Scoring Concept Maps: An Expert Map-Based Scheme Weighted for Relationships

James A. Rye
West Virginia University

Peter A. Rubba
Penn State University

The use of student-constructed concept maps in assessment is congruent with the changing emphases set forth by the National Science Education Standards. *Authorities have expressed concern about concept map scoring systems and their associated validity and reliability. They favor methods that employ expert/criterion maps as referents and emphasize the use of accurate concept relationships in deriving scores, which have been found to correlate with performance on standardized tests. In this study, student constructed concept maps (n = 17) that emerged from post-instructional interviews about chlorofluorocarbons were scored against a teacher-expert map using a scheme weighted for relationships. Interrater reliability for the scoring scheme was high (r = .959). Students' map scores correlated highly with their scores on the California Achievement Test component total (r = .729) and moderately with their Pathfinder index (r = .474), the latter believed to be an excellent measure of structural knowledge. A revised map score, derived only from relationships containing one or more of the concepts employed in Pathfinder analysis, was a statistically significant (p = .031) predictor of the Pathfinder index. The findings of this study support the recommendations of others to use expert referents and emphasize concept relationships in assessing concept maps.*

The concept map is a versatile science education tool, with utility for investigating students' understandings and planning, providing, and assessing instruction (Good, Novak, & Wandersee, 1990; Horton et al., 1993; Jonassen, 2000; Novak & Gowin, 1984; Roth, 1994; Wandersee, Mintzes, & Novak, 1994). Innovations in instructional technology have enhanced the efficiency of constructing and presenting concept maps and, accordingly, the possibilities for "concept mapping" in educational settings (Inman-Anderson & Ditson, 1999). As an assessment tool, student-constructed (as opposed to fill-in-the-blank) concept maps provide both qualitative and quantitative measures of understanding and have considerable potential for revealing changes in knowledge over time (Edmonson, 2000; Kinchin, Hay, & Adams, 2000). The use of concept maps in assessment is congruent with the *National Science Education Standards'* changing emphases, which call for engaging students in ongoing self-assessment and "assessing to learn what students *do* understand" (National Research Council, 1996, p. 100). Luft (1998) recommended, "To move our current assessment beyond a competency measure, teachers should use concept maps..." (p. 43).

Recent reviews and investigations of the concept map as an assessment tool have expressed concern about concept map scoring systems and associated validity and reliability (McClure, Sonak, & Suen, 1999; Rice, Ryan, & Sampson, 1998; Ruiz-Primo & Shavelson, 1996; Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Xiufeng & Hinchey, 1996). These articles discuss the use of expert or criterion maps as referents and the importance of concept relationships (propositions) in deriving concept map scores. This paper will review some of these concerns and subsequently report on a study of a concept map scoring method that was weighted for propositions present in a referent "teacher-expert" map. The findings have implications for science education researchers, as well as for classroom science teachers who wish to use concept maps to assess student learning.

Scoring Concept Maps

Ruiz-Primo and Shavelson (1996) and Jonassen (2000) described the use of expert maps produced by the teacher, instructor, or another expert as one standard for scoring concept maps. As students develop

domain expertise, their knowledge structures become more like domain experts or their instructors (Goldsmith & Davenport, 1990; Ruiz-Primo & Shavelson, 1996). Jonassen contended, "Learners do come to think like teachers" (p. 36); he cited research showing that the degree of similarity between the learners' and instructor's knowledge structures is a good predictor of exam performance. Rye and Rubba (1998) found that the degree of similarity between eighth-grade science students' and their teacher's knowledge structures about chlorofluorocarbons, as determined through Pathfinder analysis (Schvaneveldt, 1990), was a strong predictor ($p = .003$) of students' performance in concept interviews.

Of 21 research studies reviewed by Ruiz-Primo and Shavelson (1996) that used the concept map to assess learning, the majority was found to employ an expert or criterion map, and several examined learner maps for the presence of accurate or valid propositions. Ruiz-Primo and Shavelson concluded, "We favor scoring criteria that focus more on the adequacy of the propositions over those that focus simply on counting the number of map components (i.e., nodes and links)" (p. 595).

A chief concern expressed by Ruiz-Primo and Shavelson (1996) was that few of the studies employing the concept map in assessment reported reliability data. Interrater reliability, also known as interscorer agreement (Moore, 1983), can be expressed as a correlation coefficient or as a percentage of agreement (McMillan, 1992). However, a high correlation (reliability coefficient) between the scores produced by raters is not synonymous with agreement: Scores assigned by one rater may be significantly higher or lower than scores assigned by another (Borg & Gall, 1971), yet both sets of scores may be strongly correlated. Semantics complicate the issue of reporting reliability, in that the terms interrater reliability and interrater agreement may be used together without further qualification.

Ruiz-Primo and Shavelson (1996) had additional concerns about studies that did report reliability data. One study estimated that interrater agreement was greater than 80%, but did not disclose the procedures to establish reliability or the number of raters. Other concept map studies reporting high interrater reliabilities (at least .8 or 80%) amongst raters did so based on only a few maps or through scores that did not consider the *accuracy* of relationships in the map, or eliminated a rater prior to reporting the reliabilities. Novak and Musonda (1990), in referring to previous work, reported achieving interrater correlations of approximately .95 from scoring concept maps produced by seventh- and eighth-grade students. From their own work, Ruiz-Primo and Shavelson (2000) reported achieving high interrater reliability coefficients (.91 to .98) for "convergence scores," as well as a "proposition-accuracy score," and concluded that concept maps can be scored reliably "even when complex judgments of propositional quality are required" (p. 326).

In their study of concept maps produced by college students in an educational psychology course, McClure et al. (1999) employed holistic, structural (adapted from Novak & Gowin, 1984), and relational scoring methods with and without a master (expert) map. The "relational with master map" method was utilized to examine students' concept maps only for the accuracy of "master map" propositions; among all six methods, it achieved the highest score reliability, which was expressed as a g-coefficient. Because of the high reliability, and because scoring propositions do not require a complicated procedure, they recommended a relational scoring method with some form of a master map for classroom science teachers who wish to use concept maps to assess student learning. Additionally, McClure et al. found that of six scoring procedures employed, the "relational with master map" procedure correlated the highest ($r = .608$) to a "neighborhood" similarity index of each student's map and the master map. The concept of "neighborhood" also is employed in Pathfinder analysis, which compares Pathfinder networks derived from student and expert ratings of concept pairs to produce a Pathfinder "similarity" index (Acton, Johnson, & Goldsmith, 1994; Jonassen, 1993; Schvaneveldt, 1990). Jonassen (2000) reported that this index is an excellent measure of students' structural knowledge.

Rice et al. (1998) developed a scoring method that linked concept map scores to achievement of instructional objectives for seventh-grade science students. Their method, for which they reported a high interrater reliability (98% agreement), emphasized the correctness of concept relationships that comprised instruction. It was "reminiscent of scoring maps relative to an expert teacher's map" (p. 1124). Rice et al. also found that students' concept map scores, when derived solely from the presence or absence of correct propositions, correlated highly ($r = .82$ to $.87$) with state and national (SAT) standardized tests and, accordingly, indicated that the concept map scores had high concurrent validity. Rice et al. and Ruiz-Primo and Shavelson (2000) cited Anderson and Huang (1989), who reported substantial correlations (.49 - .74) between concept map scores based on accurate propositions and standardized test scores (e.g., Stanford Science Achievement Test).

Xiufeng and Hinchey (1996) scored concept maps on solutions, force, and interaction produced by junior high students after they completed respective units of study. They employed Novak and Gowin's scheme (Novak & Gowin, 1984), which awards points for valid hierarchies and cross-links, as well as propositions and examples. Post-instruction, students also completed conventional tests on these units. The authors reported that for two of the three units—force and interaction—the correlations between the concept map scores and conventional test scores were not significant, and accordingly, the map scores had low predictive validity for students' performance on these tests. Further, the authors reported that "Novak and Gowin's scoring scheme is not internally consistent" (p. 933) and suggested the development of a new scoring scheme. Kinchin et al. (2000) also expressed concerns about concept map scores derived by aggregating elements such as valid links, degree of cross-linking, and hierarchical structure, contending that such "creates a blurring of what the overall score actually reveals" (p. 46).

Neither Xiufeng and Hinchey (1996) or Ruiz-Primo and Shavelson (1996) made reference to the scoring method that Novak and colleagues employed in their longitudinal study of science concept learning (Novak & Musonda, 1991). Concepts and propositions were the only concept map components explicitly "counted" in the scoring, and points were deducted for the presence of misconceptions. Their algorithm was grounded in Ausubel's assimilation theory (Novak, 1992): Twice as many points were assigned to valid propositions than to relevant concepts, and these values decreased according to the level at which the concepts and propositions appeared in the concept map hierarchy. For example, first and second level propositions were awarded 20 and 10 points, respectively, whereas first and second level concepts received 10 and 5 points. The components and relative weights used in Novak and Musonda's scoring system informed the rubric employed in the study reported in this article. Further, the emphasis on concept relationships in this rubric and use of a "teacher-expert" referent in the scoring scheme finds support in the research by Rice et al. (1998) and McClure et al. (1999).

Study

Background

The research reported in this paper employed a rubric to score concept maps (Novak & Gowin, 1984) that emerged from post-instructional concept interviews (as described by authorities, e.g., White & Gunstone, 1992). Other measures of students' understandings were either regressed on or correlated to these scores. This research was a subset of a study (Rye, 1995) that examined eighth-grade physical science students' understandings about chlorofluorocarbons (CFCs) and their role in global atmospheric change (GAC). The instruction provided to the students was from a GAC unit developed through a Science-Technology-Society (STS) Leadership Institute for teachers (Rubba et al., 1995; Rubba, Wiesenmayer, Rye, & Ditty, 1996). The instruction covered the greenhouse effect, the enhanced greenhouse effect (global warming), and ozone layer depletion. CFCs are involved in global warming and ozone depletion: a robust understanding of CFCs includes both phenomena (Rye & Rubba, 2000).

As described by Rye and Rubba (1998), a chief interest of the study was to determine if embedding a concept mapping process in a post-instructional concept interview would enhance the elicitation of students' conceptual understandings during the interview. Thirty-six students (all volunteers) were randomly assigned to either a "standardized interview" group (as described by Patton, 1990) or a "standardized interview with concept mapping" group. Thus, all interviews yielded an interview transcript, and those embedding the mapping process also yielded a concept map. The interview questions posed to all students concerned (a) the nature and source of CFCs and (b) what, if any, problems were caused by CFCs. Considerable focus was on the latter: For each problem cited, students were queried as to how CFCs caused that problem and why it was a problem. Students in the "interview plus concept mapping" group were asked to map what they believed to be important from their verbal response(s) to an interview question, and to examine the emerging map while formulating verbal response(s) to subsequent interview questions.

Seventeen students in each group completed the interviews. Concepts and propositions (concept relationships) elicited from the students during the interview were analyzed against those in templates derived from a teacher-expert concept map to determine the degree to which the interviews elicited the understandings set forth in that teacher-expert map. The teacher-expert map was constructed through a knowledge engineering process (as in Gordon, Schmierer, & Gill, 1993) with the teacher of the students, and captured what the teacher believed to be an ideal understanding relative to the instruction she provided on CFCs and their role in GAC. The derivation of the teacher expert map is described in more detail and illustrated by Rye and Rubba (1998).

This expert map had nine levels in the concept hierarchy: Figure 1 provides the first, second, and a portion of the third level of that hierarchy.

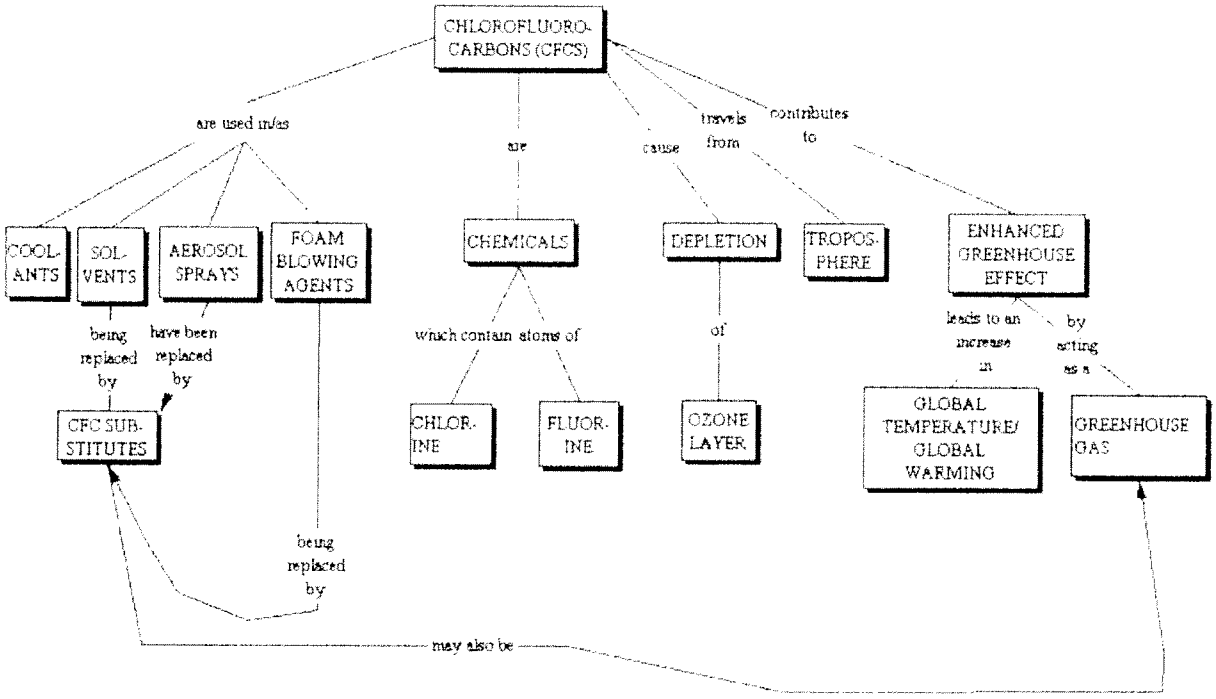
The study also included Pathfinder analysis (Schvaneveldt, 1990) as another measure of students' understandings about GAC and to determine if the Pathfinder "similarity index" was a predictor of students' understandings, as externalized during the concept interviews. A subset of eight concepts—CFCs, greenhouse gas, global warming, ozone layer, depletion, man-made, UV light, and coolants—deemed most important in the expert map were identified by the teacher, and students completed a concept pair relatedness rating activity (see Acton et al., 1994) of all possible combinations (28 pairs) of these eight concepts. The ratings were used for generating Pathfinder networks (net) via the Knowledge Network Organizing Tool (KNOT, Version 1.0, Interlink, Inc., Las Cruces, NM), and KNOT compared each student's net to the teacher-expert net to determine each student's Pathfinder index (0 = *no similarity* to 1 = *identical to teacher's net*). Rye and Rubba (1998) provided the teacher's and examples of students' Pathfinder networks and respective discussion on the comparison of teacher-to-student nets.

Rye and Rubba (1998) also detailed the research design, interview questions and process for embedding concept mapping in one-half of the interviews, and analysis of the interview transcripts. They reported findings relative to these questions: (a) Does embedding a concept mapping process in the interview increase the elicitation from students of teacher-expert concepts and relationships during that interview? (*ns*); and (b) Is the Pathfinder index a predictor of the (simple sum of) teacher-expert concepts and relationships elicited from students during the interview? ($p = .003$). The research reported subsequently examined students' scores on the concept maps emergent from the interviews that embedded a mapping process.

Research Questions

Two research questions targeted the concept map scores, which were derived from a concept map scoring rubric developed by the researchers: (a) Are students' concept map scores a statistically significant ($p < .05$) predictor of students' Pathfinder indexes? and

Figure 1. *Portion of teacher-expert concept map on CFCs.*



Note: From "An exploration of the concept map as an interview tool to facilitate the externalization of students' understandings about global atmospheric change," by J. Rye & P. Rubba, 1998, *The Journal of Research in Science Teaching*, 35, p. 521-546. Copyright 1998 by John Wiley & Sons, Inc. Reprinted with permission.

(b) Are students' concept map scores a statistically significant ($p < .05$) predictor of their performance during the interview (where interview performance is measured as the sum—1 point each—of teacher-expert concepts and relationships elicited from students and captured by the interview transcripts and respective concept maps)? Standardized achievement test scores also were obtained for students in this study, and the researchers became interested in the degree to which students' concept map and achievement test scores would correlate.

Subjects and Data Sources

The subjects in this study were the 17 eighth-grade physical science students who had been randomly assigned to the group that would complete the concept interviews that embedded a concept mapping process. Prior to these post-instructional interviews, the students received approximately 5 weeks of instruction from a GAC unit. The teacher of these students, who engaged in developing the teacher-expert map and Pathfinder network referents for data analysis, was a participant in the STS Leadership Institute that developed the GAC unit (Rubba et al., 1996). The principal data sources for this study were the students' concept maps emergent from the interviews. Other data sources from students were the interview transcripts, the concept pair similarity ratings of all possible combinations of the eight central GAC concepts, and verbal and component total scores from the most recent take of the California Achievement Test (CTB-McGraw-Hill, 1986).

Concept Map Rubric and Scores

The concept map scoring rubric was designed prior to any data analysis, in order to minimize researcher bias. The rubric considered two components of each student's map: concepts and relationships (propositions). The relative weights assigned to concepts and relationships were adapted from scoring procedures instituted by others (Arnaudin, Mintzes, Dunn, & Shafer, 1984; Malone & Dekkers, 1984; Novak & Gowin, 1984; Novak & Musonda, 1991; Shymansky et al., 1993; Stensvold & Wilson, 1990; Wallace & Mintzes, 1990). According to Novak and Gowin, any scoring key, as well as the numerical scores employed to score concept maps, are somewhat arbitrary and subjective: "We would encourage educators to experiment with different values" (p. 105). In this study, relationships were assigned a greater weight than concepts in the concept maps, because the explication of relationships was considered to be the more robust indicator of understanding (Goldsmith, Johnson, & Acton, 1991; Stensvold & Wilson, 1990). Cross-links were not weighted more heavily than relationships. Given that cross-links are indicative of integrative reconciliation, they are often (Heinze-Fry, 1987; Novak & Gowin, 1984; Wallace & Mintzes, 1990), but are not always (Arnaudin et al., 1984; Novak & Musonda, 1991) assigned a greater value than relationships. The number of instances of branching and levels within the concept hierarchy in the maps were not scored: Novak and Musonda reported difficulty in making determinations of the latter.

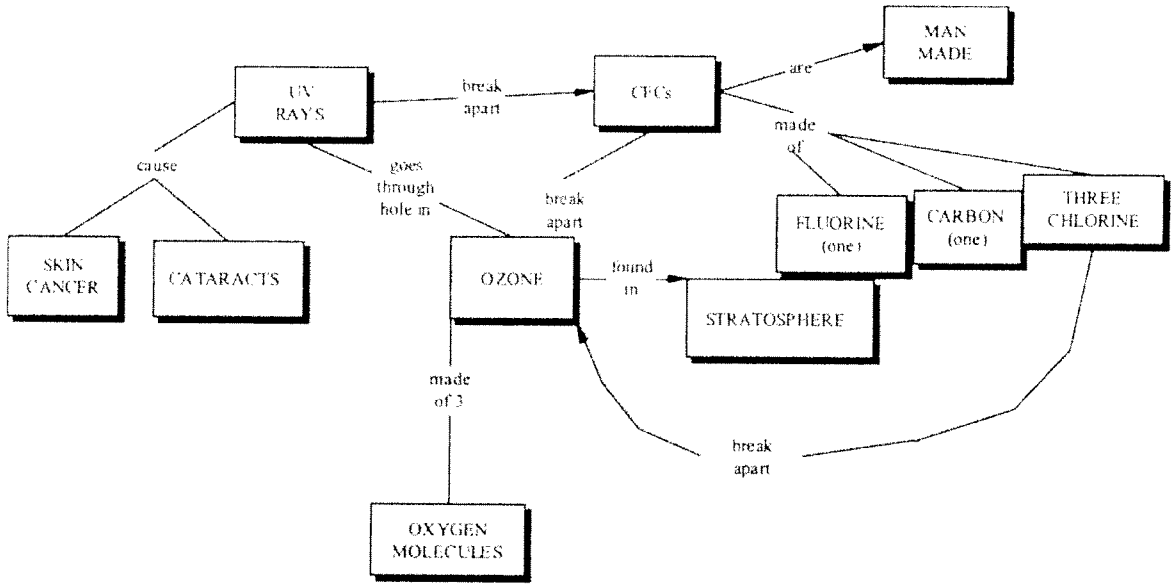
The scoring scheme employed for awarding points to concepts and relationships in student maps is described in a following section. The sum of the points awarded became the student's concept map score (CMAPSCR). CMAPSCR was derived from scoring completed by the first author. Figure 2 is an example of one student's concept map that emerged from the interviews: The derivation of this student's CMAPSCR (93 points—25 points for concepts plus 68 points for relationships) is explained as a part of the scoring scheme that follows.

Concepts. Each of the lessons that comprised the GAC instruction in this study set forth at the beginning a concept map; there were a total of about 140 different concepts in these maps. The maps were to help teachers "key in" on the lesson concepts and relationships in planning their instruction. A subset of 41 of these concepts was included by the teacher in the "teacher-expert map" on CFCs. Of the 41 concepts, the teacher deemed the following eight as "central" to an understanding of CFCs: CFCs, coolants, depletion, global warming, greenhouse gas, man-made, ozone layer, and UV light.

The concept map rubric was set up to allow for the "counting" of most of the 140 concepts: The actual number of concepts was reduced to 127 by eliminating general concepts (e.g., "amount") and those that were highly specific to laboratory procedures (e.g., "acid-base indicator"). Concepts in the map were assigned values of 3, 2, or 1 point(s). Semantic equivalency was considered in awarding points for any concepts, e.g., "man-made" was considered the equivalent of "human-made."

- *Three points* were awarded for the presence of each of the eight *central* teacher-expert concepts, with the exception of "CFCs." No points were awarded for CFCs because this concept was provided to the student as the interview topic and as a seed concept for initiating the map. The map illustrated by Figure 2 earned 12 points for these four concepts: "man-made," "break-apart" (taken

Figure 2. *Sample student concept map (CMAPSCR = 93 points) that emerged from interview.*



as semantic equivalent of "depletion"), ozone, and UV rays.

- *Two points* were awarded for each of the remaining 33 teacher-expert concepts. Figure 2 earned 12 points for these six concepts: carbon, cataracts, chlorine, fluorine, skin cancer, and stratosphere.
- *One point* was awarded for any of the remaining 87 concepts. These remaining concepts were referred to as "external" concepts to designate that they were not part of the teacher-expert map on CFCs, yet they were related to GAC and students may "bring them in" when mapping what they knew about CFCs. Figure 2 earned 1 point for this concept: oxygen molecule.

Relationships. The rubric was set up to award more points to relationships than to concepts. Relationships provide detail about the meanings students' give to the concepts and the interrelatedness of students' understandings. Relationships in the map were each worth 6, 4, or 2 points, depending on the type.

- *Six points* awarded for teacher-expert relationships. Forty-five different relationships were extracted from the teacher expert map. Scoring for the presence of each relationship allowed for the counting of semantic equivalents (e.g., "CFCs are used in coolants" is the equivalent of "CFCs are a refrigerant") and asymmetric equivalents (e.g., "ozone layer is in stratosphere" is the asymmetric equivalent of "stratosphere contains the ozone layer"). Figure 2 earned 60 points for 10 relationships, e.g., "CFCs are man-made," "UV rays go through hole in ozone," and "UV rays cause skin cancer."
- *Four points* were awarded for other valid and relevant relationships between two teacher-expert concepts *or* between one teacher-expert concept and one "external" concept. (The rubric as set forth prior to the scoring awarded only 2 points to "other" relationships between two teacher-expert concepts. During the scoring, this was found to be inconsistent with the importance given to teacher-expert concepts elsewhere in the rubric, and accordingly, a decision was made to award 4 points to these other relationships between two teacher-expert concepts.) Figure 2 earned 8 points for these two relationships: "UV rays break apart CFCs" and "ozone made of 3 oxygen molecules." The researchers accepted the latter as valid, recognizing that oxygen "atoms" is the accurate term in this relationship.
- *Two points* were awarded for valid and relevant relationships between two "external" concepts. Figure 2 earned no points.

Accordingly, the scoring of relationships in the maps was based on a student's extent of understanding (as recommended in White, 1985; White & Gunstone, 1992) of the relationships amongst a pool of concepts

that comprised the teacher-expert map and lessons from the GAC curriculum.

Interrater Reliability

The second author scored the maps independently from the first author. An interrater reliability (r) of .959 was realized by correlating the resulting CMAPSCR obtained by each author for all 17 maps. The high interrater reliability does not mean that the raters obtained almost the same score for each map. The mean (M) of all maps scored by the first author was approximately 67 (see Findings and Discussion) compared to $M = 61$ for the second rater (e.g., the second rater scored the map illustrated in Figure 2 as 89 points).

Data Analysis

Statistical analysis of the data employed STATVIEW (Version 5.0, SAS Institute, Cary, NC). Bivariate regressions were performed to determine the predictive validity of CMAPSCR for the following: (a) students' Pathfinder index, and (b) students' performance in the interview relative to the content of the teacher-expert concept map. Each student's performance in the interview was measured as the sum of teacher-expert concepts (1 point each) and relationships (1 point each) elicited during the interview: Those captured by the interview transcript and any *additional* teacher-expert concepts and relationships that the student placed in the concept map. (For more information on the measurement of interview performance, see "ACCORD Templates" in Rye & Rubba, 1998, p. 532.) A correlation matrix was generated to determine the degree to which CMAPSCR was related to students' component total and verbal scores on the California Achievement Test (CAT).

Normality and Linearity of the Data

This study applied the statistical analyses of correlation and regression to a small sample size ($n = 17$). Therefore, for each of the continuous variables (CMAPSCR, Pathfinder index, California Achievement Test [CAT] component total and verbal scores, and interview performance score), the Kolmogorow-Smirnov (K-S) Normality Test was employed to test the assumption that the actual values and ideal normal values (as computed by the K-S Normality Test Template in STATVIEW) are from the same distribution. For each continuous variable, the K-S Normality Test yielded $p > .05$. Accordingly, we assume that our data is normally distributed. Additionally, we have included two bivariate scatterplots of select data to illustrate linearity: Figure 3 plots CMAPSCR against Pathfinder index; Figure 4 plots CMAPSCR against CAT component total.

Figure 3. *Bivariate scatterplot of continuous variables: CMAPSCR vs. Pathfinder index.*

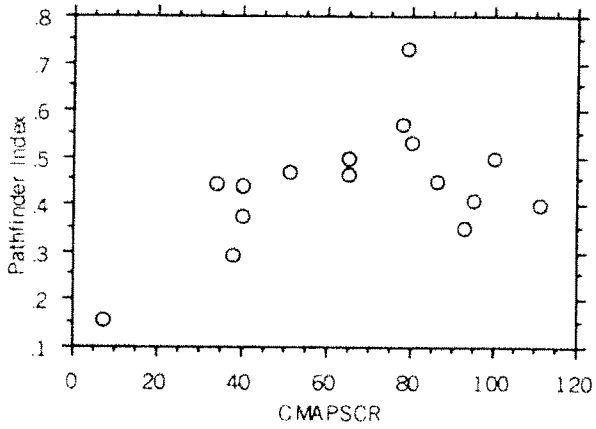
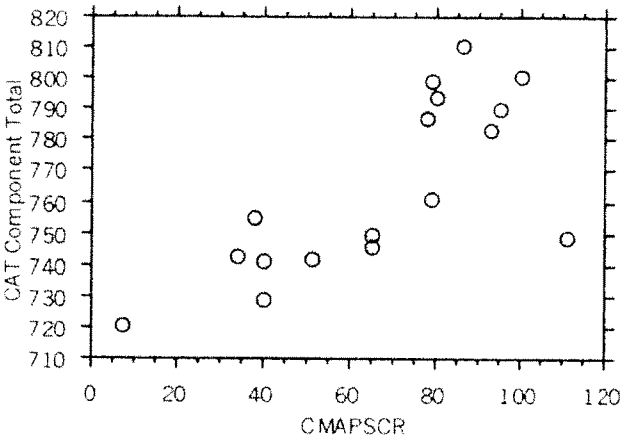


Figure 4. *Bivariate scatterplot of continuous variables: CMAPSCR vs. CAT component total.*



Findings and Discussion

For all students' ($n = 17$), the mean CMAPSCR was 67.118, with a range of 7 to 111 and a standard deviation (SD) of 28.147. Rye (1998) provided a frequency analysis of the degree to which the teacher-expert concepts and relationships were present in the students' concept maps. Rye and Rubba (1998) provided descriptive statistics for students' interview performance, Pathfinder index, and CAT scores.

CMAPSCR Predicting Pathfinder Index

The Pathfinder index was available for only 16 of the 17 students who developed concept maps. Regressing students' Pathfinder index on CMAPSCR yielded a parameter estimate (b) of .002, a probability (p) of .0638, and a proportional reduction in error (R^2) of .224. Thus, CMAPSCR was not a reliable ($p < .05$) predictor of Pathfinder index. Although statistical significance at the .05 level was not achieved, CMAPSCR did have a moderate correlation with ($r = .474$; McMillan, 1992) and reduced appreciably the error ($R^2 = .224$) in predicting Pathfinder index. This finding illustrates the importance of reporting R^2 along with the statistical inference for regression analysis, especially when a small sample size—such as in this study—limits the power of that analysis (see Judd & McClelland, 1989).

The examination of students' concept map scores as a predictor of their Pathfinder index can be considered exploratory research. Gall, Borg, and Gall (1996) reported that exploratory research may set $p < .10$, because "it might spotlight a potentially important difference, relationship, or effect that would have been overlooked had a lower p value been used" (p. 187). Thus, $p = .0638$ led the researchers to investigate a subset of CMAPSCR that was based more closely on those eight concepts used for Pathfinder analysis.

The research described previously by McClure et al. (1999) on "relationship" concept map scores and "neighborhood similarity index" suggested that a score based solely on relationships might be a better predictor of the Pathfinder index. Accordingly, the CMAPSCR subset was based solely on the presence of *teacher-expert* relationships (the score was labeled CMAPTER), and derived by counting only teacher-expert relationships in the map that included at least one of the eight central teacher-expert concepts. As in CMAPSCR, 6 points were awarded to each of the relationships that comprised CMAPTER. Descriptive statistics for CMAPTER were $M = 36.706$ (range of 0 to 66) and $SD = 18.721$. Regressing Pathfinder index on CMAPTER revealed that the latter was a reliable predictor ($b = .004$; $p = .031$; $R^2 = .290$) of and had a moderate correlation ($r = .539$) to the Pathfinder index. This correlation approached the correlation ($r = .608$) found by McClure et al. between relationship concept map scores and neighborhood similarity index and supports placing an emphasis on (accurate) concept relationships in scoring concept maps, as recommended by McClure et al. and others (e.g., Rice et al., 1998; Ruiz-Primo & Shavelson, 1996).

CMAPSCR Predicting Interview Performance

CMAPSCR correlated highly ($r = .750$) (McMillan, 1992) to and was a reliable predictor ($b = .323$, $p < .001$, $R^2 = .563$) of students' ($n = 17$) interview performance. (Interview performance was the sum—1 point each—of teacher-expert concepts and relationships captured by the transcript and any *additional* teacher-expert concepts and relationships in the concept map.) This finding was not surprising, because the concept maps constructed by students during the interview were based on their own verbal responses to the interview questions. Students had been asked to map what they believed to be "important" in each of their verbal responses. The high predictive validity ($p < .001$) of CMAPSCR for student performance in the interview suggests that the concept map tool can be important in verifying with a student his or her understanding, as verbalized in response to interview questions. Further, the emergent concept map can be compared with the interview transcript as a second piece of interview data that explicates how the student understands the topic. These findings also support and extend those reported by Wallace and Mintzes (1990), who found that students with the highest concept map scores on a science topic provided in interviews more "critical" concepts and relationships on that topic than did those students with the lowest concept map scores.

CMAPSCR and Achievement Test Scores

The validity of the concept map scoring procedures in this study is supported by the moderate to high correlations found between CMAPSCR and students' ($n = 17$) performance on the CAT. These statistically significant correlations were as follows: $r = .558$ ($p = .02$) for the CAT verbal component; and $r = .729$ ($p < .001$) for the CAT component total. Additionally, CMAPTER, which was derived exclusively from the presence of relationships, correlated highly ($r = .773$) with the CAT component total. Rice et al. (1998) also reported high correlations ($r = .82$ - $.87$) between students' concept map scores that were based solely on relationships and their performance on standardized achievement tests. Further, this study and the research by Rice et al. achieved high interrater reliabilities (reported as interrater agreement by Rice et al.) for scoring concept maps.

CMAPSCR and Dual Role of CFCs in GAC

Rye (1998) presented a frequency analysis of the relationships in these concept maps, which revealed that few ($n = 5$) of the 17 students mapped the *dual* role (ozone layer destruction and global warming) of CFCs

in GAC. Most (n = 12) students included only the problem of ozone layer destruction. The researchers became interested in the CMAPSCR of each subgroup (those who did and did not map the dual role of CFCs in GAC) and, most specifically, if there was a statistically significant difference in favor of those mapping the dual role. The mean CMAPSCR of the "dual role" subgroup was 91.2 (range of 79 to 111, SD = 13.882), whereas the mean score of the "ozone destruction role only" subgroup was 57.083 (range of 7 to 95, SD = 26.623). The Mann-Whitney U test revealed that the difference in mean CMAPSCR was statistically significant: U=6.5, Z=-2.477, p=.013. Recognizing that the small sample size limited the value of this finding, the scoring scheme employed in this study did yield on a subgroup basis concept map scores that were consistent with the degree of explication about the role of CFCs in GAC. However, on an individual basis, three of the 12 "ozone destruction role only" maps achieved a score equal to or higher than some of the "dual role" maps. The latter has implications for concept mapping rubrics used by teachers to assess classroom learning.

Conclusions and Implications

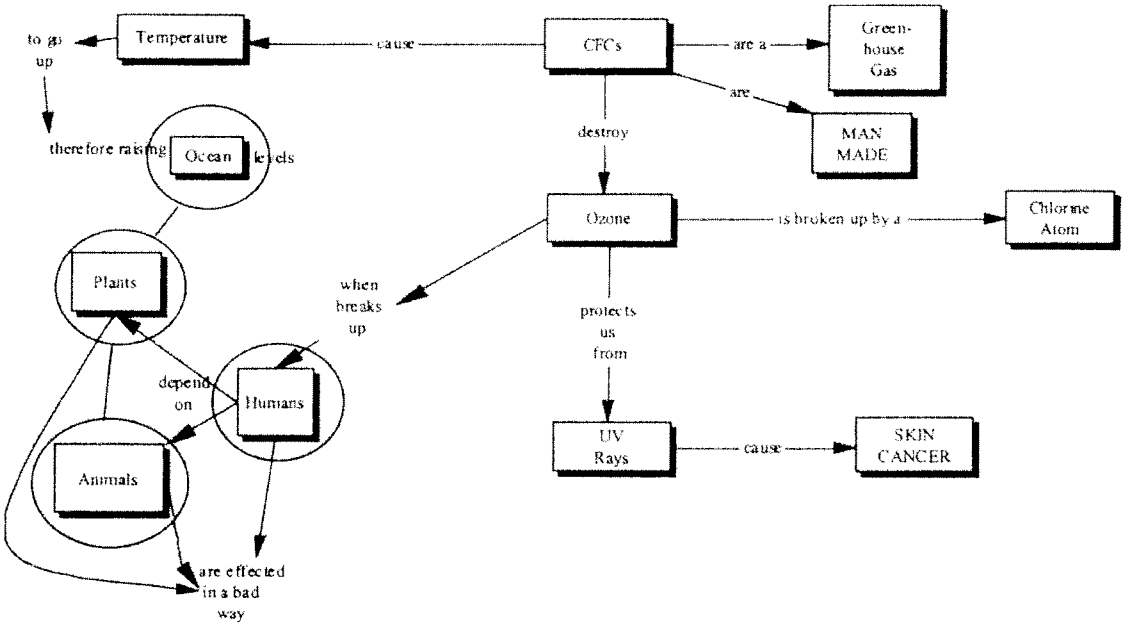
In this study, points were awarded to each concept map in a manner consistent with the research of Rice et al. (1998): The scoring scheme emphasized concept relationships and reflected the degree to which the student map "communicates understanding of concepts covered during instruction....information that the teacher thought was important and that the teacher actually taught (Lomask et al., 1992)" (p. 1124). The concept map scores (CMAPSCR) in this study had a high interrater reliability and good predictive validity for performance in the interview; and they correlated highly to students' total scores on a standardized achievement test (CAT). Further, the mean CMAPSCR of the subgroup explicating the dual role of CFCs in GAC was significantly greater than the subgroup mapping only "ozone depletion." Thus, the findings of this study respond to the call of Rice et al. and others to develop and validate scoring methods resulting "in scores that reflect a stronger relationship between concept maps and students' learning in science, scores that are reliable predictors of intended learning outcomes (Ruiz-Primo, 1996, Shavelson et al., 1993)" (p. 1104).

The concept map score that was based solely on concept relationships (CMAPTER) was a statistically significant predictor of students' Pathfinder index, which Jonassen contended to be an excellent measure of students' structural knowledge. The correlation of these "relationship" concept map scores with the Pathfinder index (r = .539) approached the correlation (r = .608) obtained by McClure et al. (1999) between their "relational with master map score" and the "neighborhood similarity index of student to teacher maps." Accordingly, the findings of this study lend further support to the recommendations by McClure et al. for science teachers to "use some form of relational scoring method, preferably with a master map" (p. 490). However, we agree with Kinchin et al. (2000) that additional aspects of scoring need consideration in order to "make the benefits of concept mapping more accessible to the classroom teacher" (p. 43).

The processes employed in this study to develop referents for and score the student maps was time consuming and rather complex. As Rice et al. (1998) also concluded, research is needed to develop more efficient concept map rubrics that teachers will view as practical for assessing student learning. In developing such rubrics, attention should be given to breadth as well as depth of understanding relative to the instruction provided, because, as Ruiz-Primo et al. (2001) pointed out, it is possible to achieve relatively high scores by explicating many accurate concept relationships about one subtopic and few to no relationships about another. This limitation can be illustrated by contrasting two concept maps that emerged from this study (see Figure 2 and Figure 5).

The map illustrated in Figure 2, which does not connect CFCs to global warming, scored 93 points: This and one other "ozone destruction role only" map scored near the high end of the range (7 - 111 points) for all maps because of the depth of understanding explicated about ozone depletion. The map shown by Figure 5, which explicates the dual role of CFCs in GAC (ozone destruction *and* global warming), scored lower at 86 points, yet had the more robust understanding about the role of CFCs in GAC. Although the scoring rubric yielded a significantly higher mean CMAPSCR for the subgroup that mapped both problems as opposed to just ozone destruction, it was possible for an individual student to explicate a conception of only one of the key problems caused by CFCs and achieve a higher score than a student with a more robust explication of those problems. A more practical and equitable assessment of classroom learning might be achieved through a rubric that set forth, for each key subtopic (e.g., composition of CFCs, sources of CFCs, problems caused by CFCs), incremental levels of performance and corresponding point values (e.g., 0 points for "no valid relationships about subtopic" up to 25 points for "several valid teacher-expert relationships showing

Figure 5. *Sample student concept map (CMAPSCR = 86 points) that included the greenhouse gas nature of CFCs that emerged from the interviews.*



breadth of understanding about subtopic"). Points earned for a specific subtopic might be weighted (e.g., problems caused by CFCs "x 2") to emphasize relative importance. Rubrics such as these that assess for breadth and depth would take advantage of the capacity of concept maps to illustrate "the knowledge transformation that occurs as students progress from the knowledge of a relative novice to that of an expert" (Edmonson, 2000, p. 36).

References

Acton, W., Johnson, P., & Goldsmith, T. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86,* 303-311.

Anderson, T., & Huang, S. (1989). *On using concept maps to assess the comprehension effects of reading expository text.* (Technical Report No. 483). Urbana-Champaign: Center for the Studying of Reading, University of Illinois at Urbana-Champaign. (ERIC Document Reproduction Service No. ED 310 368).

Arnaudin, M., Mintzes, J., Dunn, C., & Shafer, T. (1984). Concept mapping in college science teaching. *Journal of College Science Teaching, 14,* 117-121.

Borg, W., & Gall, M. (1971). *Educational research methods.* New York: David McKay Company, Inc.

CTB McGraw-Hill. (1986). CAT: California achievement test forms E & F: Norms book, Dec.-Feb., Montery, CA: CTB McGraw-Hill.

Edmondson, K. (2000). Assessing science understanding through concept maps. In J. Mintzes, J. Wandersee, & J. Novak (Eds.), *Assessing science understanding* (pp. 15 - 40). San Diego, CA: Academic Press.

Gall, M., Borg, W., & Gall, J. (1996). *Educational research* (6th ed.). White Plains, NY: Longman Publishers.

Goldsmith, T., & Davenport, D. (1990). Assessing structural similarity of graphs. In R. Schvaneveldt (Ed.), *Pathfinder associative networks* (pp. 31-52). Norwood, NJ: ABLEX Publishing Corporation.

Goldsmith, T., Johnson, P., & Acton, W. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83,* 88-96.

Good, R., Novak, J., & Wandersee, J. (Eds.) (1990). Special issue: Perspectives on concept mapping. *Journal of Research in Science Teaching, 27*(10).

Gordon, S., Schmierer, K., & Gill, R. (1993). Conceptual graph analysis: Knowledge acquisition for instructional system design. *Human Factors, 35,* 459-481.

Heinze-Fry, J. A. (1987). *Evaluation of concept mapping as a tool for meaningful education of college biology students*. Unpublished doctoral dissertation. Cornell University, Ithaca, NY.

Horton, P. B., McConney, A., Gallo, M., Woods, A., Senn, G., & Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Science Education, 77*, 95-111.

Inman-Anderson, L., & Ditson, L. (1999). Computer-based concept mapping: A tool for negotiating meaning. *Learning and Leading With Technology, 26*, 6-13.

Jonassen, D. (1993). Changes in knowledge structures from building semantic net versus production rule representations of subject content. *Journal of Computer-Based Instruction, 20*, 99-106.

Jonassen, D. (2000). *Computers as mindtools for schools*. Englewood Cliffs, NJ: Prentice-Hall.

Judd, C. M., & McClellan, G. H. (1989). *Data analysis: A models approach*. Florida: Harcourt Brace Jovanovich.

Kinchin, I., Hay, D., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research, 42*, 43-57.

Lomask, M., Baron, J. B., Greig, J., Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. Symposium presented at the annual meeting of the National Association for Research in Science Teaching, Cambridge, MA.

Luft, J. (1998). Challenging myths. *The Science Teacher, 66*(4), 40-43.

Malone, J., & Dekkers, J. (1984). The concept map as an aid to instruction in science and mathematics. *School Science and Mathematics, 84*, 220-231.

McClure, J., Sonak, B., & Suen, H. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36*, 475-492.

McMillan, J. (1992). *Educational research*. New York: Harper-Collins Publishers.

Moore, G. (1983). *Developing and evaluating educational research*. Boston: Little, Brown, and Company.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy of Sciences.

Novak, J. (1992). A view on the current status of Ausubel's assimilation theory of learning. A paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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

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

Patton, M. (1990). *Qualitative evaluation and research methods*. London: Sage Publications.

Rice, D., Ryan, J., & Samson, S. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching, 35*, 1103-1127.

Roth, W. M., (1994). Science discourse through collaborative concept mapping: New perspectives for the teacher. *International Journal of Science Education, 16*, 437-455.

Rubba, P., Wiesenmayer, R., Rye, J., & Ditty, T. (1996). The leadership institute in STS Education: A collaborative teacher enhancement, curriculum development and research project of Penn State University and West Virginia University with rural middle/junior high school science teachers. *Journal of Science Teacher Education, 7*, 1-10.

Rubba, R., Wiesenmayer, R., Rye, J., McClaren, M., Sillman, K., Yorks, K., Yukish, D., Ditty, T., Morphew, V., Bradford, C. Dorough, D., & Arabas, K. (Eds). (1995). *Global atmospheric change: The enhanced greenhouse effect, ozone layer depletion, and ground level ozone pollution. A Science-Technology-Society Issue Investigation and Action Unit for the middle grades* [Online]. University Park, PA: The Pennsylvania State University. Available: http://www.ed.psu.edu/CI/papers/sts/gac-main.html

Ruiz-Primo, M., & Shavelson, R. (2000). Psychometrics of assessing student understanding. In J. Mintzes, J. Wandersee, & J. Novak, (Eds). *Assessing science understanding* (pp. 303-341). San Diego, CA: Academic Press.

Ruiz-Primo, M., Schultz, S., Li, M., & Shavelson, R. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38*, 260-278.

Ruiz-Primo, M., & Shavelson, R. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*, 569-600.

Rye, J. (1995). *An investigation of the concept map as an interview tool to facilitate externalization of conceptual understandings associated with global atmospheric change by eighth grade physical science students*. Unpublished doctoral dissertation, The Pennsylvania State University, University Park, PA.

Rye, J. (1998). Understanding the role of chlorofluorocarbons in global atmospheric change. *Journal of Geoscience Education, 46*, 488-493.

Rye, J., & Rubba, P. (1998). An exploration of the concept map as an interview tool to facilitate the externalization of students' understandings about global atmospheric change. *The Journal of Research in Science Teaching, 35*, 521-546.

Rye, J., & Rubba, P. (2000). Students' understandings of global warming: Implications for STS education beyond 2000. In D. Kumar & K. Chubin (Eds.), *Science, technology, & society: A sourcebook on research and practice*. New York: Kluwer Publishers.

Schvaneveldt, R. (Ed.). (1990). *Pathfinder associative networks*. Norwood, NJ: ABLEX Publishing Corporation.

Shavelson, R., Lang, H., & Lewin, B. (1993). *On concept maps as potential "authentic assessments" in science.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 367 691)

Shymansky, J. A., Woodworth, G. Norman, O. Dunkhase, J. Matthews, C. & Liu, C. (1993). A study of changes in middle school teachers' understanding of selected ideas in science as a function of an in-service program focusing on student preconceptions. *Journal of Research in Science Teaching, 30*, 737-755.

Stensvold, M., & Wilson, J. (1990). The interaction of verbal ability with concept mapping in learning from a chemistry laboratory activity. *Science Education, 74*, 473-480.

Wallace, J., & Mintzes, J. (1990). The concept map as a research tool: Exploring conceptual change in biology. *Journal of Research in Science Teaching, 27*, 1033-1052.

Wandersee, J., Mintzes, J., & Novak, J. (1994). Research on alternative conceptions in science. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 177-210). New York: MacMillan Publishing Company.

White, R.T. (1985). Interview protocols and dimensions of cognitive structure. In L.H. West & A.L. Pines (Eds.), *Cognitive structure and conceptual change* (pp. 51-58). Orlando, FL: Academic Press, Inc.

White, R., & Gunstone, R. (1992). *Probing understanding.* New York: The Falmer Press.

Xiufeng, L., & Hinchey, M. (1996). The internal consistency of a concept mapping scoring scheme and its effect on prediction validity. *International Journal of Science Education, 18*, 921-937.

Editors' Note: James A. Rye, Educational Theory and Practice, West Virginia University; Peter A. Rubba, Professor of Education, Director of Academic Programs, The World Campus, Penn State University.

This research was supported in part by the National Science Foundation, Grant TEP-9150232. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to James A. Rye, Educational Theory and Practice, 604L Allen Hall, West Virginia University, Morgantown, WV 26506.

Electronic mail may be sent via Internet to jrye@wvu.edu