerical taxonomy. San Francisco, CA: Freeman.

Spence, I. & Domoney, D. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. Psychometrika, 39, 469-490.

Spence, I. (1982). Incomplete experimental designs for multidimensional scaling. In R. G. Golledge & J. N. Rayner (Eds.), Proximity and preference: Problems in the multidimensional analysis of large data sets. Minneapolis: University of Minnesota Press.

Spence, I. (1983). Monte Carlo Simulation Studies. Applied Psychological Measurement, 7, 405-425.

Spiro, R.J. (1977). Remembering information from text: The state of the schema approach. In R.C. Anderson & W.E. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, N.J.: Erlbaum.

Stanners, R.F., Brown, L.T., Price, J. M., & Holmes, M. (1983). Concept comparisons, essay examinations, and conceptual knowledge. Journal of Educational Psychology, 75, 6, 857-864.

Stanners, R.F., Price, J.M., & Painton, S. (1982). Interrelationships among text elements in fictional prose. Applied Psycholinguistics, 3, 95-107.

Stasz, C., Shavelson, R.J., Cox, D.L., & Moore, C.A. (1976). Field independence and the structuring of knowledge in a social studies minicourse. Journal of Educational Psychology, 68, 550-558.

Stein, N.L. & Glenn, C.G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), New directions in discourse processing. Norwood, N.J.: Ablex.

Sternberg, R. J. (1981). Testing and cognitive psychology. American Psychologist, 36, 1181-1189.

Surber, J. R. & Smith, P. L. (1981). Testing for misunderstanding. Educational Psychologist, 16, 163-174.

Surber, J. R. (1984). Mapping as a testing and diagnostic device. In C. Holley & D. Dansereau (Eds.), Spatial Learning Strategies: Techniques, Applications, and related issues, New York: Academic Press, Inc.

Thorndike, R.L. & Hagan, E. (1977). Measurement and evaluation in education and psychology (4th ed.). NY: Wiley.

Thorndyke, P.W. (1977). Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology, 9, 77-110.

Trabasso, T. (1978). Cognitive prerequisites to reading. Paper presented at the meeting of the American Educational Research Association, Toronto, March, 1978.

Valencia, Pearson, & Chapman, (1986). New strategies for reading comprehension assessment--Illinois initiatives. Ill: Center for the Study of Reading.

van Dijk, T. A. & Kintsch, W. (1983). Strategies for discourse comprehension. New York: Academic Press.

Wagener, M. & Wender, K. F. (1985). Spatial representations and inference processes in memory for text (p. 115-136). in G. Rickheit & H. Strohner (Eds.) Inferences in text processing. North-Holland: Elsevier Science Publishers B. V.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Weiner, H. & Kaye, K. (1974). Multidimensional scaling of concept learning in an introductory course. Journal of Educational Psychology, 66, 591-598.

Wilensky, R. (1978). Why John married Mary: Understanding stories involving recurring goals. Cognitive Science, 2, 235-266.

Wilkinson, L. (1989). SYSTAT: The system for statistics. Evanston, IL: SYSTAT, Inc.

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. Reading Research Quarterly, 19, 404-425.

Woodcock, R. W. & Johnson, M. B. (1977). Woodcock -Johnson Psycho-Educational Battery, Part Two: Tests of Achievement, Allen, TX: DLM Teaching Resources.

Jastak & Wilkinson, (1984). The Wide Range Achievement Test—Revised. Wilmington, DE: Jastak Associates, Inc.

Young, F. W. & Lewyckyj, R. (1979). ALSCAL-4 User's Guide. University of North Carolina, Chapel Hill: Data Analysis and Theory Associates

Young, F. (1984). Scaling. Annual review of psychology, 35, 55-81.

Table 1

Agreement Among Five Raters on 28-Item Concept Comparison Tests Based on Eight Passages From Science and Social Studies Basal Texts.

		Cohen's Kappa	
	Intraclass Correlation	K / K.Max[1]	Wt.K[2]
"The Heart"*	.81	.49 / .70 = .71	.60
"Igneous Rocks"*	.81	.51 / .65 = .78	.60
"Population Limits"	.71	.27 / .79 = .35	.43
"One-celled Animals"	.73	.40 / .75 = .54	.49
"The Seashore"	.73	.28 / .59 = .48	.40
"The Skeletal System"*	.87	.47 / .83 = .57	.62
"Soviet Union"	.69	.28 / .71 = .39	.39
"Texas"	.65	.32 / .90 = .36	.47

All Coefficients are significant beyond the .01 level.

*Three most reliable passages selected for Phase II.

[1] The ratio of Kappa to the maximum possible Kappa value for the given table.

[2] Weighted Kappa: linear weights of 0, .25, .50, and 1 are assigned according to degree of discrepancy between raters.

Table 2

Medians and Interquartile Ranges for Normalized Percentiles in Reading Comprehension and Vocabulary for 104 Jr. and Sr. High School Students Served by Chapter I Compensatory and Categorical Special Education Programs.

Jr High (n=53)	Special Ed. (n=20)				Chapt. I (n=33)			
	Reading Comp.		Vocabulary		Read Comp.		Vocabulary	
	Md	IQR*	Md	IQR*	Md	IQR*	Md	IQR*
	20	13	21	13	23	14	19	9
Sr High (n=51)	Special Ed. (n=23)				Chapt. I (n=28)			
	Md	IQR*	Md	IQR*	Md	IQR*	Md	IQR*
	29	20	24	16	20	13	21	10

IQR = Interquartile Range: spread of the middle half of scores clustered about the Median.

Table 3

Design Elements: Observations and Experimental Conditions Across Time by Group.

	0_1	0_2	0_3	Reading:	0_2	0_4	0_5
	Extant	Pre-test		Related: (X_R)	Post-test	Multiple	
	Ach. Scores	CC	Maze	Unrelated: (X_U)	CC	Choice	O.R.F.
Stage 1							
I. Treatment (n=54)							
Jr.: "Heart" (26) Sr.: "Skeltal" (28)	0_1	0_2		X_R	0_2		
II. Control (n=53)							
Jr.: "Heart" (28) Sr.: "Skeltal" (25)	0_1	0_2		X_U	0_2		
Stage 2							
I. Treatment (n=43)							
Jr.& Sr. "Rocks"	0_1	0_2	0_{3R}	X_R	0_2	0_{4R}	0_5
II. Control (n=49)							
Jr.&Sr. "Rocks"	0_1	0_2	0_{3U}	X_U	0_2		

Note:

C C = concept comparisons

O. R. F. = oral reading fluency

Table 4

Three-way AN. 'A for Dependent Variable, "Concept Comparison Scores" , with One Grouping Variable, "Program" , one Experimental Variable, "Reading Passage", and One Repeated Measure, "Time of Assessment", (Jr. High Grade Level: "The Heart" Passage (N=53)).

Source of Variance & (Levels)	SSbt	SSw	F (1,49)	p	n
Between Subject Effects:					
Read. (Related, Unrelated)	.264	2.51	5.16	.03	.31
Prog. (SPED, Chapt. 1)	.55	2.51	10.84	.002	.43
Read. x Prog.	.01	2.51	.24	.63	.07
Within Subjects Effects:					
Time (Pre, Post)	1.55	.53	143.08	.000	.86
Time x Read.	.42	.53	39.14	.000	.67
Time x Prog.	.06	.53	5.65	.02	.32
Time x Read. x Prog.	.003	.53	.27	.61	.07

Table 5

Three-way ANOVA for Dependent Variable, "Concept Comparison Scores" , with One Grouping Variable, "Program" , one Experimental Variable, "Reading Passage", and One Repeated Measure, "Time of Assessment". (Sr. High Grade Level: "The Skeletal System" Passage (N=50)).

Source of Variance & (Levels)	SSbt	SSw	F (1.46)	p	η
Between Subject Effects:					
Read. (Related, Unrelated)	1.12	4.74	10.89	.002	.44
Prog. (SPED, Chapt. 1)	.008	4.74	.08	.78	.04
Read. x Prog.	.04	4.74	.39	.53	.09
Within Subjects Effects:					
Time (Pre, Post)	2.34	1.23	86.93	.000	.81
Time x Read.	.839	1.23	31.14	.000	.64
Time x Prog.	.032	1.23	1.17	.28	.16
Time x Read. x Prog.	.024	1.23	.88	.35	.14

Table 6

Three-way ANOVA for Dependent Variable, "Concept Comparison Scores" , with One Grouping Variable, "Program" , one Experimental Variable, "Reading Passage", and One Repeated Measure, "Time of Assessment", Sr. Hign () and Jr. High () Grade Levels, "Igneous Rocks" Passage.

Source of Variance & (Levels)	SSbt	SSw	F (1,88)	p	η
Between Subject Effects:					
Read. (Related, Unrelated)	.707	5.16	12.07	.001	.35
Prog. (SPED, Chapt. 1)	.000	5.16	.006	.94	0.0
Read. x Prog.	.102	5.16	1.74	.19	.14
Within Subjects Effects:					
Time (Pre, Post)	1.71	.927	162.39	.000	.81
Time x Read.	.537	.927	50.98	.000	.61
Time x Prog.	.068	.927	6.43	.01	.26
Time x Read. x Prog.	.05	.927	4.74	.03	.23

Table 8.

Pre- and Post-Reading Concept Comparison Scores (Pearson r's) with Reading of Related or Unrelated Passage

"The Heart": Jr. High

	Special Ed. (n=20)				Chapt. I (n=33)				Total (n=53)			
	Pre		Post		Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Related	-.046	.151	.367	.153	.138	.167	.419	.160	.074	.182	.401	.157
Un-Related	-.029	.065	.127	.137	.176	.132	.243	.197	.095	.15	.196	.181

"The Skeletal and Muscular System": Sr. High

	Special Ed. (n=23)				Chapt. I (n=28)				Total (n=51)			
	Pre		Post		Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Related	.068	.20	.434	.273	.088	.25	.512	.21	.078	.198	.474	.24
Un-Related	.076	.139	.165	.251	.041	.154	.199	.222	.056	.146	.184	.23

"Igneous Rocks": Jr. and Sr. High

	Special Ed. (n=31)				Chapt. I (n=60)				Total (n=91)			
	Pre		Post		Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Related	.082	.21	.428	.204	.107	.173	.332	.183	.1	.181	.357	.192
Un-Related	.049	.192	.143	.149	.106	.136	.188	.146	.083	.162	.169	.147

Table 8

Descriptive Data for Pre- and Postest CC, Published Standardized Group Reading Tests, Maze Tests, Multiple Choice Tests, and Oral Reading Fluency (n = 38).

Test	Min.	M	Max	SD
Pre-Reading CC (Pearson r)	- .23	.09	.24	.31
Post-Reading CC (Pearson r)	.11	.51	.79	.26
Std. Reading (percentile)	1.	20.	62.	14.8
Std. Vocabulary (percentile)	1.	19.	60.	13.8
Maze (percent correct)	17.	66.	92.	24.0
Multiple Choice (percent correct)	20.	53.	80.	18.0
Oral Reading Fluency (wcpm)	22.	91.	146.	29.7

Table 9.

Correlation of Pre-Reading and Post-Reading Concept Comparisons with Five Criteria: the Maze, Multiple Choice, Oral Reading Fluency, and Standardized Reading and Vocabulary Tests (N = 39)

	Pre C.C.	Post C.C.	Maze	M. Choice	Read. Std.	Vocab. Std.
Post C.C.	.42*	•				
Maze	.28	.61*	•			
M.Choice	.15	.45*	.75*	•		
Read. Std.	.21	.36	.36	.38	•	
Vocab. Std.	.19	.38	.43*	.38	.66*	•
O.R.F.	.15	.57*	.60*	.51*	.37	.45*

* p < .01

Figure Caption

Figure 1. Interpretation of Concept Clusters and Concept Relationships on an MDS Map.

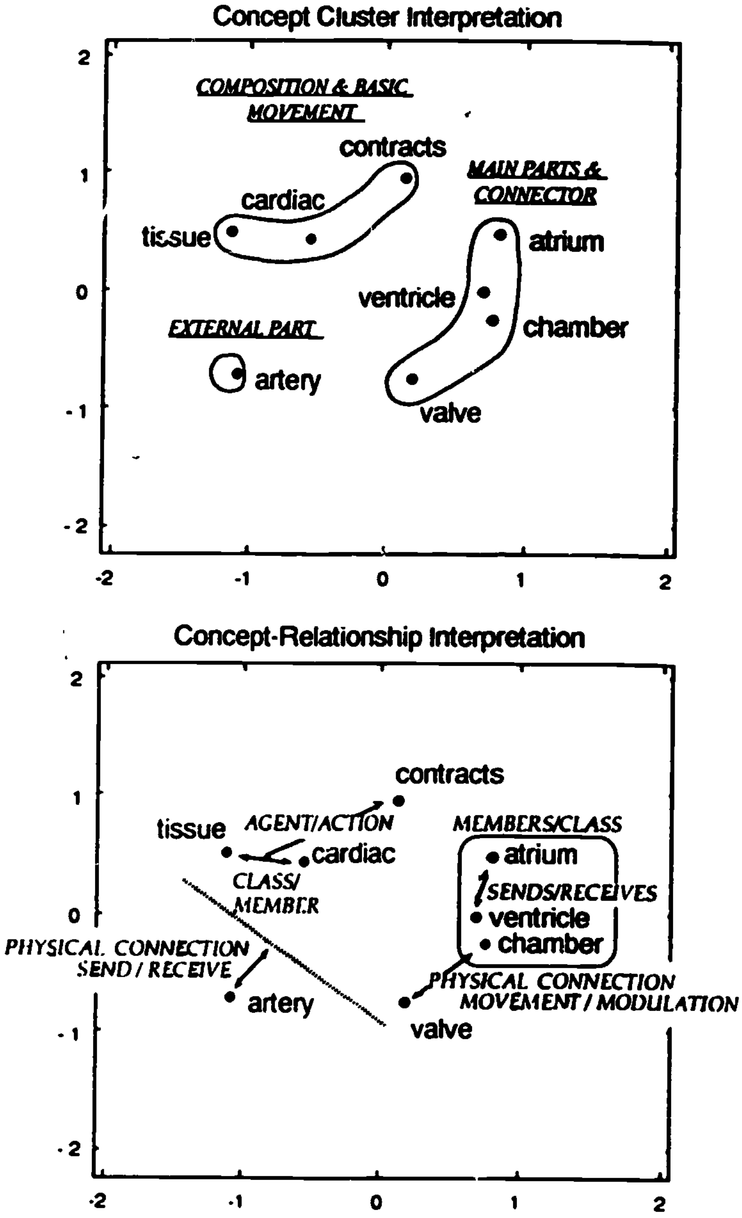


Figure Caption

Figure 2. Concept Comparison Task for Multidimensional Scaling Input.

Student:______________ Grade______ School____________

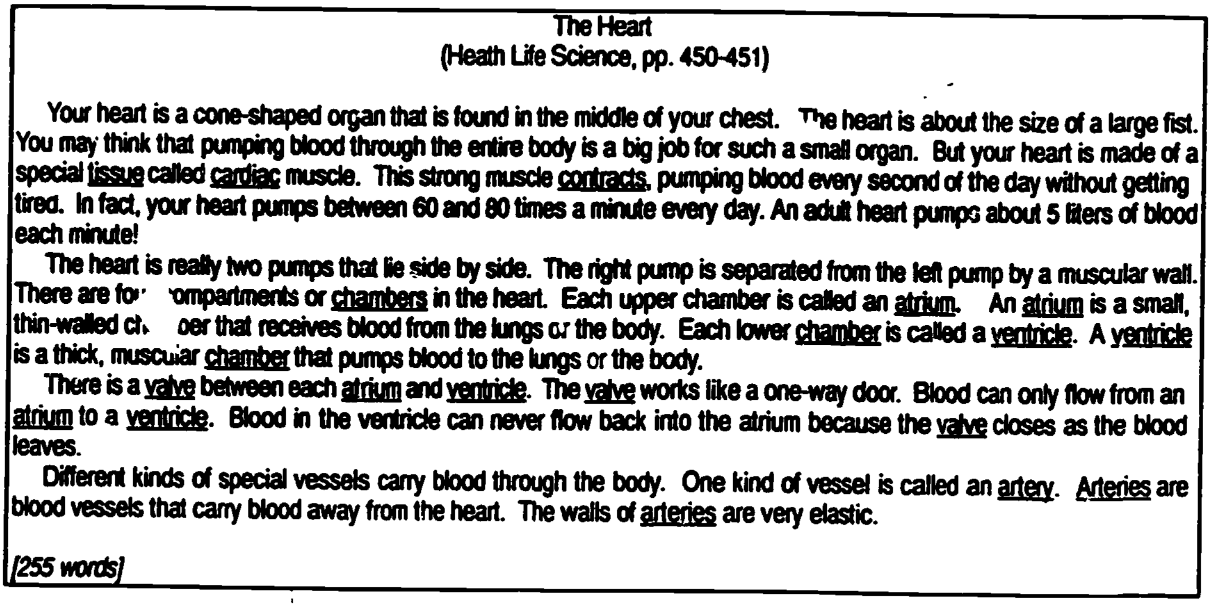
Teacher:______ ______________ Date:__ /____/____/

Passage: The Heart

	CLOSE RELATION			LITTLE OR NO RELATION
	4	3	2	1
atrium - cardiac				
tissue - cardiac				
tissue - chamber				
valve - chamber				
contracts - ventricle				
valve - tissue				
valve - cardiac				
artery - chamber				
ventricle - tissue				
artery - contracts				
ventricle - chamber				
cardiac - chamber				
ventricle - valve				
contracts - tissue				
atrium - valve				
contracts - cardiac				
atrium - ventricle				
contracts - valve				
atrium - tissue				
atrium - contracts				
artery - tissue				
artery - cardiac				
ventricle - cardiac				
artery - valve				
artery - ventricle				
contracts - chamber				
atrium - chamber				
artery - atrium				

Figure Caption

Figure 3. "The Heart" Science Text Passage with Underlined Key Vocabulary Words.



The Heart
(Heath Life Science, pp. 450-451)

Your heart is a cone-shaped organ that is found in the middle of your chest. The heart is about the size of a large fist. You may think that pumping blood through the entire body is a big job for such a small organ. But your heart is made of a special tissue called cardiac muscle. This strong muscle contracts, pumping blood every second of the day without getting tired. In fact, your heart pumps between 60 and 80 times a minute every day. An adult heart pumps about 5 liters of blood each minute!

The heart is really two pumps that lie side by side. The right pump is separated from the left pump by a muscular wall. There are fo' 'ompartments or chambers in the heart. Each upper chamber is called an atrium. An atrium is a small, thin-walled ch oer that receives blood from the lungs or the body. Each lower chamber is called a ventricle. A ventricle is a thick, muscuiar chamber that pumps blood to the lungs or the body.

There is a valve between each atrium and ventricle. The valve works like a one-way door. Blood can only flow from an atrium to a ventricle. Blood in the ventricle can never flow back into the atrium because the valve closes as the blood leaves.

Different kinds of special vessels carry blood through the body. One kind of vessel is called an artery. Arteries are blood vessels that carry blood away from the heart. The walls of arteries are very elastic.

[255 words]

Figure Caption

Figure 4. Hierarchical Cluster Analysis Solution for MDS Map Configuration.

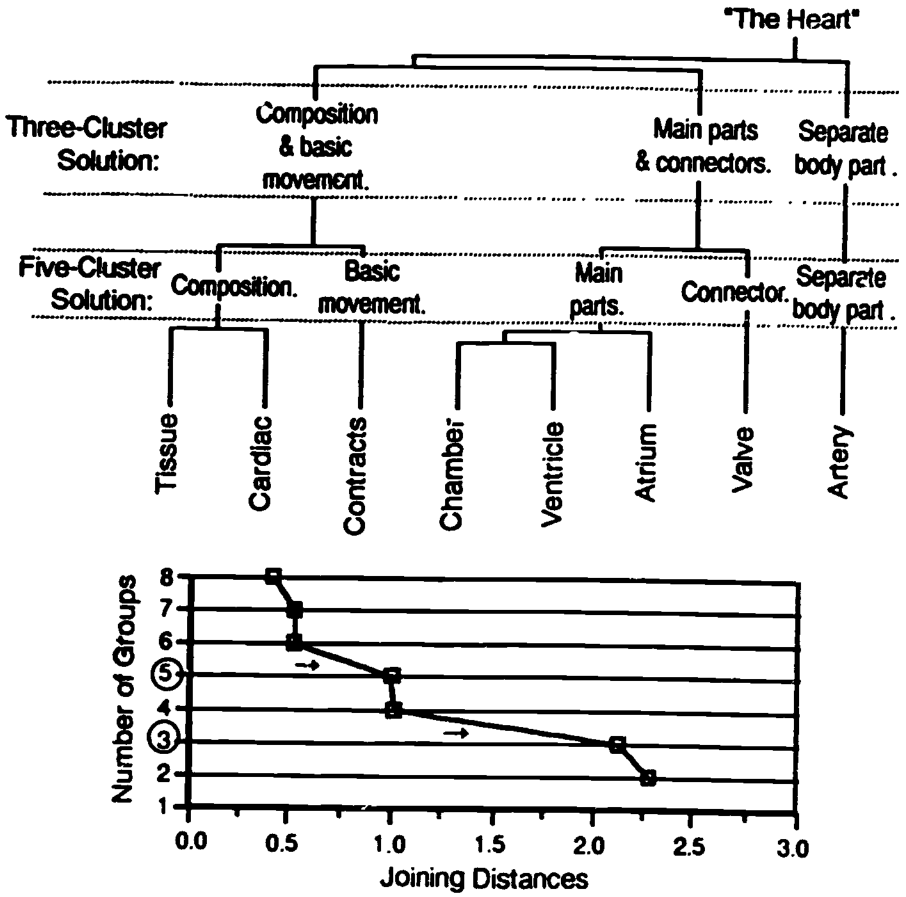


Figure Caption

Figure 5. ANOVA Interaction Plots for "Time" x "Read" for Junior High students, Senior High students, and Combined Grade Levels.

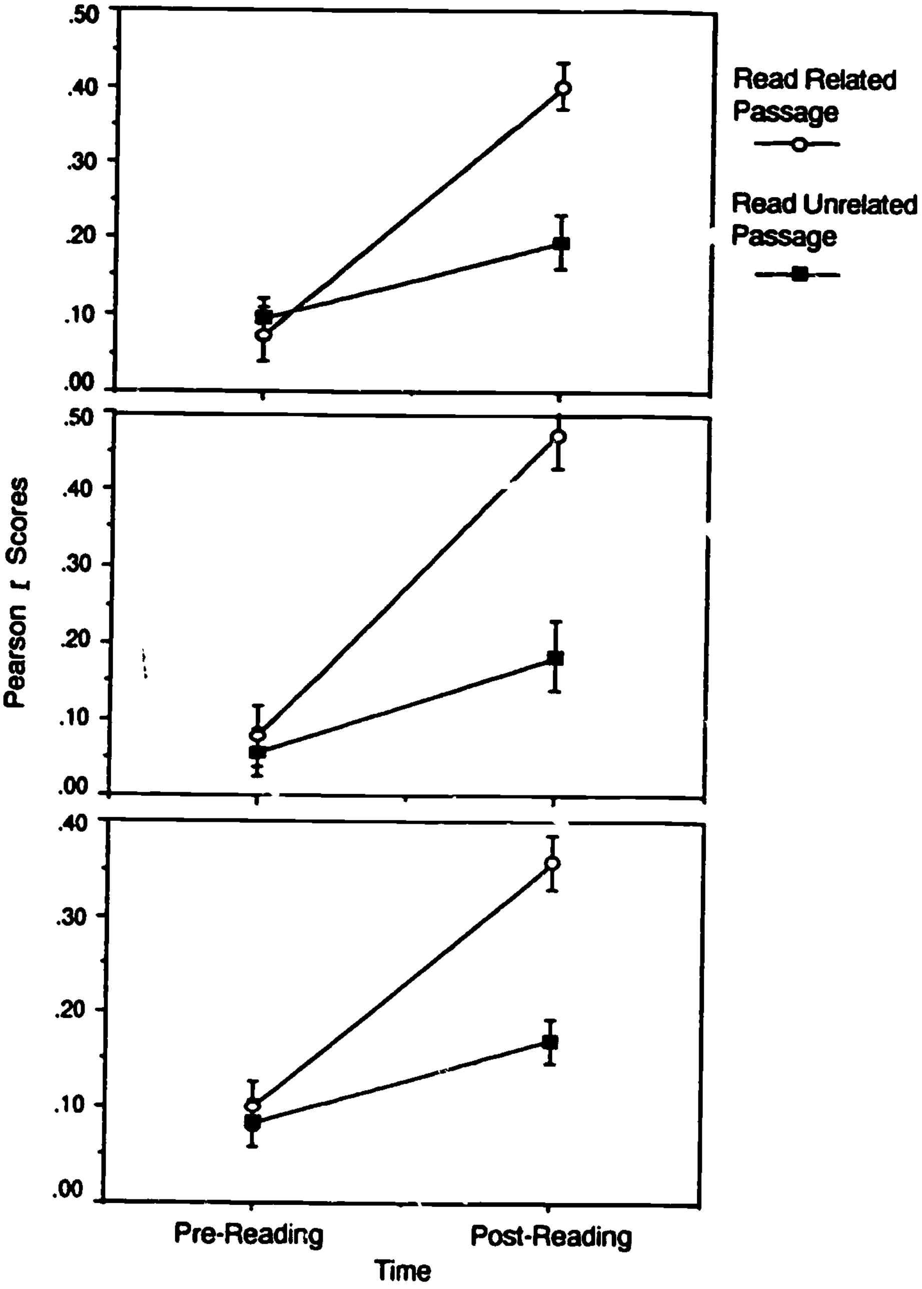
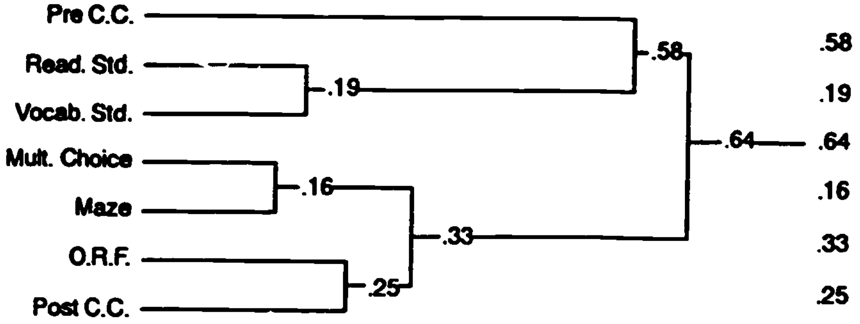


Figure Caption

Figure 6. Hierarchical Cluster Analysis of Correlation Matrix: Pre-Reading and Post-Reading Concept Comparison Scores, the Maze, Multiple Choice, Oral Reading Fluency, and Published Reading and Vocabulary Tests (N = 39).



Appendix A

Passages Which Served as Basis for Concept Comparison Tasks

The Skeletal and Muscular Systems
(Holt General Science, pp. 525-527)

Organs working together make up systems. Two of these systems are the skeletal system and the muscular system.

The human skeleton is made up of bone and cartilage. One difference between the two is that cartilage does not contain the calcium or phosphorus compounds that bone contains. This makes cartilage more flexible than bone.

There are 206 bones in the human skeleton. Some of these bones are connected to each other by ligaments. Since ligaments stretch easily, they allow the bones to move freely. This forms what is called a movable joint.

Joints can allow movement in different directions. A hinge joint allows back and forth movement. A ball and socket joint allows rotational movement.

The inside surfacs of most joints is covered with cartilage. Joints also contain a special fluid that lubricates them so they do not wear each other away.

Movement at the joints and other parts of the body is caused by the muscles. The muscles of the arms and legs are examples of muscles that aid us in movement. These are called voluntary muscles. There are some muscles like the ones found in the digestive, respiratory, and circulatory systems that are involuntary.

All muscles work only by contracting. Since they only work by contracting, they can only pull. They cannot push. If one set of muscles pulls on a tendon to bend a joint, another set of muscles must pull on a different tendon to straighten the same joint.

[243 words]

Igneous Rocks
(Holt Science, pp. 82-83)

Heat deep inside the earth causes some rocks to melt. Red-hot, melted rock under the earth's surface is called magma. Sometimes, the magma pushes out through a crack or a weak spot in the earth's crust. Red-hot melted rock coming out of the earth is called lava. The lava piles up, cools, hardens, and forms a mountain of solid rock. This kind of mountain is called a volcano.

Rocks that form from melted material that cools and hardens are called igneous rocks. The word igneous means "coming from fire". Hardened lava is one kind of igneous rock. The way the rock looks depends on how fast the lava cooled.

The lava cools slowly as a volcano becomes inactive. Rocks formed by the slow cooling of melted material have large crystals. Crystals are the structures that minerals form when they are solid. Gabbro is an igneous rock that has large crystals of many minerals.

In active volcanos, the lava is mixed with hot gases. The lava explodes, or erupts, through a small hole in the earth's surface. When this happens, the hot material often cools quickly. There is no time for crystals to form. The lava hardens and looks like a glass rock. This kind of rock is called obsidian.

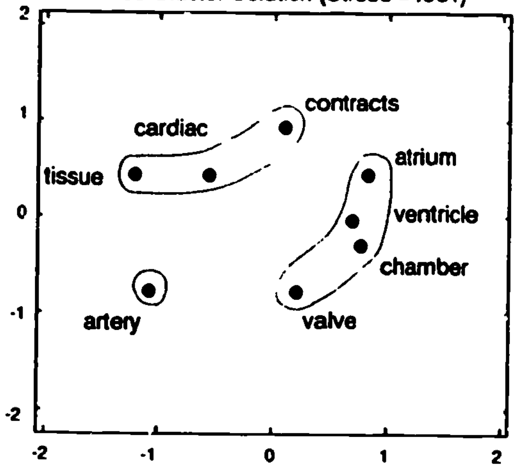
At times, lava cools so fast that the hot gases mixed with the lava do not have time to escape. They become trapped inside the hardened lava and form a spongy rock light in color. This kind of igneous rock is called pumice.

[252 words]

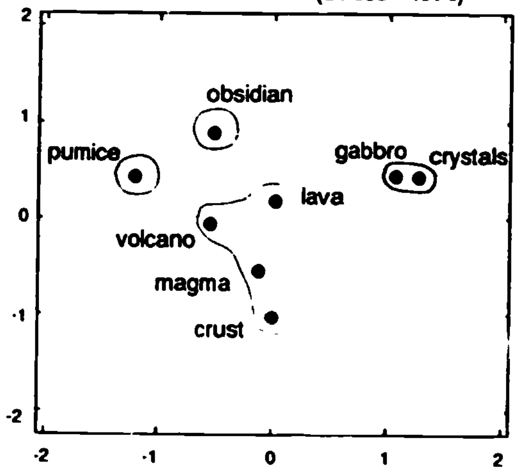
Appendix B

Average Expert Teacher Maps

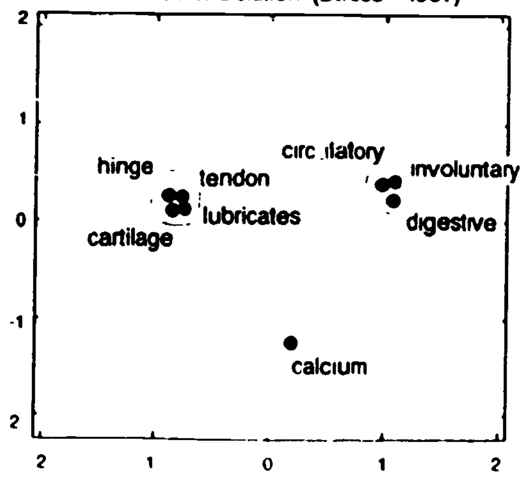
"The Heart"
Three-Cluster Solution (Stress = .031)



"Igneous Rocks"
Four-Cluster Solution (Stress = .078)



"The Skeletal and Muscular System"
Three-Cluster Solution (Stress = .007)

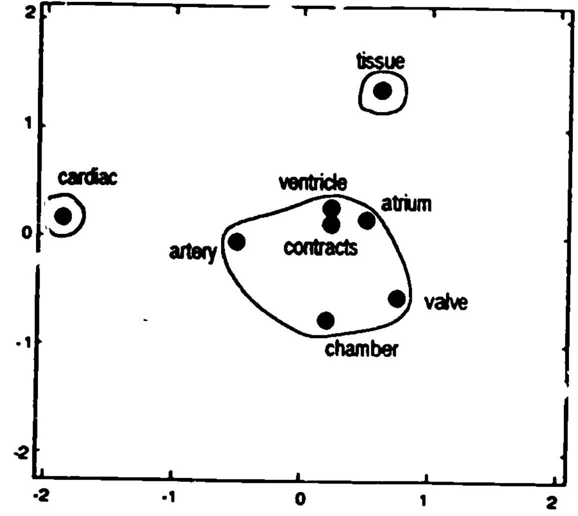


Appendix C

Pre-Reading and Post-Reading Maps for Alice and Bob

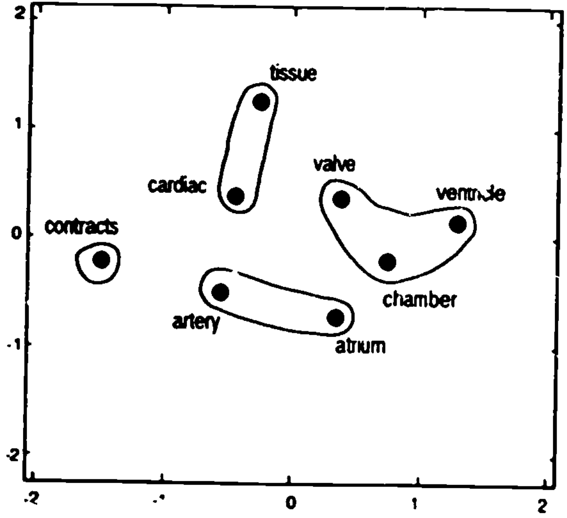
"The Heart" - Alice - Pretest

r = .12, τ - b = .08, Ω = .27 (Stress = .0095)



"The Heart" - Alice - Postest

r = .46, τ - b = .36, Ω = .61 (Stress = .046)

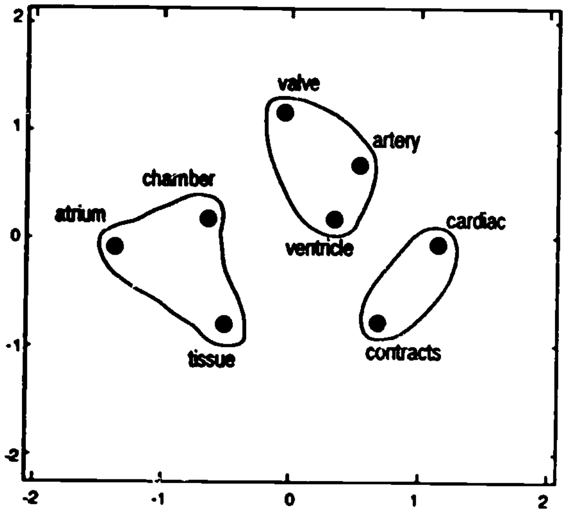


Appendix C

Pre-Reading and Post-Reading Maps for Alice and Bob

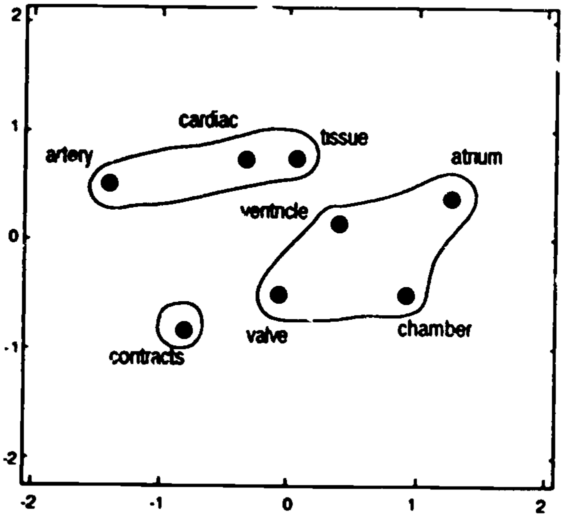
"The Heart" - Bob - Pretest

r = -.09, τ - b = .12, Ω = .33 (Stress = .064)



"The Heart" - Bob - Postest

r = .39, τ - b = 40, Ω = 74 (Stress = .079)


cardiac
tissue
artery
atrium
ventricle
valve
chamber
contracts