Chapter 17

A Structural Assessment of Classroom Learning

Timothy E. Goldsmith and Peder J. Johnson

A most basic and long standing concern of philosophers, psychologists, and educators is the problem of knowledge elicitation and representation. How do we assess and represent an individual's knowledge? Philosophers, when asking these questions, have usually expressed an interest in general or world knowledge. Psychologists and educators, on the other hand, have often been more interested in the problem of assessing and representing a person's knowledge of some particular topic or area. It is this problem, as it arises in the assessment of classroom types of knowledge, that is the concern of the present chapter. Knowledge assessment and representation, as carried out in the classroom, appears as a relatively straightforward matter. Knowledge is assessed by simply asking factual questions and is represented by presenting the individual's relative standing in terms of a percentile. We begin with a critique of this conventional approach to assessing and representing classroom knowledge.

The two processes, assessment and representation, are obviously related. In our view the approach to representation is more fundamental in that assumptions regarding the organization of knowledge have implications for how we assess knowledge. The representation of classroom knowledge is usually in terms of percentage correct, or in the case of standardized tests, performance may be converted to a percentile ranking within some designated population. In some instances performance may be analyzed into subscales, but for the most part classroom learning is represented in terms of a unidimensional scale reflecting the student's relative standing in her class. Percentage correct may be perfectly adequate for representing certain types of knowledge (e.g., a student's knowledge of the capitals of the 50 states) where the conceptual relationships among the knowledge elements are not particularly relevant. In this case it may be safe to assume that the "facts" comprising the domain are independent and additive. An important property of more interesting domains of knowledge involves the relationships or organization of the elements which compose the domain. We contend that for domains of this type it is this *configural* property of knowledge that must be assessed and represented. A percentile ranking may be very convenient in assigning grades to students, but it tells us very little regarding what the student knows or does not know. The fundamental problem with conventional educational assessment procedures is that they are in principle, incapable of explicitly representing the abstract nature of conceptual domains. A basic assumption in cognitive psychology is that knowledge entails an understanding of the interrelationships among concepts and that this organizational property of knowledge can best be captured with structural representations (e.g., Bower, 1972; Collins & Quillian, 1969). It is our aim to develop and evaluate a set of procedures that capture this configural property of knowledge.

Turning to the problem of assessing classroom knowledge, historically, evaluation has occurred by presenting various recognition (e.g., true-false or multiple choice) or recall (e.g., essay) types of questions that are directly related to the relevant content area. It is

generally recognized that this method of assessment has several potential problems. First, performance on recall and even recognition tests may be impaired by retrieval problems. Because conventional test procedures depend very heavily on episodic memory, they are subject to the influence of context specific cues on retrieval processes. As a consequence we often hear students complain that they knew the answer but could not recall it, or they did not know what was being asked or that they misunderstood the question. All of these complaints have sufficient validity that should concern us.

A second problem is that much of what we expect students to learn is implicit knowledge that is difficult, if not impossible, to state in an explicit manner. To the extent that some relevant domain-specific knowledge is implicit, it may be extremely difficult to assess that knowledge with direct questions. This problem is most obvious in the case of procedural knowledge. Third, there is the problem of developing tests that are objective and easily scored, while at the same time assessing the more abstract or conceptual aspects of knowledge. This is the obvious tradeoff between multiple choice and essay questions. Multiple-choice exams, while objective and easy to score, require considerable expertise in their development if they are to assess conceptual knowledge. On the other hand, essay questions may assess conceptual knowledge more readily, but the scoring is likely to require a considerable degree of expertise. For this reason essay exams are often considered too time-consuming and resource demanding to be used on a large scale.

To summarize, we are suggesting that the ideal approach for assessing and representing classroom knowledge would be objective, reliable, require minimal retrieval demands, and most important reflect the student's conceptual organization of the domain. We now turn to some earlier research that takes a structural approach to assessment and representation of classroom learning.

Structural Assessment Approaches to Learning

The limitations of conventional testing methods point to the need for procedures that assess and represent the conceptual properties of classroom learning. This need was clearly recognized as early as 1972 by Shavelson and his colleagues (Geeslin & Shavelson, 1975; Shavelson, 1972, 1974; Shavelson & Geeslin, 1973; Shavelson & Stanton, 1975) who took the view that classroom instruction is most properly seen as the communication of a specific structure that is implicit within the curriculum of a subject matter. It was Shavelson's assumption that "a structure of a subject matter, ultimately, rests in the minds of the 'great scientists.' This structure is communicated through the scientists' writings in journals and advanced textbooks as well as through informal communication channels" (Shavelson, 1974, p. 232).

Shavelson's goal was to assess the effects of classroom instruction by determining whether the student's organization of the material became more similar to an expert's over the course of instruction. In his research Shavelson attempted to assess the structure of a classroom topic (e.g., high school physics) using a variety of techniques, such as word associations, card sorting, and a graph construction method. Distance measures were calculated on the basis of these data, which in turn were subject to a hierarchical cluster analysis to derive an underlying structure.

Shavelson's results were generally encouraging, showing that a student's structure did become more congruent with an expert's structure. Whereas this program of research has not had an obvious impact on educational assessment, it has provided the impetus for a slowly developing literature investigating the relationship between derived cognitive

structures and domain performance (Champagne, Klopfer, Desena, & Squires, 1981; Diekhoff, 1983; McKeithen, Reitman, Reuter, & Hirtle, 1981; Naveh-Benjamin, McKeachie, Lin, & Tucker, 1986; Thro, 1978). It is not our aim to provide a comprehensive review of this literature, however, some sense of the direction taken in the more recent work can be gained by describing two of these studies.

Champagne et al. (1981) assessed students' cognitive structures of physical geology with a procedure they called the *Concept Structure Analysis Technique*. This technique involved having the students arrange a set of core concepts spatially on a large piece of paper. While the students arranged the concepts, the researcher, guided by the student, labeled subgroupings of concepts. In a related study, Naveh-Benjamin et al. (1986) used a modification of Reitman and Reuter's (1980) ordered tree technique to assess changes in student's cognitive structure of the material presented in a course on the psychology of aging. The results from both of these studies were generally positive in that they found structures to become more similar to experts as a consequence of training, and the structure of students who earned the higher grades were more similar to the experts' structure.

In summary, a number of studies have been conducted showing that individual differences in levels of domain performance are related to differences in derived cognitive structures. However, all of these studies suffered from one or more of the following problems. First, the data on which the cognitive structures were derived were averaged across subjects and therefore failed to assess individual students. If this approach is eventually to be used in the classroom it must be applicable to individual students. Second, the assessment procedures often required the students to report directly the organization of their structure. There are reasonable concerns (e.g., Nisbett & Wilson, 1977) as to whether we have direct conscious access to cognitive structure. Third, the derived structures were often assumed to be hierarchical. Hierarchical structures may be appropriate for some domains, but not all. It would be preferable not to constrain the solution to a hierarchical representation. Finally, the basis for comparing the similarity of structures was often subjective. Ideally we would have an objective quantitative means of assessing the similarity of representations. In the next section we define more precisely the nature of the knowledge representations that we employ and then go on to describe the methods to derive and use these representations to assess classroom learning.

Empirically Derived Knowledge Representations

In his paper on the fundamental aspects of cognitive representations, Palmer (1978) noted that the field is "obtuse, poorly defined, and embarrassingly disorganized" (p. 259). Although more than a decade has passed since Palmer made these observations, there is ample evidence to suggest that his observations remain valid. Heeding Palmer's criticism we attempt, in this section, to describe how we conceptualize the representations we use in the present work.

We begin, as Palmer (1978) did, by distinguishing between what we may think of as an individual's actual knowledge of some domain and some inferred representational model of this knowledge. We assume that the actual knowledge comprises a set of relevant data structures and processes that we shall refer to as the cognitive system. As an individual becomes more expert in a domain the cognitive system is assumed to be modified in some manner. Although the precise changes in the data structures and/or processes may be indeterminate (Anderson, 1978), they are assumed to result in certain changes in domain-relevant performance. Of the many behavioral changes that may occur with the acquisition of

expertise, we are particularly interested in judgments of relatedness among central concepts within the relevant domain. This set of relatedness judgments are, in this sense, a reflection or a representation of the state of the cognitive system. Although there are a variety of transformations, such as multidimensional scaling (MDS), that can be performed on these proximity data, these different transformations are simply alternative representations of the state of the cognitive system. Hence, we view both a set of proximity values and any transformation of those values as simply ways of characterizing a functioning cognitive system rather than as representations of the cognitive system's actual data structure.

In our approach, the preference of one transformation over another is determined first by its validity as a predictor of domain performance and second by its representational simplicity relative to the complete data. In the present work we are specifically interested in determining whether Pathfinder (Schvaneveldt, Durso, & Dearholt, 1987, 1989) representations have greater predictive validity than raw proximity data or MDS-derived representations.

A second distinction in our interpretation of knowledge representations concerns the relationship among the elements represented. Earlier we made the argument that an important aspect of knowledge involves the configural relationships among the concepts of the domain. That is, to be knowledgeable of a domain implies that the important concepts are interrelated and organized in some particular configuration or class of configurations. In order to discuss this issue of concept interrelatedness, we will employ network representations; however, our view of configurality also extends to other types of transformations of proximity data.

Our interpretation of the relationships among concepts as revealed in a network representation differs in two important ways from other representational approaches with which the reader may be more familiar. First, in contrast to network models of language comprehension where it was necessary to explicitly label the type of link that connected various nodes (Collins & Quillian, 1969), the link in the present networks are unlabeled. This raises the question of whether there can be any semantics in a network with unlabeled links (Woods, 1975). We assume that to the extent that a derived representation has predictive validity it also contains a degree of semantic relevance. Second, links between specific nodes in the network do not necessarily have any direct causal implications regarding domain performance. Instead, we assume that it is the pattern of links that is meaningful. This assumption is in contrast to rule-based representational approaches, where relations often do have direct performance implications. What then is implied by network representations as we are interpreting them? We hope to show that structural properties of networks reflect general associative information regarding the state of a cognitive system.

Given our assumption regarding the configural character of knowledge, it follows that we would want to assess the configural properties of network representations. In particular, in comparing the knowledge representations of two individuals we would want to assess the degree of their configural similarity. We assume that the configural properties of a network are not directly obtainable but rather must somehow be interpreted. For our purposes this interpretation process is described by a metric that assesses the similarity of two networks. Goldsmith and Davenport (Chapter 5, this volume) report a set-theoretic method for defining structural similarity between graphs, and we employ this method later to assess the configural similarity of two networks.

To summarize, a central thesis of this chapter is the idea that configural properties of representations reflect important characteristics of an individual's cognitive system. We further assume that these configural characteristics can be compared in network representations by employing a method for assessing structural similarity between graphs. Below we

test the hypothesis that representations of domain knowledge have functional utility by attempting to predict individual differences in performance within a domain from the structural similarity of an individual's representation to some idealized referent representation.

Methodological Issues

First, we turn to some general issues that arise with the method that we are proposing to use to assess domain-specific knowledge. We discuss here the choice of a procedure for collecting proximity data on a set of domain concepts, the particular type of transformations performed on these data, and the methods by which different representations are compared.

A variety of techniques for obtaining proximity data have been previously used, ranging from sorting to memory recall tasks. We have chosen to use direct judgments of concept relatedness as the basis for obtaining conceptual representations. Our choice of relatedness ratings is based in part on the past successes of this method, much of which comes from rating the similarities of directly perceived physical objects. There exists extensive literature demonstrating that such direct similarity ratings are useful for studying perception, memory, and learning (e.g., Shepard, 1974). The validity of similarity judgments as they apply to semantic concepts is more indirect, but here too there has been some success especially in discriminating experts from novices. For example, different levels of expertise have been discriminated among fighter pilots (Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker, & DeMaio, 1985) and computer programmers (Cooke, 1983; Cooke & Schvaneveldt, 1988) on the basis of representations that were in turn derived from pairwise similarity ratings. This suggests that the ratings were sensitive to properties of the cognitive system related to domain performance.

In addition, judgments of relatedness or similarity occur naturally in the course of learning and performing in a domain. The journey from novice to expert may be viewed as a continuous sequence of analysis and synthesis, with each successive cycle providing a more differentiated and integrated cognitive system. In this regard the basic processes of generalization and discrimination play a fundamental role in the acquisition of knowledge. Judgments about what is alike and what is different would appear capable of reflecting fundamental properties of the developing cognitive system. Perhaps William James (1890/1981) had something like this in mind when he said, "the sense of sameness is the very keel and backbone of our thinking" (p. 434).

The validity of relatedness judgments ultimately rests on their ability to provide meaningful results, and this issue raises the question of transforming proximity data. Are transformations of proximities more useful for understanding psychological phenomena than the raw proximities themselves? The history of scaling suggests that the answer to this question is yes, and so a major effort of our work in classroom assessment has been to determine which transformations of concept ratings have greatest predictive validity. In particular we compare raw proximity ratings, MDS spatial representations of the ratings, and Pathfinder representations of the ratings.

A third issue that arises is how to compare different representations. To assess the effects of classroom learning on cognitive structure requires that we have some means of objectively comparing a student's conceptual representation with a referent representation that is assumed to reflect a desired organization of the domain's concepts. In our work a student's structural representation of course concepts is compared to the instructor's representation of the same concepts.

In the case of raw proximity values, we simply calculate the Pearson correlation coefficient between corresponding values in the two sets of proximities. For MDS spatial representations, we first calculate the euclidean distance between all pairs of points in the n-dimensional space and then calculate the Pearson correlation coefficient between corresponding distances in different representations. For Pathfinder networks we employ two measures of similarity. The first is similar to the one just described with MDS representations, but instead of euclidean distances in space we use graph-theoretic distances between nodes in the derived networks. The second similarity measure for networks is the technique mentioned previously and described by Goldsmith and Davenport (Chapter 5, this volume). Briefly, this technique employs a set-theoretic method to compare corresponding neighborhood regions of two networks. The method computes for any two networks a single quantitative index of closeness called C. The values of C range from zero to one.

The issues of how to represent concept ratings of relatedness and then how to compare these representations have important implications for our work. As stated previously, we believe that an important property of conceptual representations is their configural nature. If true, then it is the pattern of interrelationships among a set of concepts that should prove useful for differentiating among individuals with differing levels of knowledge. Hence, we want to employ transformations of concept ratings that preserve or uncover configural relationships in the data and then to use methods for comparing these representations that are sensitive to configural information. We believe that C offers such a method for assessing structural (i.e., configural) similarity, and therefore hypothesize that C applied to Pathfinder networks will indeed result in higher validity for predicting levels of knowledge in students than other representation/comparison methods.

Assessing Classroom Learning

We turn next to an empirical study that investigated the feasibility of assessing classroom learning with empirically derived knowledge structures. The basic purpose of the study was to measure student's knowledge structures over the course of learning and to assess whether the degree of agreement between the students' structure and the instructor's structure was indicative of classroom performance as measured by conventional testing techniques. We hypothesized that students whose structures more closely match the instructor's will indeed be more knowledgeable and hence perform better on standard examinations. We further hypothesized that a measure of representational similarity that assesses configural relationships of knowledge elements will be more predictive of performance than one not based on configural information.

Method

Domain. The knowledge domain was a 16-week sophomore/junior-level college course on psychological research techniques with a primary focus on the analysis and design of experiments. Prior to taking the course, each student had completed an introductory course in probability and statistics. An initial set of concepts considered to be central to experimental design was selected by the instructor, and then suggestions from other faculty members who taught courses in statistics and design were obtained resulting in a revised set of 30 concepts. The final set of concepts is provided in the Appendix. Student performance in the course was measured by three exams and two papers totaling 480 points.

Subjects. A total of 40 students participated in the study with 20 students coming from each of two separate courses taught in different semesters by the same instructor. The students were primarily college juniors and seniors.

Procedure. The purpose of the concept rating project was explained to students at the beginning of the semester. They were told they would be rating the relatedness of 435 pairs ($n(n–1)/2$) of concepts and that these ratings would be used to assess their knowledge of the course material. Students were told that they could earn up to 20 extra course points by performing well on the tasks. At the end of the semester extra points were assigned on the basis of the degree of agreement between their structures and the instructor's.

Students were asked to judge the relatedness of each pair of concepts using a 7-point scale where 1 corresponded to less related and 7 to more related. At the beginning of each rating session, students were shown the complete set of concepts and were encouraged to pick out some pairs that were highly related and some they thought were quite unrelated to serve as anchors. They were also told to use the full range of the scale in making their ratings. Because students would be unfamiliar with some of the concepts at the beginning of the semester, they were asked to consider their confidence in their knowledge of the concepts while making relatedness judgments. Specifically, they were told that ratings from the ends of the scale (e.g., 1, 2, 6, and 7) implied that they were more certain of the meaning of those concepts, whereas ratings from the middle part of the scale (e.g., 3, 4, and 5) could reflect both medium relatedness or uncertainty about the meaning of the concepts.

Students were instructed to give quick intuitive judgments of relatedness rather than performing a lengthy and deliberate analysis of the concept pairs. On average, students took about one hour to complete the set of 435 ratings.

Each pair of concepts appeared left-right centered below the rating scale. A bar marker appeared initially at rating 4 for each concept pair. The bar marker could then be moved with the left and right directional keys on the computer keyboard until it was above the desired rating. Pressing the space bar accepted the rating for the current pair and presented the next pair of concepts. The presentation order of the concept pairs was randomized individually for each subject. Additionally, the left-right order of the concepts was randomized for each pair.

Students rated the concepts on approximately the 1st, 8th, and 15th weeks of the semester. Each student performed the task individually and at their own convenience on microcomputers located around campus. The course instructor also rated the concepts to provide a referent structure against which to compare students.

Results

Data from both classes were combined and analyzed together. The concept ratings yielded proximities by subtracting each rating from eight. These proximities were then analyzed by Kruskal's (1964) nonmetric MDS procedure and Pathfinder. In the case of MDS, an elbow criterion test yielded four dimensions as optimal and so all subsequent MDS analyses are based on four dimensions. PFNETs(∞, n–1) were derived on the same datasets.

Once the MDS and Pathfinder representations were obtained, the similarity between each student's representation and the instructor's was determined using the methods described previously. Comparisons were made from each student's data using each of four different knowledge indices: correlations on raw proximities, correlations on MDS distances, correlations on Pathfinder graph-theoretic distances, and Pathfinder networks assessed by C. To simplify reporting of the results, we abbreviate these as PRX, MDS, PFR, and PFC, respectively.

The first question we turn to is how frequently students used the seven values along the rating scale. Table 1 shows the frequency distribution of relatedness ratings for both the

instructor and students. Particularly striking about these data are the stability of the three distributions over the 15 weeks of the semester. Also, the distribution of the students' ratings was quite similar to the instructor's, with the surprising exception that the instructor tended to use extreme ratings (1 and 7) less frequently than students. Both students and the instructor were more inclined to rate concepts as related (5, 6, or 7) than as unrelated (1, 2, or 3). This may be a function of the domain or the particular set of concepts selected.

We turn next to how well students agreed with themselves and with the instructor as a function of time in the semester. Table 2 shows mean agreement measures assessed both within students and between students and instructor on each knowledge index. Keep in mind that the agreement between networks as measured by *C* units is not directly comparable to the correlation coefficients. Notice first that some agreement exists among students and with the instructor even at the beginning of the semester. This is not too surprising because all of the students had taken a previous course in probability and statistics and many of the concepts were already known.

Table 1. Frequency distribution of relatedness ratings for students at the 1st, 8th, and 15th week of the semester and for the course instructor.

	Relatedness Rating						
Dataset	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Week 1	.06	.13	.16	.14	.20	.18	.13
Week 8	.06	.13	.16	.12	.22	.17	.14
Week 15	.08	.15	.16	.11	.19	.16	.15
Instructor	.02	.14	.20	.14	.25	.21	.04

Table 2. Mean agreement of representations[a] within students and between students and instructor at the 1st, 8th, and 15th week of the semester.

	Student-Student			*Student-Instructor*		
Knowledge Index	*Week 1*	*Week 8*	*Week 15*	*Week 1*	*Week 8*	*Week 15*
PRX	.24	.35	.30	.26	.32	.34
MDS	.28	.23	.43	.39	.49	.54
PFR	.19	.23	.25	.24	.29	.32
PFC	.30	.34	.41	.43	.45	.50

[a]PRX - correlation on raw proximities
MDS - correlation on MDS distances
PFR - correlations on Pathfinder graph-theoretic distances
PFC - Pathfinder similarity assessed by C

Since students in the same course learn a common knowledge base across the semester, we would expect the degree of agreement to increase among students over the semester. This trend clearly existed for all of the knowledge indices from the beginning to end of the semester. However, the mid-semester correlations on proximities and MDS distances fall outside of their beginning-to-end ranges. One explanation for this finding is that students in the middle part of the semester may still view the material quite differently as a result of differing past orientations to the domain. Students may also differ both in their rate of assimilation of the material and in their development of strategies for organizing the material. However, by the end of the semester a sufficient number of shared conceptual experiences has occurred to ensure that a fairly homogeneous view of the domain emerges.

More important, perhaps, is the change in agreement across the semester between students and instructor. We assume that the common knowledge base learned by students is, at least to some extent, that of the instructor's and learning occurs when students agree with the instructor, not necessarily when they agree with one another. Here we find a consistent trend of increasing agreement over time for all of the indices. Notice also that the magnitude of increase over time is greater between students and instructor than within students. The change was most dramatic with MDS representations which might lead one to speculate that if MDS better reflects changes in representations across learning, the degree of agreement between a student's and instructor's MDS representations would be a better index of the student's level of knowledge. We turn to this question next.

Agreement as assessed by the various knowledge indices between each student and instructor was computed based on the student's end-of-the-semester ratings. The last set of ratings was analyzed because they should best reflect the student's overall knowledge of the domain. Pearson product-moment correlations were then computed between each knowledge index and the student's earned course points at the end of the semester. Table 3 shows the resulting correlations. The correlations were all significant ($p < .01$).

Table 3. Correlations (and squared correlations) of instructor-student agreement and final course points for students.

Knowledge Index	*Correlation*
PRX	.61 (.37)
MDS	.54 (.29)
PFR	.66 (.43)
PFC	.74 (.55)

The correlation coefficient between a student's and instructor's concept ratings (PRX) accounts for 37% of the variance associated with the student's final course grade. Therefore, concept ratings themselves appear to be an indicator of a student's knowledge. Of more interest, however, is whether scaling algorithms, such as MDS and Pathfinder, are able to extract from these ratings information that would allow even better performance predictions. The answer appears to vary. Distances from MDS were slightly poorer than proximities in predicting performance, whereas Pathfinder distances were better than the raw proximities. Comparison of Pathfinder networks with *C* (PFC) provided even better predictions than with correlations (PFR).

One way of looking closer at the relative contribution of each knowledge index to predicting final course points is to examine partial correlations. Table 4 gives the correlation

between each index and final points with the variance contributed by each other index partialed out. Consider first Pathfinder networks that have been compared using *C*. PFC correlates significantly with final points even when each of the other indices is held constant. However, none of the other indices correlate significantly with course grades when the variance contributed by PFC is held constant. This pattern of findings strongly suggests that Pathfinder networks, as assessed by *C*, are uniquely capturing important predictive variance in the concept ratings. Consider next Pathfinder networks as assessed by correlations. PFR is a significant predictor when proximities and MDS is each held constant, but not PFC.

Therefore, taken together, these results imply that Pathfinder networks do indeed contain unique predictive variance over the proximity ratings and MDS, and that a configural assessment of networks is a better index for assessing network similarity than correlations. Apparently, *C* better reflects commonalties between structures that happen to be important in assessing knowledge. We assume that the characteristics common to a student's structure and instructor's structure that are predictive of knowledge attainment exist at a global or configural level within those representations. This, of course, is exactly the type of information that *C* is assumed to be good at assessing.

Table 4. Partial correlations between Knowledge Index 1 and final course points with Knowledge Index 2 partialed out.

	Knowledge Index 2			
Knowledge Index 1	*PRX*	*MDS*	*PFR*	*PFC*
PRX	-	.34*	.30	.15
MDS	.03	-	.29	.12
PFR	.43**	.52**	-	.17
PFC	.54**	.61**	.46**	-

*$p < .05$
**$p < .01$

Consider next the results from MDS. Spatial structures did not significantly predict course performance when the other variables were partialed out. MDS has been successful in previous application for representing physical or continuous relations. Our results may indicate a specific limitation of MDS for assessing conceptual-level relations. Some corroboration for this conclusion is found in work by Cooke, Durso, and Schvaneveldt (1986) who compared MDS and Pathfinder in predicting recall data.

In summary, we conclude that knowledge representations based on college students' concept ratings do indeed offer a valid assessment of their classroom learning. Students' conceptual representations appear to reflect changes in their learning over the course of instruction and become increasingly similar to the instructor's conceptual representation. Further, the extent to which a student's representation matches that of the instructor at the end of the semester is a good index of how much knowledge the student has learned about the domain of study. As hypothesized, predictive ability depends on how the structure is represented and how structural similarity is assessed. Pathfinder networks assessed by *C* appear to offer the best indication of student performance.

General Discussion and Conclusions

The primary aim of this undertaking was to investigate the possibility of using knowledge representations derived from relatedness ratings of domain concepts as a means of assessing classroom knowledge. The primary assumption guiding the work was that expertise in abstract domains, such as statistics and experimental design, requires an understanding of the interrelationships among the concepts, which we have referred to as the configural property of knowledge. The results that bear most directly on this thesis were the partial correlations which showed that all significant variance in course performance was captured by PFC. When PFC was partialed out, none of the other predictor variables accounted for a significant proportion of variance, but with these other predictors partialed out, PFC continued to significantly account for the variance in course performance. Of particular importance is the finding that PFC accounted for variance in classroom performance that was not captured by PFR. This points up the importance of the metric by which network similarity is measured. The superiority of PFC over PFR is seen as support for the idea that configural properties of representations do indeed capture important aspects of a cognitive system.

The success of the *C* measure is, of course, dependent on the validity of the relatedness ratings and the Pathfinder representation of these relations. The significant relationship between raw proximity data and classroom performance corroborates previous results indicating that relatedness ratings are a valid measure of domain knowledge (e.g., Diekhoff, 1983; Schvaneveldt, Durso, Goldsmith, et al., 1985; Thro, 1978). The superior predictive power of the PFC measure over the raw proximity data and the MDS representations points up the value of the Pathfinder-derived network representations. Previous studies (Schvaneveldt, Durso, Goldsmith, et al., 1985; Stephens, 1987) have also found network representations to be better predictors of domain performance than MDS spatial representations. As suggested by Reitman and Reuter (1980) it may be the case that network structures are superior to spatial structures in representing conceptual domains.

Comparison to Expert-Novice Research

The cognitive structural approach to classroom assessment, as exemplified by the work beginning with Shavelson, has an obvious connection with much of the expert-novice research. Both approaches take a structural representational view of domain-specific knowledge. There is, however, an interesting difference between these two areas of research. The expert-novice work, as this descriptor denotes, usually compares groups that differ widely in skill levels. Perhaps as a consequence of the extreme differences between experts and novices in training and experience, the inferred cognitive differences have often been discussed in qualitative terms. For example, computer programmers have been shown to organize programming concepts either semantically or syntactically depending on skill level (Adelson, 1981).

Our results show that it is possible to discriminate among students within a level of expertise and describe these finer-grain differences of performance along a quantitative continuum (i.e., similarity to the instructor's representation). Although this is not to be interpreted as suggesting that qualitative distinctions are unnecessary, it does introduce the possibility that at some levels of analysis, expert-novice distinctions can be seen as a continuous transition. More important, the present findings suggest that relatively small differences in expertise can be discriminated with a cognitive representational approach.

Application of a Structural Assessment Approach

Our approach to assessment is similar in many ways to that of traditional psychometric tests. In both cases the representation is based on a sample of performance that serves as an index of knowledge. In the case of an achievement test, the performance sample is often a direct assessment of domain knowledge (e.g., do you know this fact?), whereas in the present case the performance is a sample of a person's judgments of the general associations that exist among some subset of the domain's relevant concepts.

Assuming that future work continues to support the validity of the present approach, it may be appropriate to consider some advantages of the structural assessment approach as an assessment tool. As noted earlier the present technique avoids a number of problems that are often associated with standard examination procedures. The structural assessment procedure is less dependent on: (a) recall of episodic information; (b) idiosyncratic misinterpretations of questions; and (c) students' ability to articulate the relevant knowledge. The technique also has the potential of being completely automated. This would allow for large-scale application with rapid, objective scoring. Automation would also allow students to perform self-evaluations at any juncture in training. Finally, the approach has a wide range of potential application in that it can be applied to any domain that involves understanding the interrelations of some specifiable set of concepts.

Future Directions

Over the time of conducting and reporting this research we have had numerous thoughts on improving and extending the approach used in the present work. In closing we shall mention briefly what we believe may be the more important of these ideas and also respond to some common criticisms of our work.

One of the most frequently voiced criticisms of this approach relates to the use of the instructor as the standard of comparison. In defense, one can contend that this only makes explicit what happens implicitly in the design and scoring of exams. It could also be argued that the instructor is as good a model as any for relatively novice students taking a sophomore-level class. However, a more positive reply to this criticism is that it is not necessary to use the instructor. A number of possibilities can be explored, ranging from other individual experts, top students, or even families of desired representations. In an analysis not reported here, we found that PFC comparisons among the students themselves allowed us to discriminate good from poor students. More specifically, when we selected a subgroup of students having the highest between-student PFC scores, these students turned out also to be the better students based on final course points.

A second matter concerns the reliability of our assessments. People performing the ratings task often comment on how uncertain and subjective these ratings appear to them. We have looked at the correlations between repeated ratings of the same concept pairs and found them to average around .60. This may appear somewhat low for a reliability coefficient relative to what is usually reported as test reliability. However, there are several important differences between standard test reliability and reliability of relatedness ratings. First, in the present case the correlation of .60 reflects item stability for an individual's ratings, whereas test reliability reflects subject stability within a sample of subjects. In this regard these correlations are not directly comparable measures of reliability. Second, the actual mean difference in ratings across repeated items tends to be quite small (e.g., 1.03 for subjects using a 7-point rating scale) and only 9% of the absolute differences were greater than two ratings apart. Therefore, it appears that students are more consistent in their ratings than they may think they are. Finally, if PFC is used to predict classroom

performance, then the reliability in question is of the final predictions, and this can be directly assessed only by determining the stability of Pathfinder representations as measured by *C*. In this regard, rating consistency is only indirectly related to the reliability of the predictor variable.

Another issue is the selection of the set of concepts to be rated. How many concepts should be sampled and on what basis should they be selected? In the work reported here, only 30 concepts were used for the simple reason that 435 ratings seemed to be close to the upper limits of time and fatigue for a single hour session of ratings. The basis for selecting concepts was that they were important and representative of the material covered in the course. We are currently exploring the effects of sample size on predictive validity. Not surprisingly, we find that predictive validity is much more variable with smaller samples. However, on the basis of preliminary results it appears possible to attain relatively high predictive validity with as few as 10 concepts. We are investigating whether there might be some principled basis for selecting concepts that maximize predictive validity.

Finally, we have speculated about using empirically derived knowledge representations as a guide to teaching. If a domain of knowledge, as a field of study, has a structure that is more or less shared implicitly or explicitly by experts in the field, then it should be the goal of instruction and training to communicate this structure as effectively and efficiently as possible. Is it possible that an individual's personal representation of concepts may help identify particular deficiencies in her knowledge of the domain? Perhaps so, but this type of intervention would seem to require an analysis of individual concepts in a representation. If the important property of concept representations really is of a configural nature, then such an analysis may in fact not be meaningful.

Appendix

Set of 30 Core Concepts from Experimental Design

analysis	main effect
assignment	matching
between subjects	model
block	order effect
confound	orthogonal
control	random
counterbalance	replication
covariance	significance
data	subject
design	theory
distribution	treatment
error	validity
experiment	variable
hypothesis	variance
interaction	within subjects