Resource Consultant Training Program
Research Report No. 3

RCTP

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

Richard Parker
Gerald Tindal

University of Oregon, Division of Teacher Education, Special Education Area,
Eugene, Oregon, 97403-1215

Published by
Resource Consultant Training Program
Division of Teacher Education
College of Education
University of Oregon

Parker, Richard, & Tindal, Gerald
The Reliability, Sensitivity, and Criterion-Related Validity of Concept Comparisons and Concept Maps for Assessing Reading Comprehension
Research Report No. 3

Staff

Gerald Tindal, Program Director
Jerry Marr, Editor
Denise Styer
Donna Jost
Clarice Skeen
Mike Rebar

Acknowledgments

Preparation of this document was supported in part by the U.S. Department of Education, grant numbers G008715106-89 and G008715710-89. Opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by the Department should be inferred.

Cover Design: George Beltran

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

Richard Parker
Gerald Tindal

Abstract

Most current reading assessment methods do not reflect the reading comprehension construct that 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 they are insensitive to the structural nature of text information and student knowledge. This study investigated the reliability and concurrent, 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 test 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 the continued use of concept comparison tasks and derived MDS maps in research for assessing reading comprehension with junior and senior high school disabled readers.

Background of the Problem and Research Questions

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 that 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 are often not closely related to the information demands placed on them by the text.

Construct Validity and Reading Comprehension Assessment

Because of the inadequate relationship between the "knowledge structure of the examinee and that of the test," most standardized reading comprehension tests have been characterized as 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). This shift in ascribed importance is apparent in the latest Standards for Educational and Psychological Testing (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).

The traditional view of content validity has played a large role in determining the form and content of reading comprehension tests. Two persuasions are apparent among test creators concerned with content validity (Fitzpatrick, 1983). The first has considered a test as a behavioral sample, and judges whether test performance is representative of a desired universe of behaviors (Aiken, 1979; Anastasi, 1988). The main concern is a clear description of a behavioral domain, the relevance of test responses to the domain, and sampling adequacy from the domain. Text-related variables are not considered relevant to this view of content validity. The second persuasion has emphasized test content rather than student behavior (Brown, 1976; Thorndike & Hagan, 1977). Similar judgements of domain specification and sampling adequacy are made, but in relation to a text-based rather than behavioral universe. The nature of the student response in the testing situation is therefore accorded little importance.

Researchers representing the cognitive processing view of reading comprehension have recently contributed a third point of view on what constitutes a valid test. They have suggested that the validity of a test be determined in part by what cognitive processes and structures are represented or implied in both test content and required student behaviors (Kirsch & Guthrie, 1980; Guion, 1978). The traditional content and behavior facets are thus redefined at a higher level of abstraction. Both facets are challenged to reflect the construct of reading comprehension as a largely cognitive phenomenon. 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).

A second shift in psychometric thinking about validity is related to the uses of test scores. The terms "decision validity" or "discriminant validity" refer to the interpretation of a test score and its use for decision making (grouping, instruction, curriculum modification, skill diagnosis, etc.) (Hambleton, 1980; Messick, 1981, 1989). Test use in schools has social consequences for teachers and students. A valid test helps make decisions that are valued and consistent with the schools' broader mission and goals (Messick, 1989). Johnston (1984) notes that the concept of test validity has been moved back into its instructional context, and that validation studies increasingly will include instructional interventions. The remainder of this introduction will explore how the construct validity of reading comprehension tests may be improved.

Cognitive psychology research on the reading process (Trabasso, 1978; Freedle, 1979) has implications for both the content and required student performance of construct-valid reading comprehension tests (Curtis & Glaser, 1983; Schwartz, 1984). At the most general level, Sternberg (1981) has distinguished two kinds of assessment: tests of cognitive *processing*, and tests of cognitive *content*. Tests of cognitive processing typically involve clinical recording of strategies and style exhibited while completing standard reading comprehension tasks. Tests of cognitive content involve reco-content orientation is more closely related to existing notions of achievement and content mastery as desired educational outcomes. Tests of cognitive content can potentially measure a student's grasp of in-

creasingly complex concepts and relationships in the social and physical sciences. For these reasons, the interests of this study are restricted to cognitive content assessment.

Within the cognitive content orientation, there is general agreement that a reading comprehension test must reflect both organization of prior knowledge (pre-reading schema), and selection and organization of key concepts from text (Johnston, 1984). A major goal of text researchers is 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). Reading comprehension can then be described as differences in structure and content between the text and the reader's cognition. The same goal is articulated by Kirsch and Guthrie (1980), who seek a method for matching "the knowledge structure of the examinee and that of the test" (p. 81). Three sequential assessments of conceptual content and structure are desired: pre-reading schema, text structure, and post-reading schema. However, we need to better understand the characteristics of these hypothesized schema and text structures to build appropriate assessment tools. Models for both cognitive knowledge and text-based information have come from cognitive psychology and psycholinguistics, and from the applied field of text structure analysis.

Cognitive Structures

Cognitive structure theorists are divided into those hypothesizing two distinct coding systems for imagery and for verbal processes, and those proposing a common system of propositional representations for both types of input (Richardson, 1983). Johnson-Laird (1983), speaking for the dual coding position, cites evidence that we store "cognitive maps" as analogues to the perceptual world, which capture key temporal/spatial/logical relationships, are recursive, and allow inferencing. The dual coding position is supported by a considerable body of research on human perception confirming that "...humans use frameworks similar to geometric spaces for organizing or perceiving many types of objects or concepts" (Fenker, 1975, p. 39). It is conjectured that these frameworks or "Mental Models" are richly interconnected with the independent linear verbal/linguistic system (Johnson-Laird & Wason, 1977). The introduction of this flexible and adaptable Mental Models construct appeared to spell the demise of narrower linguistic theories of information processing and storage (Levy, 1987).

In contrast, the *common coding* theorists contend that regardless of input mode, all information is stored as abstract, non-spatial, non-analogical semantic networks of propositions (Anderson & Bower, 1973; Rumelhart, Lindsay & Norman, 1972). A great advantage of these propositional networks is that they can be sufficiently described by explicit rule systems, and can be readily computer-modeled (Morris, 1987). Johnson-Laird acknowledges linguistic and proposition-based systems, but suggests that his Mental Models may include linguistic, propositional, and spatial elements (Millward, 1985). This flexibility allows the Mental Models to be applied to both text-based and non-text learning and problem-solving. In fact, Mental Models were introduced to account for the wide range of observed context-specific problem-solving behavior which was not easily explained by activation of generic schema (Brewer, 1987).

Van Dijk and Kintsch (1983) have responded to the same perceived deficiencies of overly narrow propositional networks by introducing "situation models," which are described as a mixture of instantiated schemas and Mental Models. The assessment methods investigated in the present study are consistent with Johnson-Laird's Mental Models in that they prominently include a non-propositional spatial component.

Text-based Structures

The field of text structure analysis intersects the field of cognitive structure analysis. For instance, Anderson's progress with text comprehension at the University of Illinois had a major impact on conceptions of cognitive structure outside the reading field. A cross-discipline perusal of periodicals in cognitive psychology, psycholinguistics, and reading research shows considerable cross-referencing in the past seven or eight years. Text analysis, however, is typically described and reviewed as a separate sub-field by reading researchers (Meyer & Rice, 1984). Though conducted as a branch of reading research, text analysis does not have assessment as an aim, and has not to date yielded assessment tools (Johnston, 1984). In fact, the recently described cognitive learning models developed outside of reading research may hold more promise for the complex task of assessing pre-reading schema, text structure, and post-reading knowledge.

Three prose analysis systems are most widely cited in educational research: Meyer (1975),. Frederiksen (1975, 1977), and Kintsch and van Dijk (1978). Meyer's text structure model yields a hierarchy of propositions ordered according to their rhetorical and logical relationships, and has been used to study the relationship between text structure and recall from reading (e.g., Meyer, Brandt, & Bluth, 1980). Frederiksen (1975, 1977) used a dual approach to analyze text structures: a semantic network of linguistically related propositions, and a logical network of broader idea units. Kintsch and van Dijk's (1978) "content frame" is a propositional hierarchy similar to Meyer's. Organized by logical relationships, content, repetition, and intuitive selec-

tion, it has been used to describe the semantic structure of text, and readers' recall and summarization of that structure (e.g. Winograd, 1984).

Meyer's model is closely bound to text; the inclusion of mechanical and syntactic relationships reduces the usefulness of her model for the assessment of pre-reading schema. Kintsch's semantic model is also text-bound, but to text-based propositions rather than mechanical and syntactic relationships. Van Dijk and Kintch's (1983) later "situation model" does extend beyond narrow text-based propositions to consider a broader range of concepts and logical relationships which may underlie propositions. Frederiksen's logic-based network resembles a detailed concept diagram of a topic or content area excluding the syntactic devices in Meyer's model. It is not "text-bound," and appears sufficiently flexible to measure pre-reading schema, text structure, and post-reading schema, though no such comprehensive use has been attempted. Applications of Frederiksen's networks are limited by their considerable detail and complexity (Bridge & Tierney, 1981).

Two other text analysis systems are less well known, but have future potential for contributing toward reading comprehension assessment. de Beaugrande's (1980) networks of both syntactic and conceptual relationships could be applied to all three measurement tasks (pre-reading, text structure, post-reading). However, de Beaugrande's system is also burdened by complexity, requiring extensive networks for short passages of only a few lines.

Graesser's (1981) Conceptual Graph Structure is composed of explicit and implicit content-based knowledge statements, and appears not to be overly dependent on a particular passage. However, the Graph Structure is created from answers to multiple "why," "how," and "when" questions for each text statement. That type of text dependency would make application to pre-reading schema assessment problematic. In addition, the empirical process of deriving the Graph Structure appears laborious.

In summary, text structure analysts have produced sophisticated models which cannot be directly ported to assessment, but which are rich in suggestions for types of measurable structures and relationships. Text structure models currently appear to be too detailed and complex, too tied to micro-structures, and too tied to syntax and text conventions to suit our flexible assessment needs. Our quest for greater construct validity mandates procedures which are equally suited to the organization of information in text and knowledge in the mind of the reader.

Classroom Semantic Mapping

Concurrent with the more theoretical text structure analysis, teachers have applied vocabulary- and text-mapping techniques in classroom instruction (Heimlich & Pittelman, 1986). For at least the past few decades they have used a variety of mapping techniques to teach content vocabulary and explain key concepts in text (Niles, 1965; Hauf, 1971). In common classroom use are two-dimensional "webs" of key concepts, characters, topics, or events (sometimes including labeled connecting lines) and hierarchical branching trees, with a broad topic as the trunk, and details or subordinate concepts at the ends of branches.

Text mapping presented to older students is often in the form of a topical outline or flowchart. Students are taught to recognize and discriminate among a limited number of formal outline types, e.g. narration, explanation, comparison/contrast, problem/solution (Schank & Abelson, 1977; Wilensky, 1978; Raphael, Englert, & Kirschner, 1986; Sinatra, Stahl-Gemake, & Morgan, 1986). These classroom techniques are often borrowed directly from text structure research; both Raphael et al. (1986) and Sinatra et al. (1986) relied on text structure research by Armbruster and Anderson (1984).

Geva (1980, 1983) has developed and tested "flowcharting" at the college level as an instructional tool for training less skilled readers to "pay close attention to the structure of expository texts and to utilize the logical-structural information ... provided by conjunctions" (p. 386). A two-dimensional hierarchical diagram contains both individual words and propositions linked by seven types of drawn lines which specify "...topic, elaboration, cause-effect, process, example, detail, and conclusion..." relationships (p. 387).

For younger students, text structure often is presented graphically as a web of interrelationships among elements of a text passage or story. Called "concept maps," "semantic maps," or "story maps," these displays are composed of nodes and relationships. The nodes may be key content vocabulary, main and supporting ideas, events, characters, goals and motives, etc. Relationships may include cause-effect, time order, subordinate-superordinate, action-reaction, major-minor importance, etc. Graphic displays of these relationships have included "hub and spokes" diagrams (for main and peripheral ideas), hierarchically branching "trees," and complex "webs," or matrices of relationships among vocabulary words or facts (Calfee & Drum, 1986; Reutzel, 1986).

Perhaps the most extensive and systematic application of semantic mapping has been developed by Holley and Dansereau (1984) within a broader learning strategies framework. The flexible mapping tools include hierarchies, sequential chains, and clusters of concepts. Semantic mapping has been found to help students concentrate while reading, and organize and memorize text information. Dansereau's methods are supported by a series of empirical studies extending over several years (Dansereau, et al., 1979).

Whether detailed and systematic or more intuitive, semantic maps have demonstrated usefulness as teaching tools (Reutzel, 1986). An ironic limitation in studies of these mapping techniques is the lack of a suitable dependent measure, one which is sufficiently sensitive to treatment effects. Earlier it was established that most existing reading tests are not designed to reflect structural changes in knowledge, whereas semantic mapping interventions often aim to do just that. Assessment unfortunately has lagged behind instruction in this area.

Researchers lack proven, replicable methods for (a) producing maps and hierarchical diagrams from text for instructional use, and (b) measuring student learning from this instruction with methods which are structurally sensitive. At least two measurement methodologies (one primary and one supplemental) can potentially address this need—both with very limited application to reading comprehension to date.

Structurally Sensitive Assessment

To measure and reliably compare cognitive and text structures implies a technology of testing beyond that commonly used in education (Johnston, 1984). One important limitation of present assessment techniques is their lack of structural sensitivity (Surber, 1984). The tractability of this problem depends on whether the hypothesized structure is a narrow, abstract linguistic or propositional model, or a broader semantic, logical, or temporal/spatial (or mixed) model. The former case is the less formidable; individual elements and connectors of the linguistic or propositional model can be separately scored (as individual items), and scores aggregated for a macro level index. Error patterns of item clusters can then be interpreted at a particular level of the network or map (Surber & Smith, 1981).

In the latter case, structural insensitivity poses a bigger problem. The spatial dimension must be preserved, as it may represent a number of relationships (time, physical space, similarity, class membership, joint functions, etc.). Not only must a range of relational connections be summarized, but also the spatial array, including grouping or clustering and distances among individual concepts and concept clusters (Holley & Dansereau, 1984).

The intent of the present study was to explore the use of a structurally sensitive methodology for the problem of assessing reading comprehension through pre-reading schema, text structure, and post-reading schema. Our goal was to use a measurement method which is sensitive to the spatial dimension of semantic networks. We hypothesize, after Johnson-Laird (1980), that spatial arrays can be analogues for one or more dimensions of perceptual reality. Given the current cognitive processing view of reading, we aimed to produce a more construct-valid alternative to the current dichotomy of norm-referenced reading "ability" tests (Allington, 1982; Kavale, 1981) versus criterion-referenced tests based on behavioral domain sampling (Bloom, 1976). The third alternative is a class of tests which can reflect the structure of information and knowledge and changes in that structure due to reading.

Two available methodologies for specifying the structure of a set of concepts are multidimensional scaling (MDS) and hierarchical cluster analysis (HCA) (Preece, 1976; Shavelson, 1974). Both MDS and HCA result in graphic displays of key concepts or vocabulary, 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 relationship of pairs of important concepts or key vocabulary words; correlation coefficients may be used instead of similarity judgments. The input for cluster analysis may be more varied, but includes similarity ratings and measured distances between objects.

Three main applications have been identified for MDS: (a) improved comprehension and communication of complex relationships among concepts, (b) verifying hypothesized concept patterns (through comparison with outside criteria or externally produced maps), (c) interpreting map dimensions (Davison, 1983). The first two applications are relevant to this study. Two-dimensional displays of concepts can help provide and communicate meaning through (a) the identification and labeling of concept clusters, (b) the identification and labeling of relationships among concepts and concept clusters, and/or (c) the identification of subordinate relationships among concepts and concept clusters. Interpretation of clusters and relationships is demonstrated on maps derived from a concept comparison task after students read a short science passage. A more detailed explanation of the concept comparison task will be provided later.

Maps in Figure 2 depict MDS plots of eight key concepts from a science textbook passage, "The Heart" (Barufaldi, Ladd, & Moses, 1981) (Figure 1). In the top map, objective decision rules were used to identify meaningful clusters of the concepts; they were then outlined and labeled. In the bottom map, interpretation of the same map highlights *relationships* rather than *clusters*. Relationship labels were selected from taxonomies of semantic relationships developed by Holley & Dansereau (1984), Frederiksen (1975), de Beaugrande (1980), and others. The bare maps are produced from MDS analysis of initial concept comparison ratings through mainly objective procedures. Map interpretation, however, entails subjective judgments, along with knowledge of the content area and the particular text passage.

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 four compartments or *chambers* in the heart. Each upper chamber is called an *atrium*. An *atrium* is a small, thin-walled chamber that receives blood from the lungs or the body. Each lower *chamber* is called a *ventricle*. A *ventricle* is a thick, muscular *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 1. Science Text Passage with Key Vocabulary Terms Underlined for MDS Mapping

Whereas identification and labeling of concept clusters and relationships is performed directly on the MDS map, identification of subordinate relationships among concepts and clusters requires a supplemental procedure: hierarchical cluster analysis (HCA). Analysis of map inter-concept distances produces hierarchical cluster trees or dendograms. Cluster trees depict the concepts not as points uncategorized on a dimensional space, but as members of discrete categories at multiple nested levels (Everitt, 1988). Interpretation of the cluster tree involves (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; Kruskal & Wish, 1978; Griffeth, Hom, DeNisi, & Kirchner, 1985). Figure 3 displays a cluster solution of the MDS maps from Figure 2. 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 demonstration purposes. In Figure 3 the main branch is labeled with the passage title, "The Heart." Note that the branch labels for the three-cluster solution are the same as the cluster labels on the top MDS map in Figure 2. The cluster tree thus provides a third, complementary interpretation of the MDS map clusters.

Figures 2 and 3 demonstrate the first main use of MDS: for improved comprehension and communication of complex semantic relationships. For the second main MDS use, verifying concept patterns, externally produced text-based "expert" maps can be used as evaluative *standards* for an individual student's concept map. The standard map and learner maps can be quantitatively compared through 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 trivial or critical.

The examples in Figures 2 and 3 demonstrate the strengths and limitations 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 those which are perceptual and those which are more abstract. In this way, they most closely approximate Johson-Laird's Mental Models construct. Although the mapped elements in Figure 2 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.

As measurement techniques, neither MDS nor HCA 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 number of psychometric studies, summarized in recent research reviews (Carrol & Arabie, 1980; Young, 1984), textbooks, (Davison, 1983; Schiffman, Reynolds, & Young, 1981) and dedicated journal issues (*Applied Psychological Measurement, 7*[4], 1983: *Psychometrika, 51*[1], 1986). While MDS lacks the statistical power associated with normal distribution assumptions and interval/ratio measurement scales, it does offer 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).

Hierarchical cluster analysis (HCA), which is relegated in this study to supplementary analyses, is

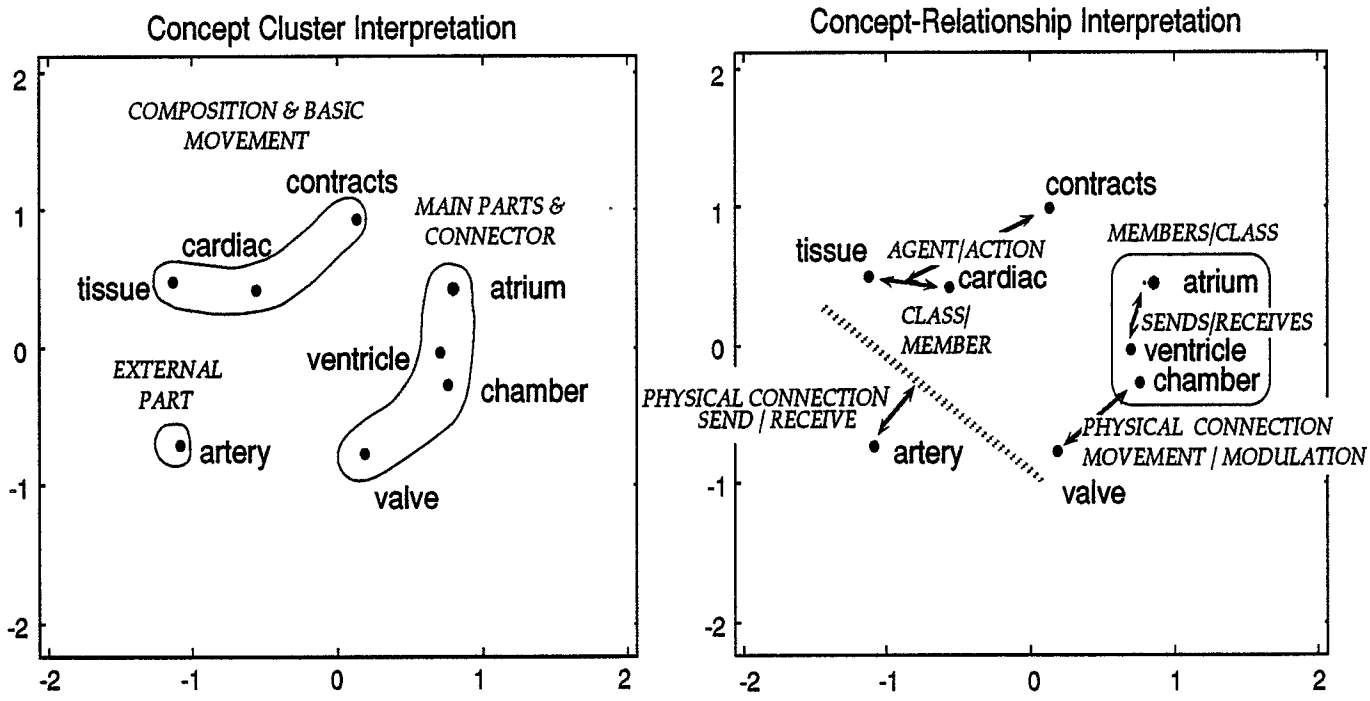


Figure 2. Two Interpretations of an MDS Map Concept Configuration

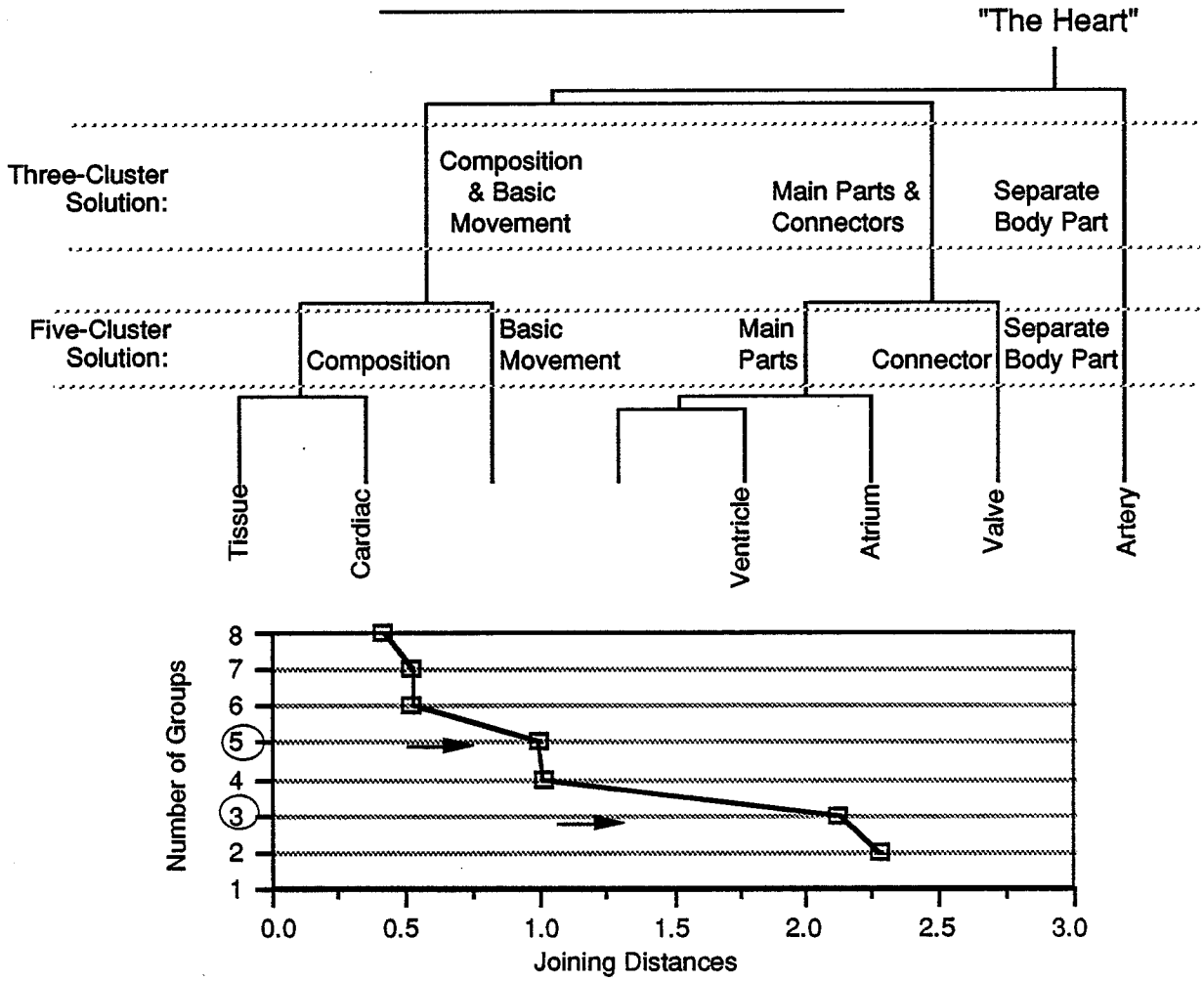


Figure 3. Hierarchical Cluster Interpretation of an MDS Map

considered an "exploratory" technique—seldom recommended for primary analyses (Everitt, 1988, p. 604). Clustering inter-concept map distances yields a hierarchical structure for individual concepts and higher order categories. Together, MDS and cluster analysis offer spatial maps and hierarchical trees similar to Johnson-Laird's (1980) "mental models" and van Dijk and Kintsch's (1983) "situation models," as well as the maps and networks traditionally constructed by teachers intuitively and by hand. The suitability of these methods for reading comprehension assessment can be judged in part from the literature on applications to reading and other student learning. Unfortunately, relatively few such studies exist.

MDS Maps and Learning Assessment

Multidimensional Scaling has been used to study changes in students' knowledge 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 20 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 derived MDS maps (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. CC scores correlated .66 with a composite of the three essay question types. Based on these results he authors concluded the following:

> 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)

MDS Maps and Reading Comprehension Assessment

More directly related to the present study are the few applications of MDS to expository and fictional reading passages (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 systematically 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.

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 interpretation of structures formed by spatial proximity. Map summary through dimensional interpretation overlooks the complex configurations of concept clusters and inter-concept relationships, thus reducing the method's sensitivity and diagnostic utility (Shepard, Kilpatrick & Cunningham, 1975).

The few studies applying MDS to student learning and reading comprehension in particular are encour-

aging. However, the reading studies have employed a very limited number and variety of passages, mainly from adult-level reading material and with able readers Maps most often have been summarized by two or three dimensions, rather than much richer configuration summaries, reducing their potential for diagnosis and instruction. Some evidence supporting the validity of MDS procedures is offered by the few studies just cited. However, evidence supporting the reliability of these procedures in reading assessment is not available. Fundamental reliability and validity questions must be addressed before MDS can be considered seriously for reading comprehension assessment.

Purpose

The present study is an initial investigation of the use of structurally sensitive MDS procedures for assessing comprehension of intermediate-grade science and social studies reading passages by students with reading disabilities. The study was conducted in two phases, addressing (a) instrument reliability, and (b) instrument sensitivity and concurrent, criterion-related validity. The central question of the first phase was: After reading 250-word intermediate-grade science and social studies passages, will teachers independently produce similar concept comparison (CC) ratings and MDS maps? The usefulness of MDS in assessing reading comprehension depends in part upon the reliable identification of "expert" maps with which pre- and post-reading student maps can be compared. In the validation phase, pre- and post-reading CC scores and MDS maps obtained from middle and high school students in Special education and Chapter 1 compensatory programs were compared with the expert teacher maps and with four external criterion reading tests.

Phase I: Instrument Reliability

Method

Reading Passages

Eight 250-word passages were selected from elementary level social studies and science texts (Holt Science, Ramsey, Gabriel, McGuirk, Phillips, & Watenpaugh, 1985; Barufaldi, Ladd, & Moses, 1981; Hagus, Reque, & Wilson, 1985). 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). The passages are included in Appendix A.

Selection criteria required that passages be cohesive and self-contained within a 250-word limit. Passages of this length typically permitted development of at least one central idea and included 8 to 12 key content-related vocabulary terms. Passages were minimally edited to delete "asides," references to charts, figures, and text located elsewhere, and longer sentences that were only peripheral to the selected passage. Editing was required for approximately five sentences per passage.

From each selected passage, eight key vocabulary terms were drawn and paired for similarity judgments within a concept comparison (CC) test. "Key vocabulary" consisted of words with central importance to the passage—both content words with meanings defined or implied, and non-content words with content-specific meanings within the passage. The selected key vocabulary included all of those words highlighted by text publishers through bold/italic type, underlining, or margin notes. In this paper, the terms "key vocabulary" and "concepts" are used interchangeably.

Concept Comparison Tasks

For each selected passage, all pairwise combinations of the eight key vocabulary terms were listed in a "Ross ordering" sequence to avoid contaminating order effects (Cohen & Davison, 1973; Davison, 1983). Although a minimum of nine concepts is recommended for a two-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.

Respondents used a 4-point scale to judge, for each pair of concepts, 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 Appendix A).

Respondents

The 15 "expert" respondents, all employed by a rural Pacific Northwest school district, included 2 district coordinators and 13 reading specialists and special education teachers from 6 junior (Grades 6-8) and senior (Grades 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 taught 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 were presented with 250-word passages and related CC tasks which they independently completed after reading each passage. While making concept-comparison judgments, they were encouraged to look back at the passage and 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 both 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). These agreement indices were used to identify those passages with highest CC interrater reliability. For three of the most reliable passages only, HCA was performed on MDS maps to assess agreement on the cluster membership (Rand, 1971; Morey & Agresti, 1984).

Results

Concept Comparison Reliability

Concept comparison ratings (scaled 1-4, from "close relation" to "little or no relation") from five raters were analyzed for each of the eight passages, using the intraclass correlation coefficient and Cohen's Kappa. The intraclass correlation is based on a partitioning of the rating variance accounted for by items, raters, and error (Brennan, 1983). Kappa is an index of categorical agreement for nominal data which conservatively reflects only agreement beyond chance (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 the degree of agreement/disagreement (from the ordinal rating scale) (Cohen, 1968). Table 1 presents the continuous (intraclass correlation) and categorical (Kappa) measures of agreement for five "expert" raters on the eight passages.

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,

Table 1. Agreement Among Five Raters on 28-Item Concept Comparison Tests From Eight Science and Social Studies Text Passages

	Intraclass Correlation	Cohen's Kappa	
		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.

"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 Agreement

While the intraclass correlation and Kappa assessed interrater CC score agreement, agreement among the MDS-produced maps required different indices. For only the three most reliable passages, MDS maps were produced for each of the five raters, using the stand-alone ALSCAL-4 (3rd ed.) software (Young & Lewyckyj, 1973), with the classical non-metric (CMDS) algorithm. The five raters' maps for "The Heart," "Igneous Rocks," and "The Skeletal System" are presented in Appendix B. The goodness of fit of each map to the rating data was first assessed through Kruskal's Formula 1 Stress. The raw ratings are monotonically matched to inter-concept map distances; *Stress* is the sum of squared residuals from this match (Davison, 1983; de Leeuw & Stoop, 1984). All but one of the 15 stress values were below .02, representing a very good fit for two dimensions and at least nine concepts (Kruskal & Wish, 1978). However, since only eight concepts were mapped, the Stress values may be somewhat reduced due to overfitted data.

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 pairs (28 in all) provided by the ALSCAL program are analogous to the original 28 CC ratings. The same intraclass correlation statistic which was applied to the CC ratings was also applied to the map distances, yielding the following reliability estimates for map configurations: "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, the more systematic procedures of hierarchical cluster analysis (HCA) with scree plots were used (Davison, Richards, & Rounds, 1986; Coxon, 1982). The Group Average clustering algorithm (Sneath & Sokal, 1973), was used, as it performed well in Monte Carlo studies (Milligan, 1980, 1981), and produced interpretable solutions for these data.

On a cluster tree, each branching level is a different potential clustering solution. The optimal clustering levels 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. Alternative clustering solutions for a particular map are mutually consistent, as they are derived from the same cluster tree. One or two optimal cluster solutions are outlined on each map in Appendix B, C, and D.

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" (W), was used, which is scaled from 0 (chance agreement) to 1 (perfect agreement) (Morey and Agresti, 1984). The W 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, average "expert maps" were 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) was also used to create an aggregate map, producing 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 E.

Phase 2: 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 assay 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 the logging industry as its economic base. At the junior and senior high levels the lowest-achieving 15% of each grade cohort in reading and language arts (approximately 330 in all) was enrolled in Chapter 1 (compensatory) or special education (learning disability category) programs. 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 the 104 students, drawn from thirteen classrooms within four junior (Grades 6-8) and two senior (Grades 9-12) high schools.

Fifty-three of the 104 students were enrolled in junior high, and 51 in senior high. Forty-three attended special education resource rooms for reading/language arts, and 61 received pull-out Chapter 1 assistance for these skills. 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 (1977) (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).

ANOVA performed on the extant vocabulary and reading comprehension scores showed no significant differences among grades at either the junior or senior 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 senior 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.

Table 2. Standardized Test Percentiles (Reading Comprehension and Vocabulary) for 104 Junior and Senior High School Students Served by Chapter I Special Education Programs

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

*IQR = Interquartile Range: Spread of the middle half of scores clustered about the Median

Concept comparison (CC) tasks. Three of the CC tasks completed by teachers were also completed by students. Each CC task consisted of 28 ratings of concept pairs drawn from a passage. Ratings were performed on a 4-point scale to indicate the relatedness of each pair of concepts (see Appendix A).

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 made syntactic and semantic sense within the sentence. For each blank in the text, students selected one of five options (see Appendix I).

Multiple choice questions. A set of ten, four-option multiple choice questions was created for each passage. One question dealt with "main idea," 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 (see Appendix H).

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

Procedure

The CC assessments were 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 junior and senior 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 test 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.

This control group design effectively controls for several of the eight potential challenges to internal validity (Campbell & Stanley, 1963, pp. 5-6). The short test-retest interval reduces the likelihood that events in *history, maturation,* or differential *experimental mortality* could confound interpretation of results. The ability to use the same CC test for pre- and post-testing also eliminates validity concerns due to changes in *instrumentation.* Although only the lowest achieving students were selected for the study, the selection instruments (standardized test scores) were not used for pre- or post-testing, so statistical regression to the mean is not an issue. Random assignment eliminates group selection as a possible confound. The potential for unwanted test reactivity is not eliminated, but the design ensures that any such effects will be equally distributed across groups.

Stage 1. On Day 1 of the first stage, during reading/language arts classes, teachers described and demonstrated the CC task from scripted instructions. Students then were instructed to complete the CC test for the passage assigned to their grade level. Fifteen minutes were allowed for the test; 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 to students in alternating order, according to classroom seating. The unrelated passages were from the same

Table 3. Design Elements: Observations and Experimental Conditions by Group Across Time

	O_1 Extant Ach. Scores	O_2 Pre-test CC	O_3 Maze	Reading: Related: (X_R) Unrelated: (X_U)	O_2 Post-test CC*	O_4 Multiple Choice	O_5 ORF**
Stage 1							
I. Treatment (n=54) Jr.: "Heart" (26) Sr.: "Skeletal" (28)	O_1	O_2		X_R	O_2		
II. Control (n=53) Jr.: "Heart" (28) Sr.: "Skeletal" (25)	O_1	O_2		X_U	O_2		
Stage 2							
I. Treatment (n=43) Jr.& senior "Rocks"	O_1	O_2	O_{3R}	X_R	O_2	O_{4R}	O_5
II. Control (n=49) Jr. & Sr. "Rocks"	O_1	O_2	O_{3U}	X_U	O_2		

*CC = Concept Comparisons **ORF = Oral Reading Fluency

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 (see Table 3). Stage 2 required reassignment reassignment of students 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 set 25 minute 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", while 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

Pre- and post-reading CC data were analyzed through 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 correlation coefficient between student and expert CC scores served as the dependent measure. 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 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, differences between pre- and post-reading CC expert correlations were tested with the Hotelling-Williams Test of correlation equality (Darlington & Carlson, 1987).

The second major analysis was the intercorrelation of scores from (a) pre- and post-reading concept comparisons (Fisher Z scores), (b) published 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, unlike pre-reading scores.

Results

Qualitative Interpretation of Students Maps

Qualitative interpretation of typical pre- and post-reading student maps is presented before the quantitative data from the two-stage control-group design. The purpose of this order is to emphasize that the sometimes-elaborate quantitative analyses play a supportive role in mapping interpretation. Qualitative analysis of the spatial map offers the greatest potential for diagnosing students' understandings and misinterpretations of text, and planning relevant remedial instruction.

Maps produced for six students (code-named A through F) with typical pre- and post-reading CC correlations (with expert maps) are presented in Appendix C. Mean Pearson r's ranged from .07 to .10 before reading, and from .36 to .47 after reading. The maps show outlines of the most defensible cluster solution based on HCA. For the six pairs of maps, Table 4 presents one index of raw score correlation (Pearson r) and two measures of map congruence—based on interpoint map distances (Kendall Tau-B), and on clusterings (Omega transform of Rand's statistic). The table is presented to highlight similarities among the three agreement indices calculated for a single set of CC scores and map. These indices also help demonstrate the relationship between agreement of CC scores and configurations of the maps in Appendix C.

The student maps can be qualitatively interpreted by comparing (a) expert teacher maps with both pre- and post-reading maps, and (b) pre- with post-reading maps. Interpretations can be based on either the map distances among individual concepts or outlined cluster membership. Both methods are supported by the agreement indices presented in Table 4; they are complementary, and can be used together.

Both the average teacher map for "The Heart" (Appendix E) and student A's pre-reading map (Appendix F) suggest a three-cluster interpretation. Figure 2 contains the same expert map as Appendix E, but elaborated with cluster and relationship interpretations. 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. Student A'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 student A's pre-reading map, "cardiac" and

Table 4. Agreement with Teacher Experts by Six Representative Students on Pre- and Post-Reading Concept Comparison Tasks and MDS Maps

	Pre-Reading CC			Post-Reading CC		
		Map Indices			Map Indices	
Passage & Student	*r*	τ-*b* *	Ω **	*r*	τ-*b* *	Ω **
"The Heart"						
A	.12	.08	.27	.46	.36	.61
B	-.09	.12	.33	.39	.40	.74
"Igneous Rocks"						
C	.08	-.02	.13	.44	.46	.69
D	.18	-.04	.33	.39	.42	.79
"Skeletal & Muscular System"						
E	.03	.20	.39	.56	.67	.62
F	-.03	-.20	.20	.50	.29	.73

*τ-*b* = Kendall Tau-B Index of Concordance **Ω = Omega: transformation of Rand Statistic of Cluster Agreement

"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 perform a more micro-level analysis of Student A's pre-reading map. Within Student A'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.

Student A'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 A's post-reading map shows greater differentiation of concepts toward interpretable, higher-order clusters.

While changes in student A'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 diagnose misunderstandings and/or plan remedial instruction.

The main similarity between Student B's pre-reading map (Appendix F) and the expert map for "The Heart" (Figure 2 and Appendix E) 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.

Student B'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 con-

firm the interpretations made on the basis of clustering and inter-concept distances.

In summary, the comprehension problems inferred from student B's post-reading map appear less severe than those of student A. Student A's most fundamental misunderstanding appears to be a confused "artery"-"atrium" connection), while student B'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 (τ-b) and cluster membership (Ω) show that student B (τ-b = .40, Ω = .74) slightly outperfomed student A (τ - b = .36, Ω =.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 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 are toward closer agreement with the expert teacher CC scores. Results from three-way ANOVA are presented for junior high ("The Heart") in Table 5, for senior high ("The Skeletal and Muscular System") in Table 6, and for both levels together ("Igneous Rocks") in Table 7. Table 5 presents main effects and interactions for the three variables in accounting for junior 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 when there are 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 as the tendency by *all* students to slightly improve in their CC scores at Post-testing (the Time variable), presumably due to a practice effect, or instrument reactivity (as will be noted in Table 8).

Tabled ANOVA Results for senior high on "The Skeletal and Muscular System" were similar to those for junior high, and consistent with our hypotheses (see Table 6). At the senior high level, only one of the three first-order interactions was significant—"Time x Read." (41% of the variance). Again both Time and Read. main effects were significant, with the much larger effect for Time (66% of the variance). The variable, Prog., did not contribute significantly.

The replication study in Stage 2, with junior and senior high students together ("Igneous Rocks"), produced results similar to the previous two analyses (see Table 7). 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%

Table 5. Three-way ANOVA for Dependent Variable, *Concept Comparison Scores*, with Independent Variables, *Program*, *Reading Passage*, and *Time of Assessment* (Junior High; "The Heart" [*N*=53])

Source of Variance & (Levels)	*SS bt*	*SS w*	*F (1,49)*	*p*	η
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 6. Three-way ANOVA for Dependent Variable, "Concept Comparison Scores ," with Independent Variables, "Program," "Reading Passage," and "Time of Assessment" (Senior High; "The Skeletal System" [*N*=50])

Source of Variance & (Levels)	*SSbt*	*SS w*	*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

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 4. For the plots, the Fisher Z scores used in ANOVA were re-converted to Pearson *r*'s. The three very similar interaction plots indicate that both junior and senior 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 8 presents CC mean scores 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. Although the ANOVAs discussed above provided pre- and post-reading CC score comparisons at the group level, they did not provide information at the individual student level.

Table 7. Three-way ANOVA for Dependent Variable, "Concept Comparison Scores," with Independent Variables, "Program," "Reading Passage," and "Time of Assessment" Senior High (43) and Junior High (47); "Igneous Rocks"

Source of Variance & (Levels)	*SS bt*	*SS w*	*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

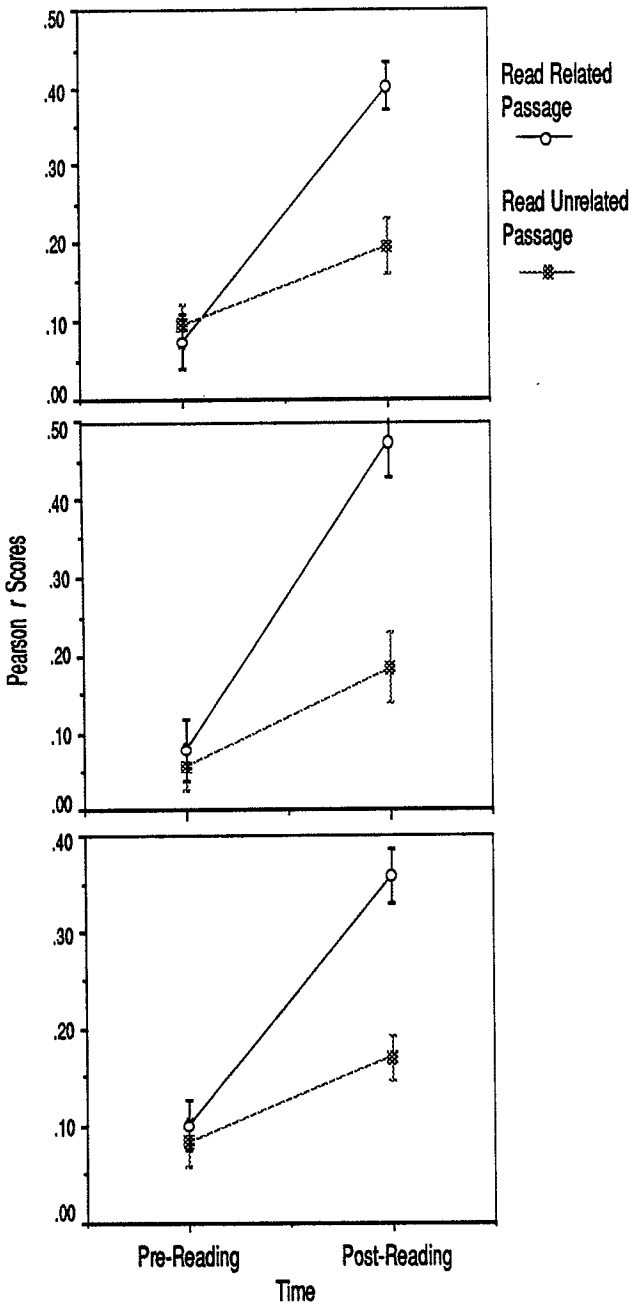


Figure 4. Interaction Plots of "Time x Read." for "The Heart" (Junior High), "The Skeletal and Muscular System" (Senior High), and "Igneous Rocks" (Junior and Senior High)

Individual-level results are essential when diagnosis or placement decisions are based on test results. Therefore, for only those students who read the "related" passages, the null hypothesis of no significant difference between pretest and posttest 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 postest-expert correlations (Darlington & Carlson, 1987).

Pretest-expert correlations were stronger than posttest-expert correlations for only 6 of the 97 treatment-group students, and none of these differences was statistically significant. In contrast, posttest-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. Therefore, a Chi-square test was performed on the proportion of significant versus non-significant findings. The resulting coefficient, $c^2(1, N=97) = 84.75$, $p < .0001$, indicated that the number of obtained significant Pearson r differences was far beyond chance level.

Criterion-Related Validity

The second major analysis was comparison of pre- and post-reading CC scores of Stage 2 treatment group students (those who read the related passage) with external measures of reading comprehension. Table 9 contains descriptive information on the CC scores, published standardized tests, maze tests, multiple choice tests, and oral reading fluency, which were intercorrelated. Intercorrelations among these seven measures are presented in Table 10 for 38 students in the Stage 2 treatment group. It was hypothesized that post-reading CC scores would be substantially related with the other measures, unlike pre-reading scores. The correlation matrix in Table 10 shows small, non-significant relationships between the pre-reading CC scores and external measures. 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 scores, 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 (see Figure 5). Ward's method, which minimizes the variance within clusters (Ward, 1963; Blashfield, 1980) is strongly supported by Monte Carlo studies (Milligan, 1980). 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 and Conclusions

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

Table 8. Pre- and Post-Reading Concept Comparison Scores (Pearson *r*'s), with Reading of Related or Unrelated Passage

"The Heart," Junior Hgh

	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," Senior 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	.20	.512	.21	.078	.198	.474	.24
Un-Related	.076	.139	.165	.251	.041	.154	.199	.222	.056	.146	.184	.23

"Igneous Rocks," Junior and Senior 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

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; i.e., 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 new assessment method: reliability, sensitivity, and validity: The first refers to the *reliability* of expert teacher concept comparison (CC) scores and MDS maps; the second refers to the *sensitivity* of CC scores to response changes following relevant reading; and the third refers to concurrent criterion-related *validity*, i.e., the relationship between CC scores and other reading measures.

The reliability of only teacher CC scores and maps was studied directly; reliability of student CC scores and maps was not. In addition, the stability of expert

Table 9. 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

teacher ratings over time needs to be studied. Only indirect evidence for stability of student scores and maps over time was available through a pre-, posttest lag of one day. Evidence from students reading the unrelated passage suggests that CC tests are reactive; pre-testing appeared to systematically influence posttest results in the direction of greater similarity to the expert map. The important question of student score stability requires further study.

The question of reliability of expert teacher CC scores and maps receives 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 for teachers were weaker. Only two of the Kappa/Kappa Max. ratios were substantial (above .70). However, the implication of this size of reliability coefficient for decision making based on a mapping test is 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 our knowledge of disagreements among teachers on a story's main idea. 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 the 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

Table 10. 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 CC	Post CC	Maze	Mult. Choice	Read. Std.	Vocab. Std.
Post CC	.42*	•				
Maze	.28	.61*	•			
M.Choice	.15	.45*	.75*	•		
Read. Std.	.21	.36	.36	.38	•	
Vocab. Std.	.19	.38	.43*	.38	.66*	•
ORF	.15	.57*	.60*	.51*	.37	.45*

* $p < .01$

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 instrument sensitivity, 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. A larger set of concept comparisons would have increased the number of individual "improvements" that reached significance.

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 is encouraging on measurement sensitivity for use with disabled readers. Clearly, only one type of sensitivity was assessed. Treatment validity, or the sensitivity for measuring score changes due to an instructional intervention, was not investigated. Such an investigation would have direct implications for probable future classroom use of assessment maps.

The third research question, assessment of concurrent, criterion-related validity, also tentatively receives an 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. As a group, the four types of validity measures reflect virtually all of the deficiencies of existing reading comprehension tests as defined by cognitive researchers and described earlier. These deficiencies, and tests representing them are: (a) lack of structural sensitivity (maze, ORF, multiple choice, standardized tests), (b) inability to account for pre-reading knowledge (maze, ORF, standardized tests), (c) information processing demands inconsistent with those hypothesized by schema theory (maze, ORF, multiple.choice, standardized tests), (d) unintentionally cuing of responses by test questions (maze, ORF, multiple choice, standardized tests). Comparing a new measure against deficient criterion measures will always result in less than satisfactory validity coefficients.

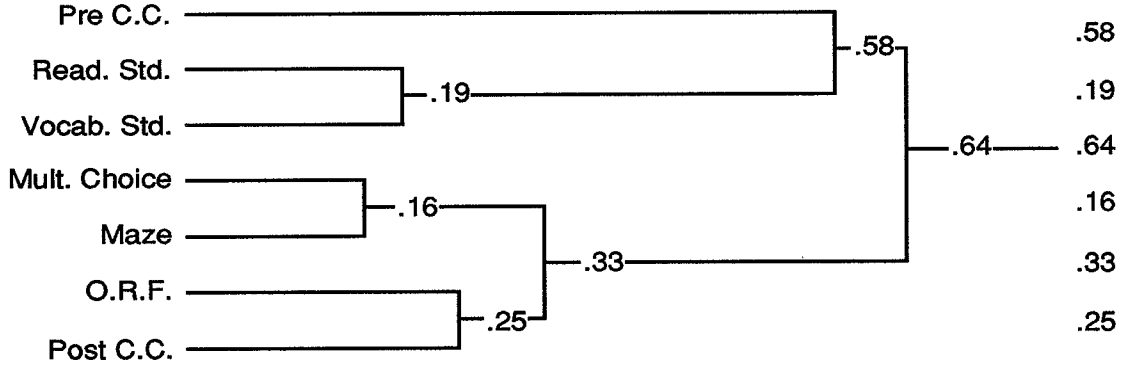


Figure 5. Hierarchical Cluster Analysis of Correlation Matrix: Pre-Reading and Post-Reading Concept Comparison Scores, the Maze, Multiple Choice, Oral Reading Fluency, and Standardized Reading and Vocabulary Tests (N = 39)

Conclusions

This study served as an initial investigation of the reliability, sensitivity, 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 a research setting. Foremost is the question of the range of types of relationships among concepts which can be reflected on a two-dimensional space. MDS map Interpretation

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 limit the variety of relationships 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. Alternatively, some constraints may need to be set on the initial selection of the key vocabulary from the passage. The available taxonomies of relationships would make selection criteria feasible.
Under conditions of different vocabulary selection guidelines, map reliability and interpretability could be studied

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. Nor 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 device (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 should 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 potentially can 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 primarily 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

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. (1988). *Psychological testing* (3rd ed.). New York: Macmillan Publishing Company.

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.

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.

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

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. L. J. (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.* New 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.). New York: Holt, Rinehart & Winston.

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

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, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measures.* New York: 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.* New York: 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, NJ.: 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.

Fitzpatrick, 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, NJ.: Ablex.

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

Geva, E. (1980). *Meta textual notions and reading comprehension.* Unpublished doctoral dissertation, University. 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.

Graesser, A. C. (1981). *Prose comprehension beyond the word.* New York: Springer-Verlag.

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,* 813-840.

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

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 application.* Newark, DE: 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-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4,* 71-115.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard 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* (pp. 66-83). 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, 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. In P. Morris (Ed.), *Modeling Cognition* (pp. 3-20). New York: John Wiley and Sons.

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. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-105). New York: American Council 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. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of Reading Research* (pp. 319-352). 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 effect 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.

Morris, P. (1987). *Modeling cognition.* New York: John Wiley & Sons.

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

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

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

Ramsey, W. L., Gabriel, L. A., McGuirk, J. F., Phillips, C. R., & Watenpaugh, F. M. (Eds.) (1985). *Holt general 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, MI: The Institute for Research on Teaching, Michigan State University.

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.), *Schooling and the acquisition of knowledge* (pp. 99-135). Hillsdale, NJ: Erlbaum.

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

Schank, 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.* New York: 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.

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, NJ: 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.

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.). New York: Wiley.

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

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

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.

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

Young, F. (1984). Scaling. *Annual Review of Psychology, 35,* 55-81.

Appendices

A. Concept Comparison Rating Task

B. Expert Teacher Maps for "The Heart"

C. Expert Teacher Maps for "Igneous Rocks"

D. Expert Teacher Maps for "The Skeletal and Muscular System"

E. Average Teacher Maps

F. Student Pre- and Post-Reading Maps

G. Reading Passages

H. Multiple Choice Tests

I. Maze Tests

Appendix A

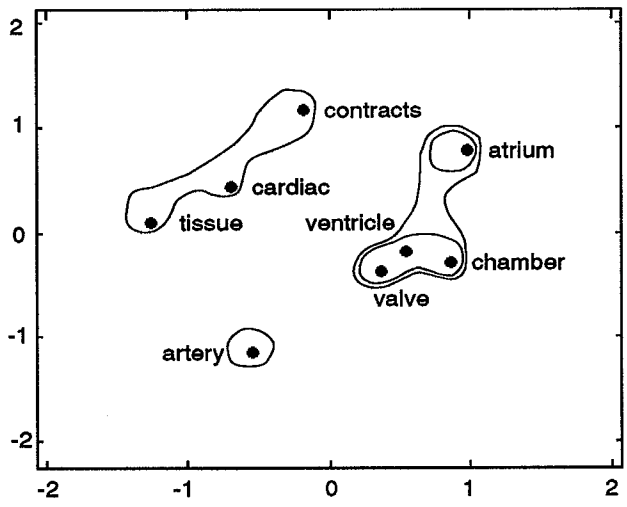
Concept Comparison Rating Task

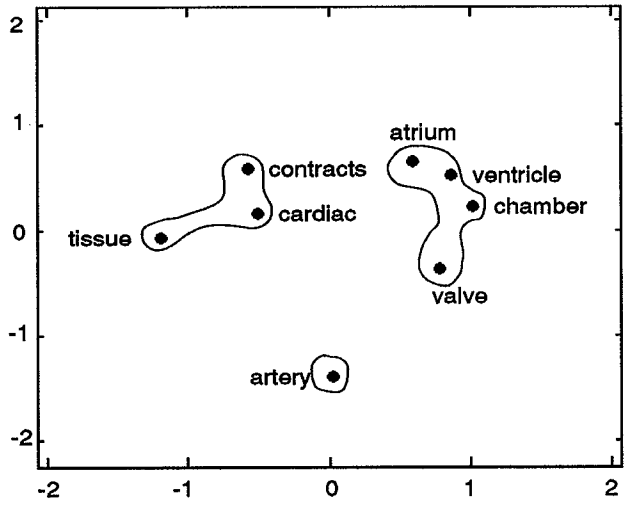
Student:_____ Grade_____ School_____
Teacher:_____ Date:____/_____/_____/
Passage: *The Heart*

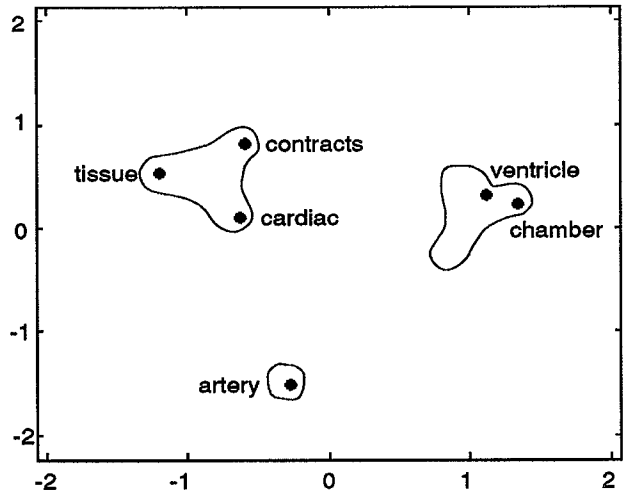
	close relation 4	3	2	little or no relation 1
atrium - cardiac				
tissue - cardiac				
tissue - chamber				
valve - chamber				
contracts - ventricle				
valve - tissue				
valve - cardiac				
artery - chamber				
ventricle - tissue				
artery - contracts				
ventricle - chamber				
cardiac - chamber				
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				

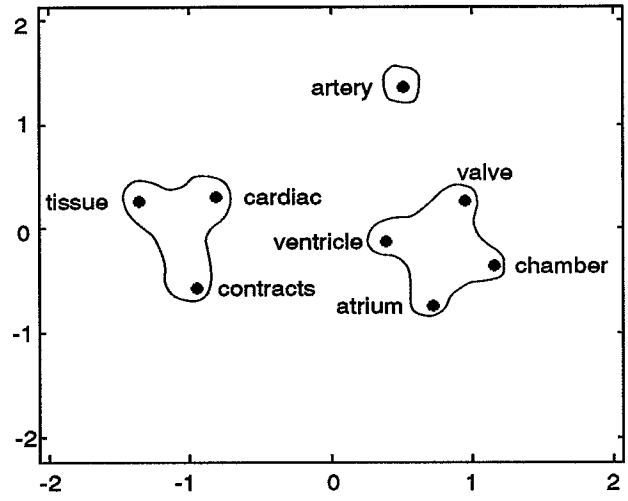
Appendix B

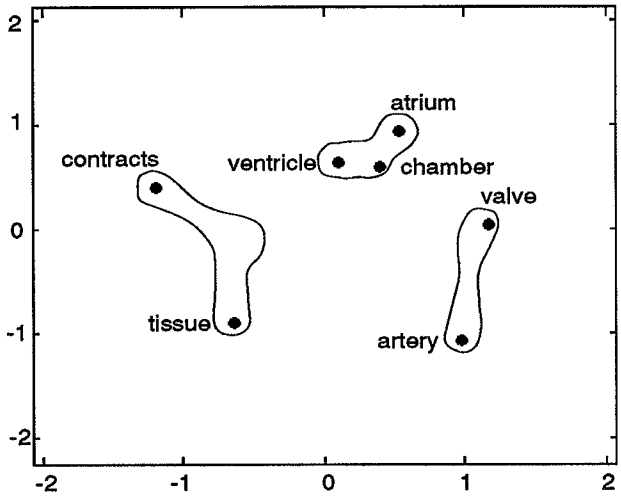
Expert Teacher Maps for "The Heart"





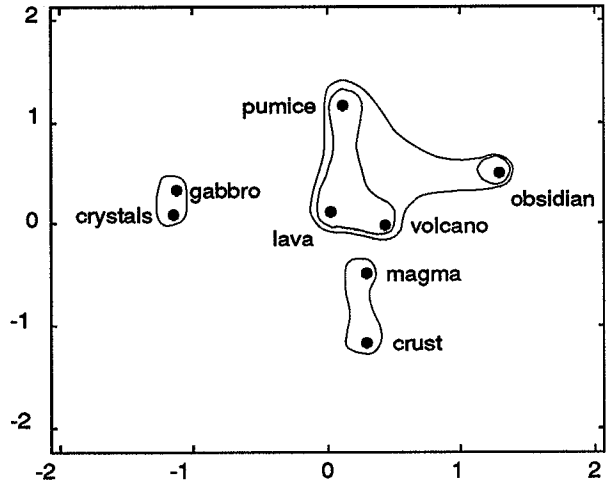


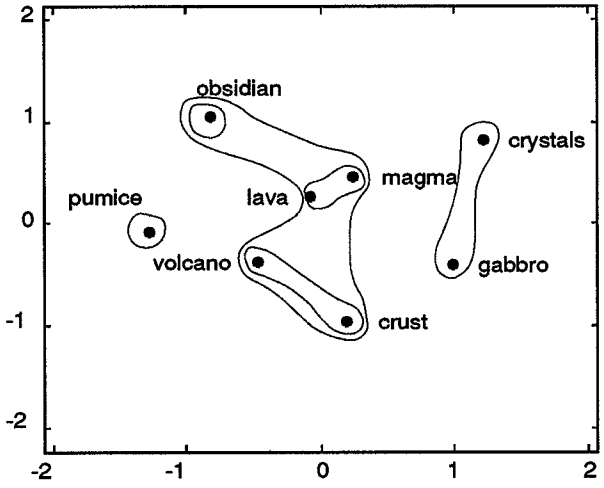


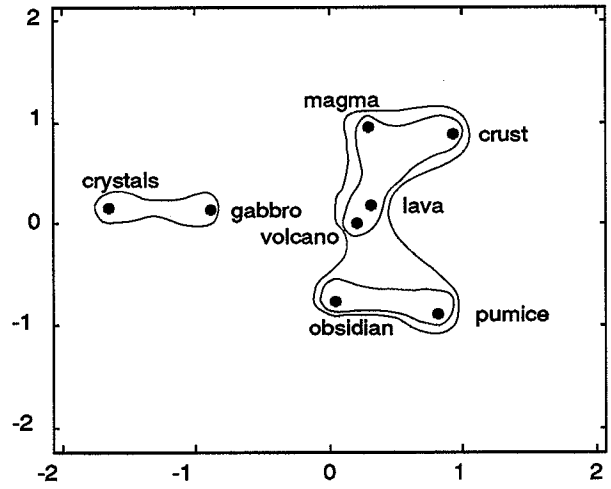


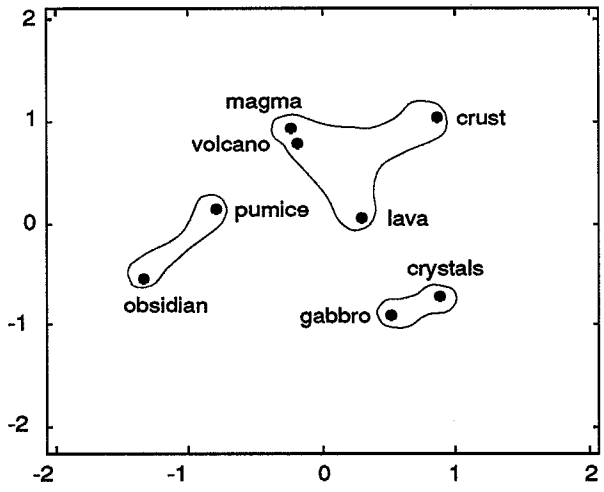
Appendix C

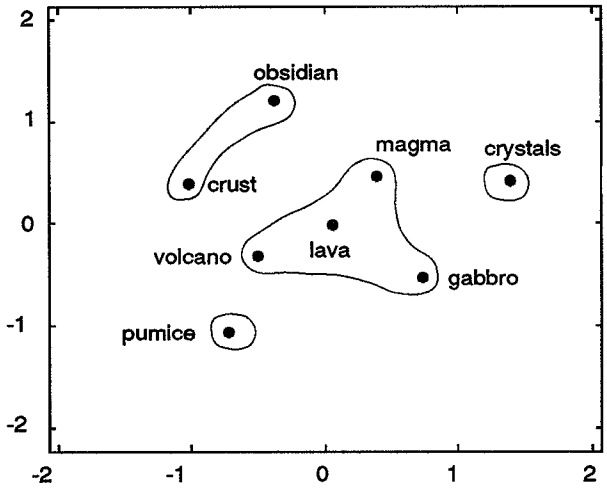
Expert Teacher Maps for "Igneous Rocks"





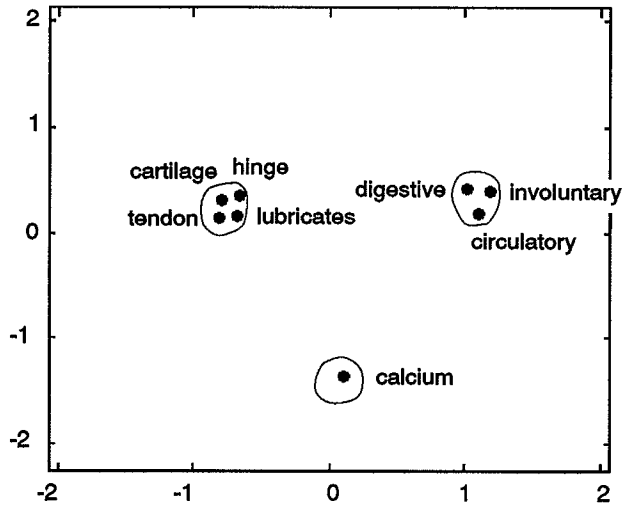


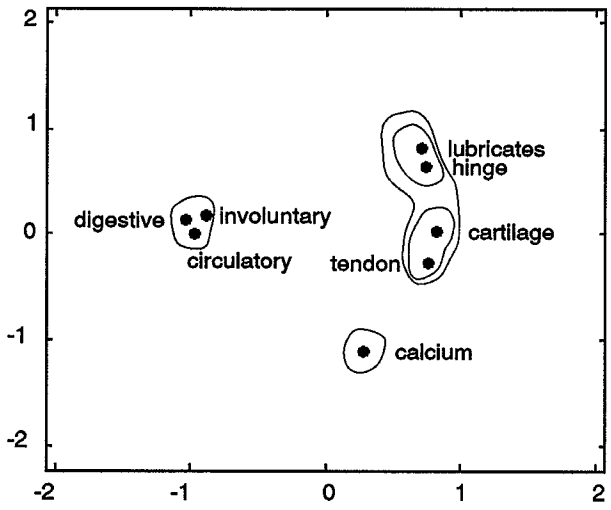


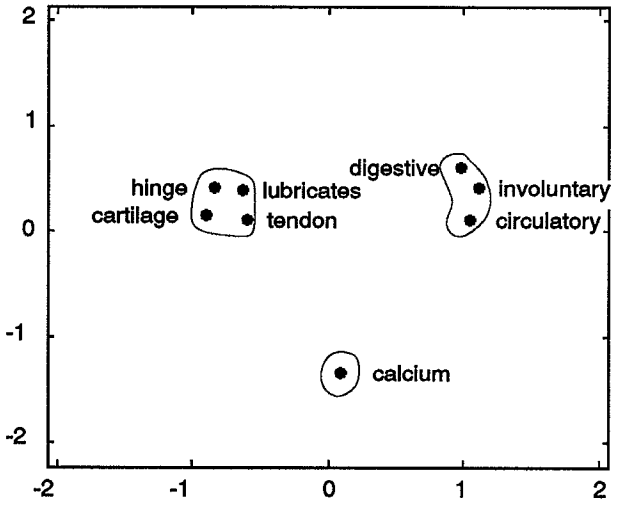


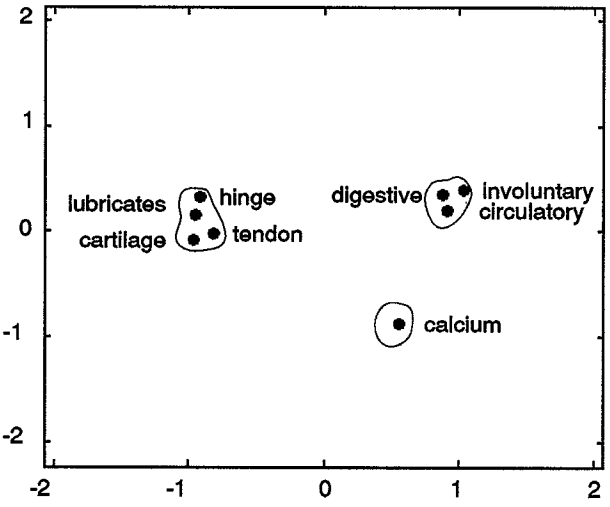
Appendix D

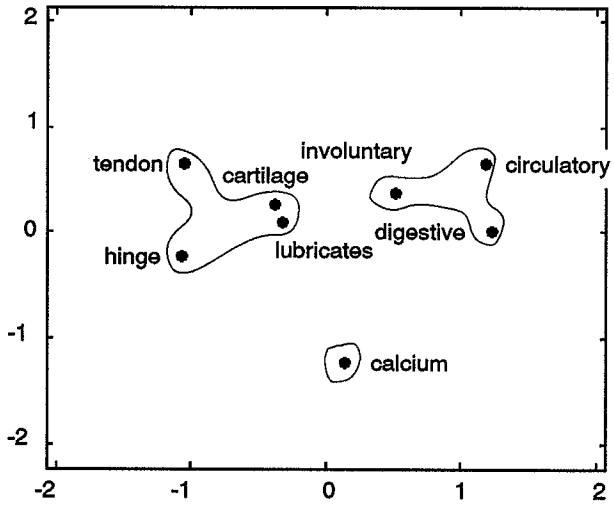
Expert Teacher Maps for "The Skeletal and Muscular System"





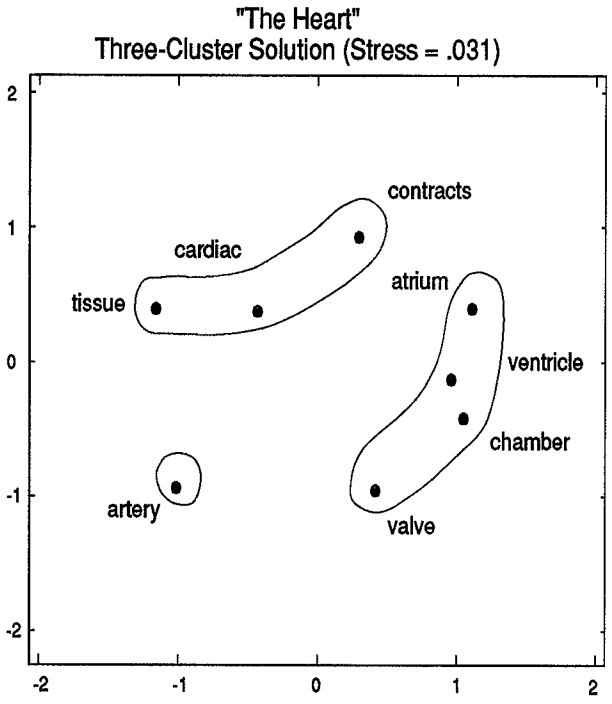


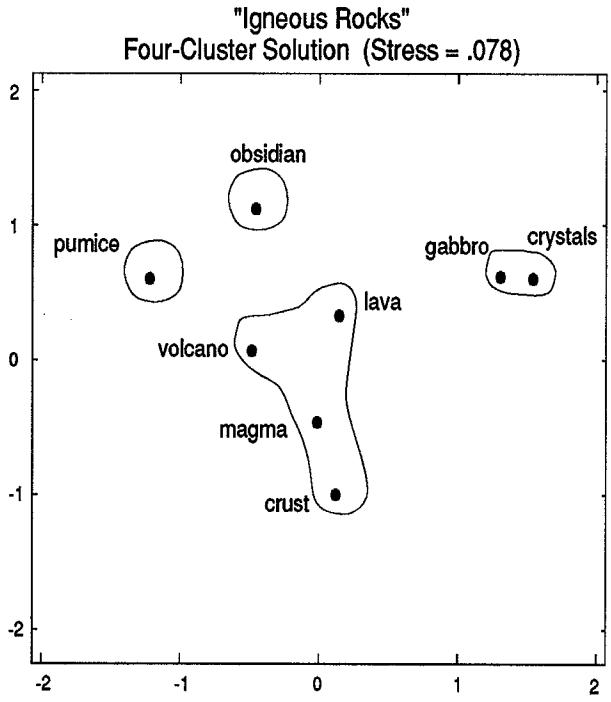


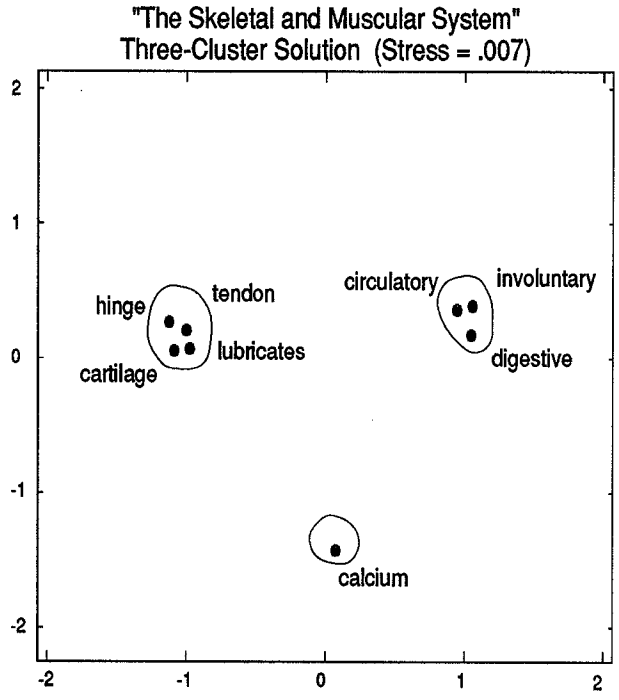


Appendix E

Average Teacher Maps





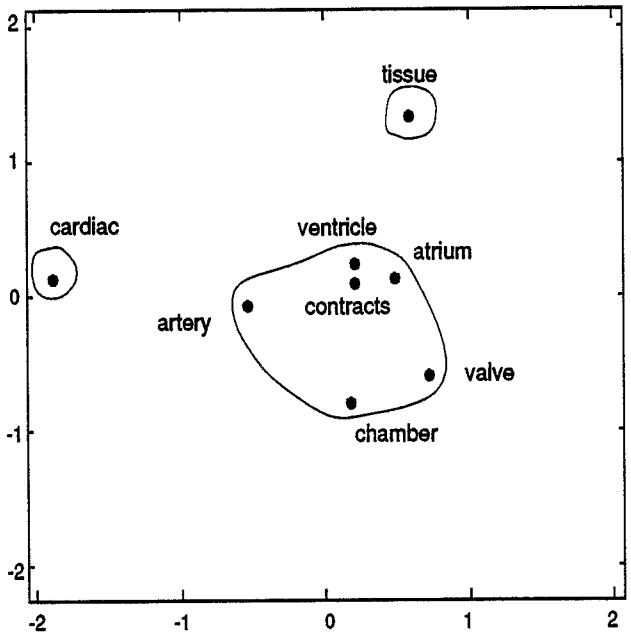


Appendix F

Student Pre- and Post-Reading Maps

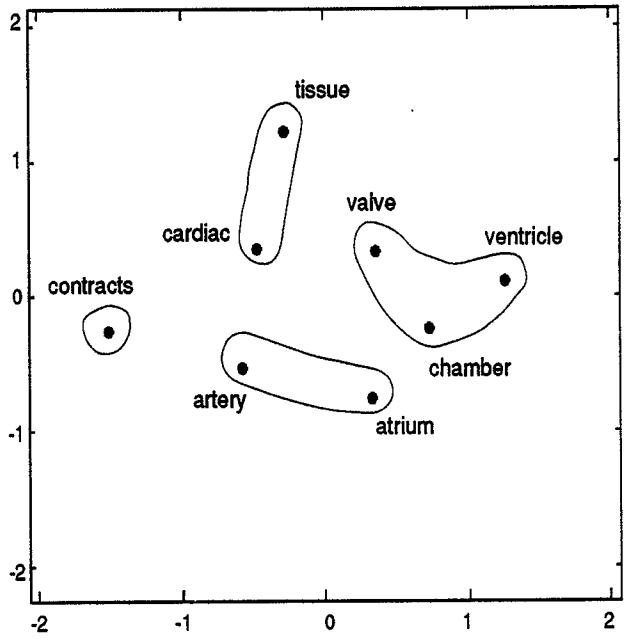
"The Heart" - Student A - Pretest

$r = .12, \tau\text{-}b = .08, \Omega = .27$ (Stress = .0095)



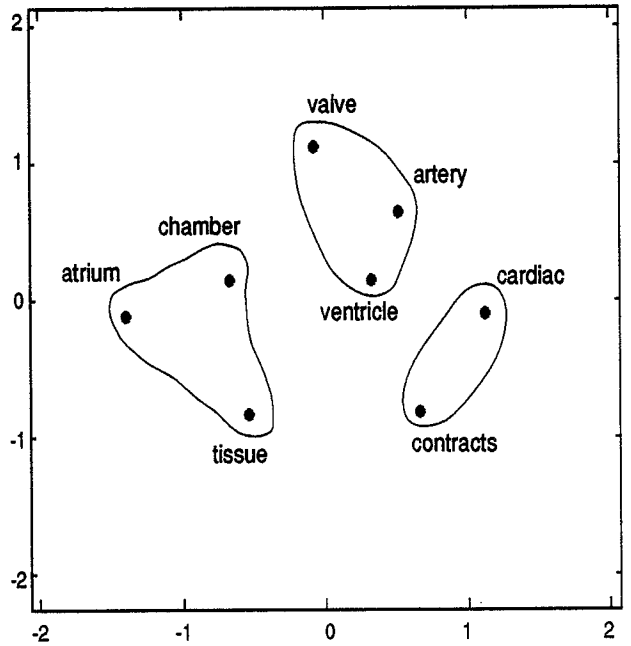
"The Heart" - Student A - Postest

$r = .46, \tau\text{-}b = .36, \Omega = .61$ (Stress = .046)



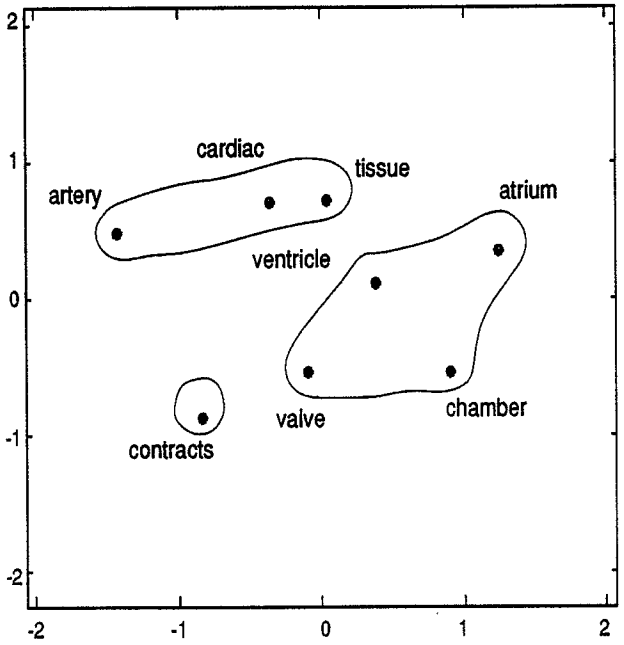
"The Heart" - Student B - Pretest

$r = -.09, \tau\text{-}b = .12, \Omega = .33$ (Stress = .064)



"The Heart" - Student B - Postest

$r = .39, \tau\text{-}b = .40, \Omega = .74$ (Stress = .079)

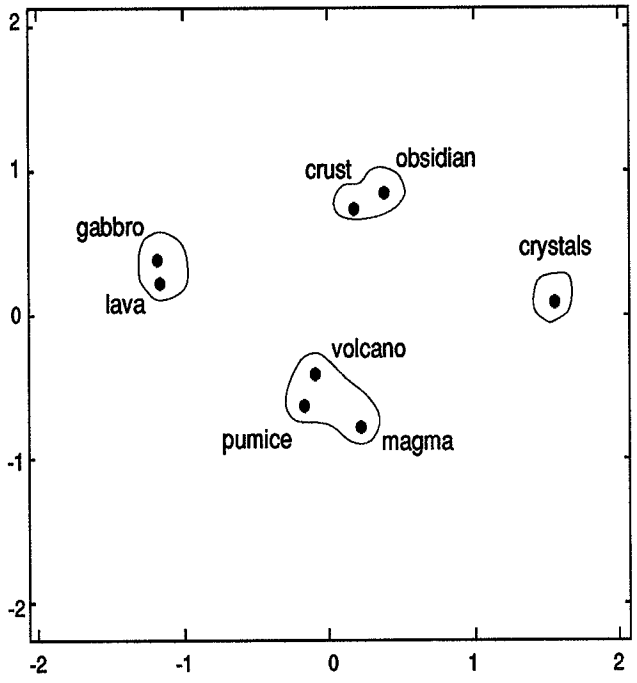


Appendix F (cont.)

Student Pre- and Post-Reading Maps

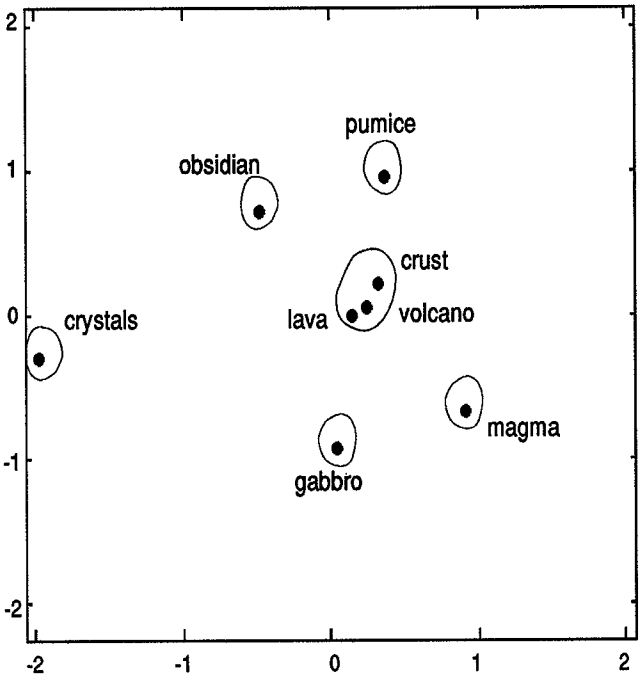
"Igneous Rocks" - Student C - Pretest

$r = .08$, $\tau - b = -.02$, $\Omega = .13$ (Stress = .023)



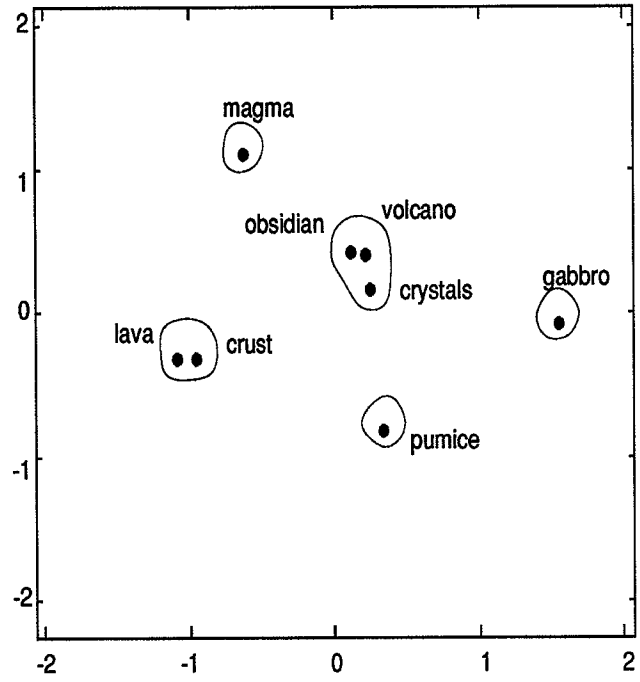
"Igneous Rocks" - Student C - Postest

$r = .44$, $\tau - b = .46$, $\Omega = .69$ (Stress = .023)



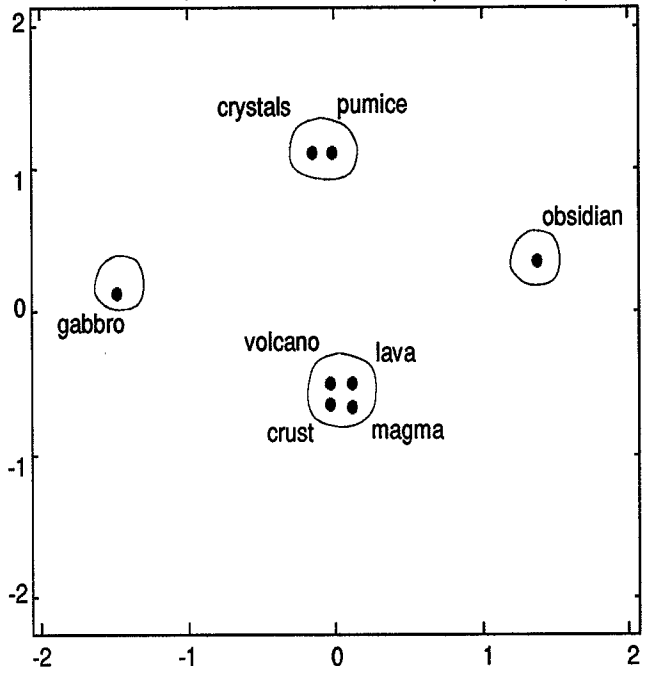
"Igneous Rocks" - Student D - Pretest

$r = .18$, $\tau - b = -.04$, $\Omega = .33$ (Stress = .024)



"Igneous Rocks" - Student D - Postest

$r = .39$, $\tau - b = .42$, $\Omega = .79$ (Stress = .011)

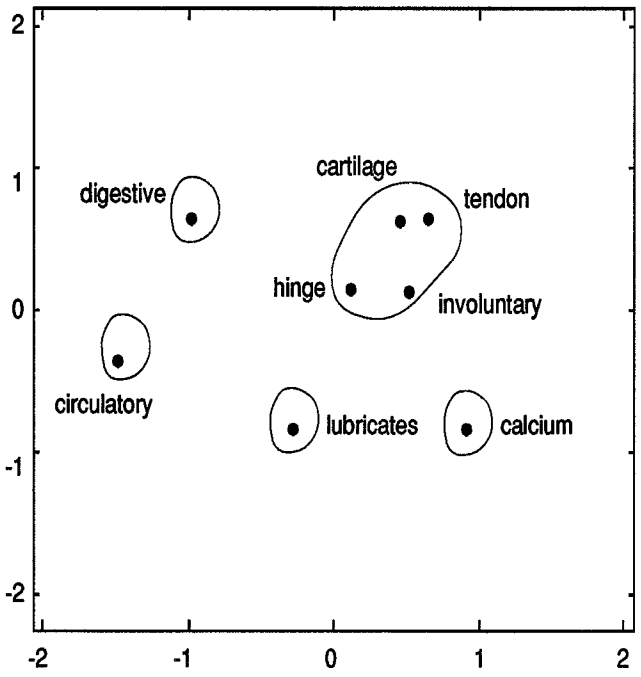


Appendix F (cont.)

Student Pre- and Post-Reading Maps

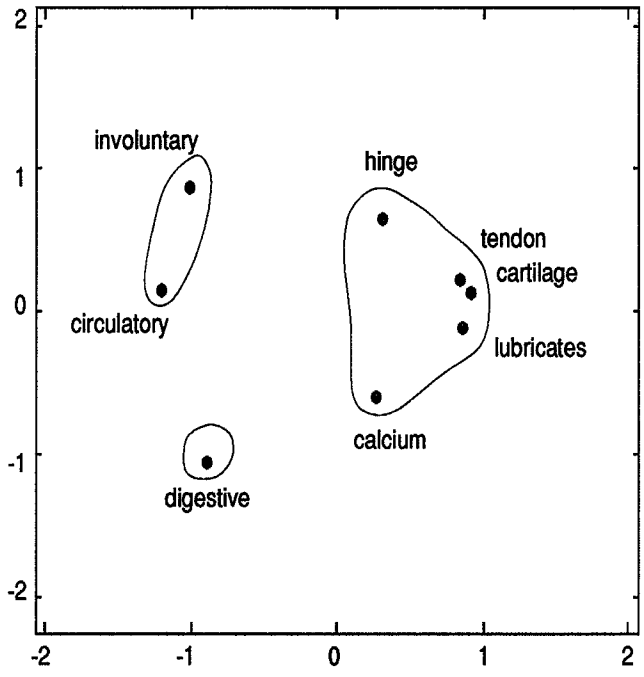
"Skeletal & Muscular System" - Student E - Pretest

r = .038, τ - b = .20, Ω = .39 (Stress = .025)



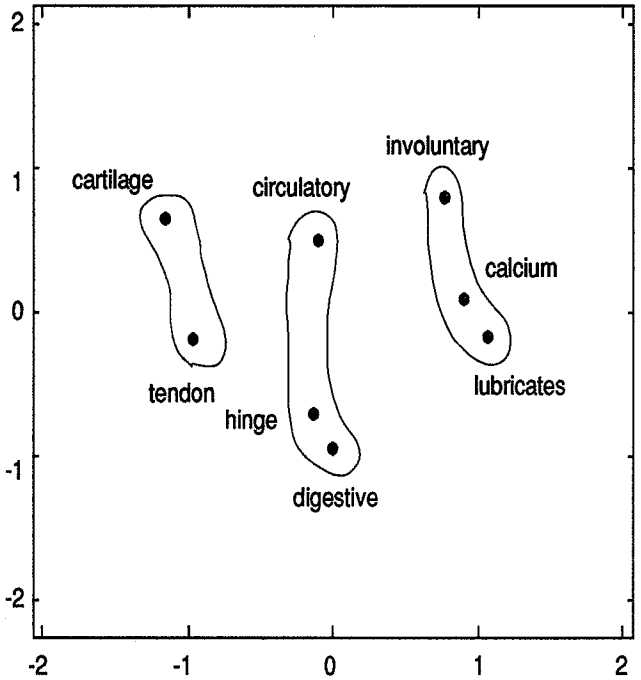
"Skeletal & Muscular System" - Student E - Postest

r = .56, τ - b = .67, Ω = .62 (Stress = .0000)



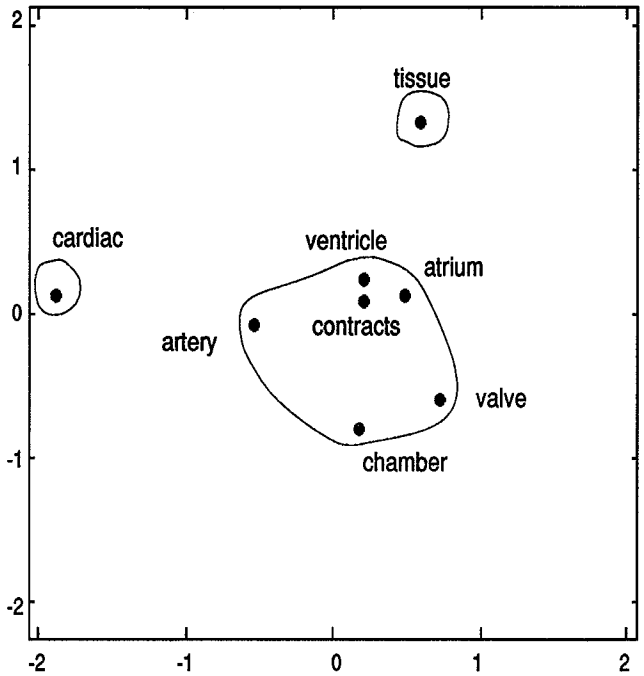
"Skeletal & Muscular System" - Student F - Pretest

r = .034 τ - b = .20, Ω = .20 (Stress = .026)



"The Heart" - Student A - Pretest

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



Appendix G

Reading Passages

The Heart
(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 surface 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 G (cont.)
Reading Passages

Limits on Animal Populations
(Holt General Science, pp. 491-493)

Scientists are always observing what happens to populations within communities. They try to find out what populations are present, the size of each population, and if the numbers are changing. They also try to find out which species is the dominant species.

By the 1900's there were four moose for each square kilometer of land on Isle Royale. The density of the moose population was so great that it caused a food shortage. In the next few years, all but a few hundred moose died. Most of them died of starvation. Such a sudden decrease in numbers is called a population crash.

In the summer of 1936, a forest fire burned over a quarter of the island. Over the next several years, the burned area began to fill with lichen and moss. The herd began to thrive on this food source. The number of moose steadily increased again. By the 1940's park rangers again began to find dead moose.

Moose are not the only animals that have extreme population changes. A well-known example is a tiny mouselike rodent, the lemming. Every three or four years, the population of lemmings gets very large. They must migrate to find new food sources. During these migrations, thousands plunge into the ocean and drown.

The density of the moose and lemming populations became a limiting factor. The high density resulted in a shortage of food. There are other limiting factors involved with population changes. Higher population densities also cause an increase in disease. Diseased animals become weakened and are easy prey for predators.

[259 words]

One-Celled Organisms
(Holt Science, pp. 43-47)

In the summer, a thick, green scum may cover part of a pond. The scum is so thick, you cannot see through it. Pond scum is really chains and chains of plant-like cells. The cells contain little green bits called chloroplasts. The chloroplasts are shaped like coils. They give the pond scum its color, and also help make food.

Is pond scum a kind of plant? No, not quite. Pond scum is a kind of algae. Like plants, algae have chloroplasts. But algae do not have stems, leaves, and roots as true plants do.

Some algae are one-celled organisms. Often, they float by themselves in water. Some, like pond scum, live together in chains. Others live together in large groups and have root-like ends. These help them hold on to rocks or soil.

When animals are hungry, they eat other organisms. Most plants, however, do not need to eat other organisms. Most plants can make their own food. Algae can make their own food, too. Algae make their food out of non-living things.

To make food, a cell needs energy. The chloroplasts inside a cell trap energy from the sun. Next, water is needed. Water enters a cell through the cell wall. Finally, a type of gas, called carbon dioxide, is needed. This gas enters the cell through the cell wall. Using energy from the sun, the cell turns the water and the gas into food and oxygen. The food can then be used by the cell. The oxygen can be used by other living things.

[256 words]

Appendix G (cont.)
Reading Passages

The Seashore
(Heath Life Science, pp. 248-249)

Two kinds of seashores are the rocky shore and the sandy shore. Due to wave action, many plants and animals living on rocky seashores have special structures for clinging to the rocks. With tube feet, starfish attach to rocks. Mussels make ropelike strands that anchor them during tide changes. Kelp have special cells at their bases which hold them to rocky coasts.

When the tide is out, starfish and crabs follow the water's movement. To stay moist, mollusks and crustaceans living high on rocks withdraw inside their coverings. Kelp have tough, leathery structures that prevent them from losing water at low tide.

Because sand shifts constantly, sandy seashores have fewer organisms than rocky seashores. Nearly all organisms live or hide in burrows when the tide is out. Parchment worms and lugworms dig U-shaped burrows. Mollusks burrow into the sand and leave only their breathing tubes above the surface.

The seashore is really a land platform that extends into the ocean. This land platform is the flat part of the ocean floor, called the continental shelf. The outer edge of the continental shelf ends in a deep slope. Since sunlight reaches the waters over the shelves, the waters of this ecosystem are rich in plankton. In addition, nutrients from the continents are washed into the continental shelves. The rich plant life and the movement of nutrients throughout the water support large populations of fish, sharks, crabs, scallops, oysters, and coral in this ecosystem.

[242 words]

Regions of the Soviet Union
(Heath Social Studies: The World Past to Present, pp. 366-367)

People in the Soviet Union speak dozens of different languages. They live in many different ways. Some work at jobs that did not exist a few years ago. A few live much as their ancestors did.

The Soviet Union has five regions, each with a different kind of climate. As a result, the country has five very different kinds of vegetation.

The region farthest north is called the tundra. Here winter lasts from October to May. For more than a month of this time, the sun does not rise above the horizon in the northern part.

South of the tundra lies a forest region that is as large as the entire United States. This immense forest region is known as the taiga. In the northern part of the taiga, most of the trees are fir, larch pine, and other evergreens.

Much of the taiga has a continental climate. Temperatures in the taiga vary greatly from summer to winter because of taiga's location. It lies far from any body of water. Winds that blow over this region heat up and cool off more quickly than winds that blow over water. So winters in Moscow are much colder than winters in coastal cities.

Most of the farms in the Soviet Union lie on the steppe. The steppe runs from the Black Sea to the borders of China. The black fertile soil of the steppe, especially in an area known as the Ukraine, is among the best in the world for farming.

[249 words]

The History of Texas
(Heath Social Studies: Regions Far and Near, pp. 227-228)

By the 1800's, Spain had been ruling much of the Southwest for nearly 300 years. In 1803, the United States bought lots of land bordering on the Southwest. It was called the Louisiana Territory. Settlers from the U.S. poured into the Louisiana Territory. South of the Territory lay Texas, which was mostly wilderness. By 1820, the Territory became crowded, and leaders of the Spanish government allowed people from the U.S. to settle in Texas.

Five years later, Mexico won its freedom from Spain. Mexico's new leaders were pleased with the U.S. settlers in Texas. By 1830, almost 30,000 Americans had moved to Texas. Some of the settlers raised cattle and sheep on the grasslands. Others grew cotton on large plantations.

The people of Texas paid taxes to Mexico. In return, Texans could vote for Mexican leaders. For a while, all went well. However, peace in Texas did not last for very long.

By the 1830's, Mexico's leaders were beginning to worry about Texas. To keep Texas from becoming too much like the U.S., Mexico's government passed two new laws. One law stopped more people from the U.S. from moving to Texas. The other law stopped Texans from trading with the U.S.

In 1834, Santa Anna took over the government of Mexico. Santa Anna raised the people's taxes and took away their right to vote. Both Mexican and American Texans decided to break away from Mexico and form their own country.

[247 words]

Appendix H

Multiple Choice Tests

The Heart

1. Your heart is made of ...

(a) cardiac muscle
(b) blood vessels.
(c) atriums and vessels
(d) arteries and blood vessels.

2. Your heart pumps about____ times per minute.

(a) 10
(b) 30.
(c) 120.
(d) 70.

3. The heart is really ...

(a) one large pump
(b) four pumps.
(c) two pumps
(d) not a pump, but a ventricle.

4. There is a valve between each atrium and ventricle so blood can flow ...

(a) both directions at once
(b) from a ventricle to an atrium.
(c) to the lungs
(d) from an atrium to a ventricle.

5. Arteries are vessels that ...

(a) carry blood away from the heart
(b) carry blood to the ventricles.
(c) carry blood between chambers of the heart
(d) return blood to the heart.

6. The upper chambers of the heart ...

(a) pump blood to the arteries
(b) receive blood from the lungs or body.
(c) pump blood to the lungs or body
(d) receive blood from the arteries.

7. Your heart is found ...

(a) in the right side of your chest
(b) in the left side of your chest.
(c) in the center of your chest
(d) in the atrium of your chest.

8. Which statement best describes the main idea of the story ?

(a) The heart's many parts each do a specific job.
(b) The heart is delicate, and can easily be damaged by habits like smoking.
(c) The heart and the atrium are important partners.
(d) The heart and lungs are two important organs.

9. Which best describes the shape of your heart ?

(a) circle-shaped
(b) triangle-shaped.
(c) cone-shaped
(d) atrium-shaped.

10. Which best describes the parts of your heart ?

(a) four chambers: two atriums and two ventricles.
(b) two chambers: one atrium and one ventricle.
(c) three parts: an atrium, a ventricle, & an artery .
(d) two chambers: one upper and one lower.

Appendix H (cont.)
Multiple Choice Tests

Igneous Rocks

1. *Melted rock under the earth's surface is called ...*
   (a) obsidian
   (b) magma.
   (c) lava
   (d) gabbro.

2. *When melted rocks cool quickly ...*
   (a) large crystals form
   (b) pumice crystals form.
   (c) no crystals form
   (d) glass crystals form.

3. *Obsidian could also be called ...*
   (a) glass rock
   (b) gabbro.
   (c) magma
   (d) melted crystals.

4. *When minerals are solid, they form ...*
   (a) obsidian
   (b) lava.
   (c) pumice
   (d) crystals.

5. *When lava cools fast, hot gases ...*
   (a) cause the lava to rise to the earth's surface
   (b) become trapped in the lava.
   (c) cause the lava to explode
   (d) form crystals.

6. *The appearance of pumice is ...*
   (a) covered with large crystals
   (b) glass-like.
   (c) black and heavy
   (d) spongy and light in color.

7. *Find the word that best fits this definition: "Rocks that form from melted material that cools and hardens."*
   (a) igneous rocks
   (b) gabbro.
   (c) pumice
   (d) obsidian.

8. *Which statement best describes the main idea of the story ?*
   (a) Several types of rocks are formed by heating and cooling
   (b) Inactive volcanos are a excellent rock-hunting sites.
   (c) Our earth produces rare and valuable minerals.
   (d) The three major igneous rocks: Gabbro, Pumice, and Lava.

9. *What two main substances come from under the earth's crust?*
   (a) gabbro & lava
   (b) lava & gases.
   (c) gases & gabbro
   (d) pumice & obsidian.

10. *Gabbro has ...*
   (a) large crystals of many minerals
   (b) a spongy look, with many holes.
   (c) a smooth, glass-like appearance
   (d) gases trapped in glass.

Appendix H (cont.)

Multiple Choice Tests

The Skeletal and Muscular Systems

1. The human skeleton is made up of ...

(a) cartilage and calcium.
(b) joints and tendons.
(c) bones and sockets
(d) bone and cartilage.

2. Bones contain ...

(a) cartilage and calcium
(b) ligaments and calcium.
(c) calcium and phosphorus
(d) ligaments and tendons.

3. Some bones are connected by _____, which can easily stretch.

(a) joints
(b) ligaments.
(c) cartilage
(d) muscles.

4. All muscles work by ...

(a) contracting and expanding
(b) pulling and pushing.
(c) only contracting
(d) only extending.

5. Movement of the joints and other parts of the body is caused by ...

(a) ligaments
(b) muscles.
(c) cartilage
(d) bones.

6. The inside surface of most joints is covered with ...

(a) cartilage
(b) calcium.
(c) tendons
(d) ligaments.

7. Some involuntary muscles are found in the ...

(a) arms, legs, and back muscles
(b) muscles that aid in movement.
(c) muscles that straighten joints muscles.
(d) digestive and circulatory

8. Which joints allow rotational movement?

(a) hinge joints
(b) ball-and-socket joints.
(c) ligament joints
(d) voluntary joints.

9. About how many bones are in the human skeleton?

(a) 200
(b) 500.
(c) 1,000
(d) 35.

10. A set of muscles pulls on a tendon to bend a ...

(a) ligament
(b) muscle.
(c) cartilage
(d) joint.

APPENDIX I

MAZE TESTS

Student_____ Grade_____ School_____

Teacher_____ Date___/___/____

The Heart

Your heart is a cone-shaped organ that is found in the middle of your chest. The heart is about the (1) of a large fist. You (2) think that pumping blood through (3) entire body is a big (4) for such a small organ. (5) your heart is made of (6) special tissue called cardiac muscle. (7) strong muscle enables the heart (8) pump every second of the (9) without getting tired. In fact, (10) heart pumps between 60 and (11) times a minute every day. (12) adult heart pumps about 5 (13) of blood each minute!

The (14) is really two pumps that (15) side by side. The right (16) is separated from the left (17) by a muscular wall. There (18) four compartments or chambers in (19) heart. Each upper chamber is (20) an atrium. An atrium is (21) small, thin-walled chamber that receives (22) from the lungs or the (23). Each lower chamber is called (24) ventricle. A ventricle is a (25), muscular chamber that pumps blood (26) the lungs or the body.

(27) is a valve between each (28) and ventricle. The valve works (29) a one-way door. Blood can (30) flow from an atrium to (31) ventricle. Blood in the ventricle (32) never flow back into the (33) because the valve closes when (34) ventricle contracts.

Different kinds of (35) vessels carry blood through the (36) . One kind of vessel is (37) an artery. Arteries are blood (38) that carry blood away from (39) heart. The walls of arteries are very elastic.

DIRECTIONS:
For each blank in the story, circle the word below that fits the best.

(1) blood, atrium, lie, size, are
(2) a, may, heart, pump, day
(3) a, the, atrium, there, to
(4) special, the, job, blood, day
(5) Blood, Like, This, Size, But
(6) called, an, a, are, size
(7) This, To, Atrium, Pump, Liters
(8) pump, body, like, may, to
(9) may, lie, a, the, day
(10) but, the, called, thick, your
(11) only, liters, 80, a, day
(12) Lie, Pump, 80, Liters, An
(13) may, liters, job, blood, like

(14) size, job, called, heart, a
(15) but, blood, to, lie, job
(16) vessels, only, a, job, pump
(17) your, body, thick, there, pump
(18) called, are, only, a, the
(19) size, but, 80, the, may
(20) body, there, like, called, your
(21) your, job, a, vessels, special
(22) blood, an, lie, liters, pump
(23) called, body, only, size, liters
(24) to, only, like, a, job
(25) heart, like, thick, called, blood
(26) there, to, the, an, lie

(27) Pump, Atrium, Only, Heart, There
(28) body, pump, atrium, your, job
(29) like, pump, this, your, may
(30) pump, there, a, heart, only
(31) atrium, a, to, an, body
(32) can, body, special, size, pump
(33) pump, atrium, your, lie, this
(34) like, but, 80, there, the
(35) special, lie, day, only, your
(36) thick, the, blood, body, this
(37) heart, called, but, job, liters
(38) body, your, like, may, vessels
(39) the, a, may, atrium, like

APPENDIX I (CONT.)

MAZE TESTS

Student_____ Grade_____ School_____

Teacher_____ Date___/___/_____

Igneous Rocks

Heat deep inside the earth causes some rocks to melt. Red-hot, melted rock under the (1) surface is called magma. Sometimes, (2) magma pushes out through a (3) or a weak spot in (4) earth's crust. Red-hot melted rock (5) out of the earth is (6) lava. The lava piles up, (7) , hardens, and forms a mountain (8) solid rock. This kind of (9) is called a volcano.

Rocks (10) form from melted material that (11) and hardens are called igneous (12) . The word igneous means "coming (13) fire". Hardened lava is one (14) of igneous rock. The way (15) rock looks depends on how (16) the lava cooled.

The lava (17) slowly as a volcano becomes (18) . Rocks formed by the slow (19) of melted material have large (20) . Crystals are the structures that (21) form when they are solid. (22) is an igneous rock that (23) large crystals of many minerals.

(24) active volcanos, the lava is (25) with hot gases. The lava (26) , or erupts, through a small (27) in the earth's surface. When (28) happens, the hot material often (29) quickly. There is no time (30) crystals to form. The lava (31) and looks like a glass (32) . This kind of rock is (33) obsidian.

At times, lava cools (34) fast that the hot gases (35) with the lava do not (36) time to escape. They become (37) inside the hardened lava and (38) a spongy rock light in (39) . This kind of igneous rock is called pumice.

DIRECTIONS:
For each blank in the story, circle the word below that fits the best.

(1) gabbro, earth's, explodes, have, color
(2) called, of, cools, has, the
(3) mixed, color, so, crack, crystals
(4) cooling, cools, hole, the, called
(5) called, mountain, coming, gabbro, explodes
(6) called, cools, fast, inactive, for
(7) this, cools, called, explodes, color
(8) explodes, rock, of, have, so
(9) gabbro, called, mountain, cools, crack
(10) from, crystals, hardens, have, that
(11) fast, cools, inactive, mountain, so
(12) rocks, hardens, color, earth's, the
(13) explodes, hardens, from, cooling, trapped
(14) trapped, this, color, called, kind
(15) crystals, trapped, so, the, for
(16) kind, rock, fast, mixed, from
(17) in, cooling, cools, this, mountain
(18) mixed, trapped, cooling, form, inactive
(19) form, gabbro, cooling, inactive, have
(20) hole, crack, explodes, crystals, hardens
(21) color, form, the, of, minerals
(22) Rock, Form, Minerals, Gabbro, Crack
(23) color, has, have, fast, mountain
(24) Of, Coming, In, This, The
(25) has, hole, explodes, called, mixed
(26) from, called, inactive, explodes, inactive
(27) hardens, crystals, crack, hole, kind
(28) earth's, rock, crack, this, so
(29) cools, called, the, that, rocks
(30) this, so, for, explodes, crack
(31) cooling, the, of, crystals, hardens
(32) gabbro, trapped, rock, that, rocks
(33) called, mixed, this, inactive, of
(34) from, the, called, so, mixed
(35) minerals, cooling, mixed, gabbro, cools
(36) hardens, color, form, have, the
(37) earths, trapped, this, fast, has
(38) coming, color, form, crystals, in
(39) called, from, mountain, color, hole

APPENDIX I (CONT.)

MAZE TESTS

Student_____ Grade_____ School_____

Teacher_____ Date___/___/____

The Skeletal and Muscular Systems

Organs working together make up systems. Two of these systems are (1) skeletal system and the muscular (2).

The human skeleton is made (3) of bone and cartilage. One (4) between the two is that (5) does not contain the calcium (6) phosphorus compounds that bone contains. (7) makes cartilage more flexible than (8).

There are 206 bones in (9) human skeleton. Some of these (10) are connected to each other (11) ligaments. Since ligaments stretch easily, (12) allow the bones to move (13). This forms what is called (14) movable joint.

Joints can allow (15) in different directions. A hinge (16) allows back and forth movement. (17) ball and socket joint allows (18) movement. The inside surface of (19) joints is covered with cartilage. (20) also contain a special fluid (21) lubricates them so they do (22) wear each other away.

Movement (23) the joints and other parts (24) the body is caused by (25) muscles. The muscles of the (26) and legs are examples of (27) that aid us in movement. (28) are called voluntary muscles. There (29) some muscles like the ones (30) in the digestive, respiratory, (31) circulatory systems that are involuntary.

(32) muscles work only by contracting. (33) they only work by contracting, (34) can only pull. They cannot (35). 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.

DIRECTIONS:
For each blank in the story, circle the word below that fits the best.

(1) most, at, muscles, the, since
(2) difference, joints, system, found, all
(3) all, up, arms, and, push
(4) rotational, the, difference, they, by
(5) they, joints, not, muscles, cartilage
(6) found, all, they, joint, or
(7) This, These, They, Bone, Most
(8) they, freely, bone, the, system
(9) that, these, the, found, all
(10) bones, difference, joint, are, cartilage
(11) or, of, by, they, push, muscles
(12) movement, rotational, they, push, found
(13) rotational, joints, they, joint, freely
(14) that, a, cartilage, movement, bone
(16) are, joint, a, they, arms
(17) Not, Rotational, Muscles, A, Up
(19) freely, this, movement, system, most
(20) Joints, System, That, At, The
(21) joints, bones, since, push, that
(22) and, not, since, all, this
(23) since, system, freely, joints, of
(24) found, all, they, cartilage, of
(25) cartilage, the, bones, joint, arms
(26) these, arms, bones, rotational, bone
(28) Are, Movement, These, Difference, Or
(29) are, not, found, movement, freely
(30) the, a, movement, and, found
(31) joints, all, bones, and, rotational
(32) Not, Found, Muscles, Freely, All
(33) Freely, They, Since, At, System
(34) or, this, bones, that, they
(35) most, rotational, muscles, up, push