DOCUMENT RESUME

ED 370 992 TM 021 588

AUTHOR Liu, Xiufeng
TITLE The Validity and Reliability of Concept Mapping as an Alternative Science Assessment when Item Response Theory Is Used for Scoring.
SPONS AGENCY Saint Francis Xavier Univ., Antigonish (Nova Scotia).
PUB DATE Apr 94
CONTRACT UCR192
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Assisted Testing; Correlation; *Educational Assessment; Estimation (Mathematics); Foreign Countries; *Item Response Theory; Junior High Schools; Junior High School Students; Reliability; *Scoring; Validity
IDENTIFIERS *Alternative Assessment; Canada; *Concept Mapping; Concept Maps; Science Achievement

ABSTRACT

Problems of validity and reliability of concept mapping are addressed by using item-response theory (IRT) models for scoring. In this study, the overall structure of students' concept maps are defined by the number of links, the number of hierarchies, the number of cross-links, and the number of examples. The study was conducted with 92 students in four classes at a junior high school in Canada. Results show that IRT scoring of concept maps is generally valid and reliable. The correlation between IRT ability estimates and the total concept-mapping scores based on a scoring scheme proposed by J. D. Novak is significant, demonstrating that it is valid to score concept maps on the basis of structural characteristics defined by links, hierarchies, cross-links, and examples. The advantage of IRT scoring is reliability. Some computer packages with concept-mapping facilities and IRT scoring are described. One figure and eight tables present study findings. (Contains 29 references.) (SLD)

The Validity and Reliability of Concept Mapping as an Alternative Science Assessment when Item Response Theory is Used for Scoring

Xiufeng Liu, Ph. D.
Department of Education
St. Francis Xavier University
P. O. Box 5000
Antigonish, N. S.
B2G 2W5
Tel. (902) 867 -2384
e-mail: liu@essex.stfx.ca

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 8, 1994.

The Validity and Reliability of Concept Mapping as an Alternative Science Assessment when Item Response Theory is Used for Scoring*

Introduction

Concept mapping as an alternative science assessment has been discussed intensively in the literature. Concept mapping is a technique used to represent the relationships between concepts in a two-dimensional graph. It was originally used by Novak and his colleagues (Novak and Gowin, 1984) as an instructional and assessment tool for science learning during the 1970's. Concept mapping has been primarily used as a diagnostic tool to assess students' conceptions (Moreira, 1985; Ross and Mundy, 1991; Wallace and Mintzes, 1990). More recently, concept mapping has been used as an alternative science classroom achievement assessment. For example, Gaffney (1992) used concept mapping to evaluate students' achievement on botany and natural communities in a fifth grade class. Tippins and Dana (1992) used concept mapping as a culturally relevant assessment. The use of concept mapping for assessing learning processes has also been reported. Fleener and Marek (1992) used concept mapping to assess student's learning in the three phases of a learning cycle (exploration, conceptual invention, and expansion). Roth (1992) also used concept mapping to assess student's learning/investigation process. The comprehensive use of concept mapping in designing instruction and assessment has been reported by Barenholz and Tamir (1992).

Although considerable effort in concept mapping as an alternative assessment has been made as reviewed above, the empirical findings on the validity and reliability of using concept mapping as an alternative achievement assessment are very preliminary and far from conclusive. In Liu's (1993) study, students' concept mapping scores correlated significantly with students' scores on the conventional tests. This result is consistent with other studies. For example, Bousquet (1982) found that concept map scores could predict

students' achievements in a college natural resources class. Fraser and Edwards (1985) found that students who scored at high levels on end-of-unit tests showed high levels of concept mastery as indicated by the concept maps they made. However, opposite conclusions about the prediction validity of concept mapping have also been reported. For example, a poor correlation between students' concept map scores and their scores on standardized tests was reported by Novak, Gowin and Johansen (1983). In Trigwell and Sleet (1990), it was also found that concept mapping scores had a low correlation with traditional examination scores.

The diversified conclusions about the predication validity of concept mapping reported may be due to the different scoring schemes used. There are various scoring schemes of concept maps reported in the literature, such as that in Cleare (1983), in Novak and Gowin (1984), in Schreiber and Abegg (1991), in Vargas and Alvarez (1992), and in Wallace and Mintzes (1990). The scoring schemes proposed so far are based on the evaluation of concept map aspects, such as the number of correct links, hierarchies, cross-links and examples. For example, Novak and Gowin (1984) proposed to measure valid links (1 point each), valid hierarchies (5 points each), valid cross-links (2 or 10 points each depending on whether or not the cross-link is significant), and valid examples (1 point each). Schreiber and Abegg's (1991) scoring scheme includes the hierarchical structure of a concept map, identified propositions, and the actual validity versus implied validity of concept map components. The overall score of a concept map is defined as

$$X = [x - n\,(b+c)] + b/c, \qquad (1)$$

Where

X is the overall concept map score;

x is the initial tally of points (ratios) awarded for recognition of hierarchical, propositional and valid constructs on a concept map;

n is the number of strands in a concept map;

b is the summed ratios of number of vocabulary terms to number of hierarchical levels (per strand); and

c is the summed ratios of number of valid connecting lines to total number of connecting lines drawn.

Because different scoring schemes emphasize different concept map aspects and award different weights to concept map aspects, a same concept map produced by a student may be given different concept map scores under different scoring schemes.

Reliability is another issue which has not been addressed intensively in the literature. Although the internal consistency among valid links, valid hierarchies, valid cross links and valid examples when using Novak's scoring scheme was fairly high (.65) according to Liu's (1993) study, the inter-rater reliability has not been reported in the literature. A low inter-rater reliability may be expected. One reason for the expected low inter-rater reliability is that, in students' concept maps, some links and cross-links may be connected without linking words. This situation is common in novices' concept maps. Therefore, it is difficult to judge a link to be correct or incorrect since the correctness or incorrectness may depend on the assumption made by the rater that what linking word might be implied by the student. A more fundamental reason for the expected low inter-rater reliability is that students' conceptions as demonstrated by links, hierarchies, cross-links and examples in a concept map are intrinsically difficult to judge as being totally incorrect or totally correct, because many studies have showed that students conceptions may make sense in some aspects but may not be completely consistent with scientific views. Therefore students' conceptions may be better considered along a continuum from nonsense to scientific conceptions (Driver and Erickson, 1983; Driver and Bell, 1986; O'Loughlin, 1992; etc.). The above discussion is not to object to the possibility that a high inter-rater reliability may be achieved if an intensive training of raters is provided and sufficient discussion among raters is allowed.

The present study addresses the above validity and reliability problems by employing Item Response Theory (IRT) models for scoring. IRT is a mathematical attempt to model students' responses to test items into item characteristics (such as item difficulty) and students' abilities. Two advantages have been claimed for IRT applications: item parameter estimates are invariant from the sample used to calibrate, and ability parameter estimates are invariant from the test used to calibrate. Thus, when IRT models are used for scoring, students' achievements can be compared even if they do not write the same test (Hambleton, 1989). Samejima (1969) extended the traditional IRT models to graded IRT models, thus students' responses to an item no longer have to be scored dichotomously (as right or wrong), they can be graded as categories. When IRT is used for scoring concept maps, the four aspects of a concept map, links, hierarchies, cross-links and examples, are considered as "test items", and the numbers of links, hierarchies, cross-links and examples are considered as students' categorical responses to the "test items". Therefore, by applying graded IRT models to students' responses, it is possible to obtain students' ability estimates.

IRT scoring emphasizes the overall structure of students' concept maps, instead of the correctness of a specific concept map aspect. In this study, the overall structure of students' concept maps are defined by the number of links, the number of hierarchies, the number of cross-links and the number of examples. The analysis of structural characteristics of students' concept maps was reported by Wilson (1993). In Wilson's study, a 24 x 24 matrix representing the inter-relationships between the 24 concepts provided for concept mapping was created. The matrix was defined by whether or not a connection between the two concepts existed. By applying non-parametric multidimensional scaling, the coordinates on the three dimensions were obtained. The canonical correlation between the coordinates and students' conventional achievement test scores was found to be significant. However, Wilson's study did not provide a scoring scheme based on the structural characteristics of concept maps. This study proposes a

scoring scheme according to the structural characteristics of concept map and studies the validity and reliability of this scoring scheme.

Methodology

Procedure

The following procedures were used to obtain students' ability estimates and item characteristics:

1. each concept mapping task is considered as a test which contains four "test items": links, hierarchies, cross-links, and examples;

2. students' concept maps are measured by the number of links, number of hierarchies, number of cross-links, and number of examples. Those numbers are categorical responses to the "test items" in procedure 1;

3. apply Samejima's graded IRT model to students' categorical responses to obtain students' ability estimates and characteristic estimates of "test items".

The software used to estimate students' abilities and item characteristics based on Samejima's graded IRT model was MULTILOG (Thissen, 1991). MULTILOG has been widely used for categorical IRT analysis for years.

Two aspects of validity, construct validity and consequential validity, were studied following the conceptual analysis by Messick (1989). As for the construct validity, two aspects were examined: internal construct validity was assessed by examining the difficulty and discrimination of the four "test items" and by examining the inter-relationship between the four "test items"; external construct validity was assessed by examining the relationship between students' IRT ability estimates and students' concept mapping scores according to Novak's scoring scheme and by examining the relationship between students' IRT ability estimates and their scores on the conventional tests.

The reliability in IRT applications is defined by Standard Error of Estimation (SEE) when maximum likelihood estimation is employed. Since SEE is defined at each

ability estimation level, an average SEE over all ability estimates can be calculated as the SEE for a test. Based on the average SEE, the reliability of a test can be calculated according to the conventional formula defined as

$$\rho_{xx} = 1 - \sigma_e^2/\sigma_x^2, \qquad (2)$$

where

ρ_{xx} is the reliability of the test;

σ_e is the standard error of estimation (SEE); and

σ_x is the standard deviation of students' ability estimations.

Data source

This study was conducted in four classes at a junior high school in a Canadian Atlantic province. Four grade 7 general science classes taught by two teachers participated in the study. In this school, students are randomly assigned to classes after the top 60 high achievement and bottom 30 lower achievement students are assigned. The two classes of the top 60 higher achievement students are called enriched classes, the class of bottom 30 lower achievement students is called an adjusted class, and the classes randomly formed are called regular classes Among the four classes participating in this study, one was an adjusted class, one a regular class and two were enriched classes.

Concept mapping technique was introduced to the classes by the two teachers according to the procedures outlined in Novak and Gowin (1984) at the beginning of the term. The two teachers are very familiar with concept mapping and have used concept mapping in their instruction for years. The data used in this study was from an end-of-the-unit test administered toward the end of the first term during the academic year when the students had grasped basic concept mapping techniques. After finishing the unit, the teachers gave the classes a conventional test as before, and also a concept mapping test.

The concept mapping was administered based on a list of concepts provided by the teachers. The list of concepts was identified from the conventional test and students were asked to use some or all of the concepts provided to draw concept maps. Students might also use any concepts not included in the lists. The conventional test took one class period (45 minutes) and concept mapping took one class period as well. Figure 1 is a sample student's concept map. From the existence of cross-links and frequent usage of linking words, we may infer that the student grasped concept mapping techniques quite well. Students' concept maps were then evaluated by the numbers of links, hierarchies, cross-links, and examples, and those numbers constituted the students' responses to the four "test items": links, hierarchies, cross-links, and examples. Samejima's graded IRT model was applied to analyze students' responses.

Insert Figure 1 about here

Results

Validity

Internal validity

Columns 1 to 4 in Table 1 list numbers of links, hierarchies, cross-links and examples in students' concept maps. From the bottom of Table 1 (*M* and *SD*), we know that the average number of links is 22, the average number of hierarchies is 6.2, the average number of cross-links is 1, and the average number of examples is 2.2. Also from the table, a higher variation can be observed in the number of links and the number of examples in students' concept maps. The inter-correlation between the numbers is listed in

Insert Tables 1 and 2 about here

Table 2. From Table 2, it can be seen that the number of links, the number of hierarchies and the number of cross-links are significantly correlated with each other, but the number of examples is not significantly correlated with any of the numbers.

Since MULTILOG can only process a maximum of 10 categories, but the maximum numbers of links, hierarchies and exampies are 39, 14, and 21 respectively; a transformation was conducted before applying MULTILOG to the data file. The rationale for the data transformation was that students' categories would be extensively distributed between 0 and 10 so that a higher discrimination power could be expected. Based on this rationale, the numbers of links were divided by 4 and rounded to the nearest whole number. All the numbers of hierarchies and examples greater than 10 were re-coded as 10 (the highest category). Only 6 out of 92 students had a number of hierarchies greater than 10, and only 4 out of 92 students had a number of examples greater than 10. After applying MULTILOG to the transformed data file, the a and b parameter estimates for each "test item" (links, hierarchies, cross-links and examples) were obtained and are listed in Table 3.

In Samej.. ıa's graded IRT model, an Item Characteristic Curve (ICC) is defined as

$$P(x=k) = \frac{1}{1+\exp[-a(\theta-b_{k-1})]} - \frac{1}{1+\exp[-a(\theta-b_k)]}, \qquad (3)$$

where

k is the category an examinee responds to an item, $k = 1, 2, \ldots, m$, m is the highest category;

θ is an examinee's ability;

a is the slope which defines the discriminating power of an item;

b_j is the threshold which defines the difficulty of an item at category k;

$p(x=k)$ is the probability of an examinee with ability θ answering an item of discrimination power a and difficulty b_j with category k.

From Table 3, it can be seen that the number of links, the number of hierarchies, and the number of cross-links have a relatively high discriminating power (>1.0), but the number of examples has a relatively low discriminating power (<1.0). It can also be seen that, for the numbers of links and hierarchies, the mean θ (which is -0.256 from Table 1) is between the difficulties of category 5 and category 6, meaning that an average ability student is likely to have 20 to 24 (5 x 4 to 6 x 4) links in their maps. Similarly, an average ability student is likely to have 4 to 5 hierarchies, 1 to 2 cross-links and 2 to 3 examples. The marginal reliability for the estimation is .78, and the negative twice the loglikelihood is 11.4, indicating that the graded IRT model fits the data quite well.

Insert Table 3 about here

External validity

Columns 5 to 8 in Table 1 list the number of valid links, the number of valid hierarchies, the number of valid cross-links, and the number of valid examples. Also in Table 1, students' ability estimates after applying MULTILOG are included in column 9. By employing Novak's scoring scheme, i.e. awarding each valid link 1 point, each valid hierarchy 5 points, each valid cross-link 10 points, and each valid example 1 point, the total concept mapping scores were calculated. The inter-correlation between the IRT ability estimates, the number of valid links, the number of valid hierarchies, the number of valid cross-links, the number of valid examples, and the total concept mapping scores were computed, the correlation matrix is listed in Table 4.

From Table 4, it can be seen that the IRT ability estimates are significantly correlated with the number of valid links, the number of valid hierarchies, the number of valid cross-links, and the total concept mapping scores. The total concept mapping scores are significantly correlated with the number of valid links, the number of valid hierarchies, and the number of valid cross-links.

Insert Table 4 about here

Also included in Table 1 are students conventional test scores (column 10). The inter-correlation between students' IRT ability estimates, their conventional scores, the number of links, the number of hierarchies, the number of cross-links, and the number of examples was computed. The correlation is included in Table 2. From Table 2, it can be seen that students' ability estimates are significantly correlated with their conventional test scores, with number of links, with number of hierarchies, and with number of cross-links. The correlation between students' ability estimates is not significantly correlated with the number of examples. From Table 2, it can also be seen that students conventional test scores are significantly correlated with the number of cross-links, in addition to a significant correlation with their IRT ability estimates.

The inter-correlation between student conventional test scores, the number of valid links, the number of valid hierarchies, the number of valid cross-links, the number of valid examples, and the total concept mapping scores was also computed, and the correlation is listed in Table 5.

From Table 5, it can be seen that students' conventional test scores are significantly correlated with student total concept mapping scores and the number of the valid cross-links.

Insert Table 5 about here

In order to study the effect of different student groups, another MULTILOG application was conducted by dividing the examinees into three different groups: enriched class, regular class, and adjusted class. The inter-correlation matrices for the three groups are listed in Tables 6 to 8. From Tables 6 to 8, it can be seen that students' IRT ability estimates are not significantly correlated with their conventional test scores.

Insert Tables 6-8 about here

Consequential validity

When IRT is used for scoring concept maps, the immediate advantage is to free teachers from the uncertainty about whether or not a proposition in a student's map is correct or incorrect, it is straight forward to count the numbers of links, hierarchies, cross-links and examples. Concept map scoring time will also be reduced when IRT is used for scoring, it is possible to score students' concept maps at a rate of one map per minute. Teachers' preparation time for tests will be reduced as well when concept mapping is used as an alternative assessment. Teachers only need to provide a list of concepts, and students feel free to add any concept not provided in the list.

As for students, IRT scoring could make adaptive testing possible. Theoretically, the difficulty of a concept mapping test is appropriate for any student with any level of ability. In this sense, a concept mapping test is adaptive to students' ability levels. Students also feel less intimidated on a concept mapping test, since it allows students more freedom to construct and express their conceptions.

A criticism of conventional concept mapping scoring is that students' scores are concept-dependent. IRT concept mapping scoring provides a concept-free ability estimates. For example, one group of students concept mapping scores on Acid and Base will hardly be compared with another group of students concept mapping scores on Forces when Novak's scoring scheme is used, because their concept maps are based on different concepts. If IRT is used for scoring, two groups of students can be compared by estimating both students' abilities concurrently. In this situation, there will be eight "test items", four for each group. The categorical responses to the four "test items" not responded to by each group are coded as "Not-reached". The final ability estimates for both student groups will be on the same scale and therefore can be directly compared.

The practical difficulty for using IRT as scoring is that an IRT parameter estimation program and computer are necessary. IRT parameter estimation programs, such as MULTILOG, cost a few hundred dollars. The computer skills needed to prepare and run an IRT parameter estimation program are minimum, a few step by step instructions will do the job. IRT parameter estimation usually entails a computer of at least 386, although a slower computer such as 286 may also work if it is not rush to have results, as is the case for most classroom assessments. Most schools have been equipped with at least one IBM compatible computer.

The meaning of IRT ability estimates can hardly be understood immediately. For example, how to interpret an ability estimate of -.005? We have been using percentage scores for decades and refer percentage scores to the percentages of items answered correctly when multiple choice items are used. IRT ability estimates are not on a ratio scale, at best on an interval scale, this is the same as percentage scores. The IRT ability estimates may be better referred to proficiency levels as suggested by Hambleton (1991). It is possible to establish a transformational relationship between IRT ability estimates and conventional test scores such as percentage scores. Examples of transformation are those in Woodcock (1978) and Wright (1977).

Reliability

The last column in Table 1 lists the Standard Error of Estimation for all the abilities estimated. The average standard error of measurement for the test is calculated as follows

$$\overline{SEE} = \sqrt{(\sigma_1^2 + \sigma_2^2 + \ldots \sigma_i^2 \ldots + \sigma_{92}^2)/92}, \qquad (4)$$

where σ_i is the standard error of estimation for a student's IRT ability estimate.

By substituting the values in Table 1 to formula 4, the average standard error of measurement for the test was calculated as .635. By substituting the average standard error of estimation (.635) and the standard deviation of IRT ability estimates (1.653) into formula 2, it can be calculated that the reliability of the test is .85.

Discussion

The results presented above show that IRT scoring of concept maps is generally valid and reliable. The correlation between IRT ability estimates and the total concept mapping scores based on Novak's scoring scheme is significant. The significant correlation between IRT ability estimates and students' conventional test scores is also consistent with the results when Novak's scoring scheme was used. This demonstrates that it is a valid approach to score students' concept maps based on the structural characteristics as defined by the numbers of links, hierarchies, cross-links, and examples. The advantage of IRT scoring is the reliability. As discussed above, the reliability is as high as .85. This reliability level should be sufficient in most classroom testing situations.

The number of examples in the scoring scheme does not contribute to the validity very much. It does not have a high discrimination power as indicated in Table 3. It is not significantly correlated with numbers of links, hierarchies and cross-links. This indicates

that the inclusion of the number of examples in the scoring scheme does not contribute significantly to the internal validity. The numbers of examples is not significantly correlated with students' IRT ability estimates, nor with their conventional test scores. Even in Novak's scoring scheme, the number of valid examples is not significantly correlated with the total concept mapping scores. This indicates that the number of examples in the scoring scheme does not contribute significantly to the external validity. The number of examples and valid examples in Table 1 show that there is a high variation in the number of examples in students' concept maps. Most students do not have any examples in their concept maps at all, while a few have more than 10 examples. This seems to indicate that using examples in concept mapping may not be a stable characteristic of students' conceptual understanding. It may also be possible that the use of examples is closely related to the topic of the concept mapping, i.e., some topics may entail more examples and some topics may entail fewer examples. If this is the case, the universality of examples as a characteristic of the students' conceptual framework is questionable.

Although there is a significant correlation between IRT ability estimates and students' conventional test scores, this significant correlation does not exist within specific student groups such as enriched, adjusted and regular classes. This seems to suggest that concept mapping scores can not predict students' conventional test scores if the student group is sufficiently homogeneous. In Table 1, students 1 to 59 are enriched class students, students 60 to 66 are adjusted class students, and students 67 to 92 are regular class students. From the distributions of student conventional test scores and IRT ability estimates, we can see that the variation in student conventional test scores in each group appears to be less than that in IRT ability estimates in each group. It seems that the conventional test is not as discriminating as concept mapping is. These results may explain the insignificant correlation between concept mapping scores and conventional test scores reported before (such as Novak, Gowin and Johansen, 1983; Trigwell and Sleet,

1990). If concept mapping is more discriminating, a sensible hypothesis is that concept mapping as an alternative science assessment may be more appropriate in large scale assessment when students are more heterogeneous if prediction validity is of interest. Of course, the insignificant correlation is also possibly due to the hypothesis that concept mapping and conventional tests assess different aspects of knowledge as suggested by Trigwell and Sleet (1990).

A complete computer package including concept mapping facilities and IRT scoring may be more convenient for classroom use. Currently, a few concept mapping computer packages, such as SemNet (Fisher, 1990) and *Inspiration* by Inspiration Software Inc. on MacIntosh, are available. IRT parameter estimation programs, such as MULTILOG and ManyFacet developed at University of Chicago, have been used on IBM compatibles for years. An integrated system of concept mapping and IRT scoring on popular IBM compatibles will be much more convenient. Once this system is available, the test will be much more flexible than it is now. For example, a student may do concept mapping test any time he/she likes: in school or at home, during daytime or in the evening, because the concern for the confidentiality of concept mapping tests is much less than in traditional testing situations.

---

*The author sincerely thanks the participating teachers, Mr. Mike Hinchey and Ms. Leona Williams, for their help with the data collection. This study was made possible by a research grant (#UCR192) from St. Francis Xavier University .

References

Barenholz, H., & Tamir, P. (1992). A comprehensive use of concept mapping in designing instruction and assessment. Research in science and technological education, 10(1), 37-52.

Bousquet, W. S. (1982). An application of Ausubel's learning theory to environmental education: A study of concept mapping in a college natural resources management course. Unpublished doctoral dissertation: Ohio State University.

Cleare, Catherine C. (1983). Using concept mapping to detect intervention effective in improving pre-service elementary education majors' understanding of science topic. In Hugh Helm and Joseph Novak (Eds.) Proceedings of the International Seminar on Misconceptions in Science and Mathematics (vol. 1). Ithaca, NY: Cornell University. 272-81.

Driver, R., & Bell, B. (1986). Students' thinking and the learning of science: A constructivist view. School science review, March, 443-455.

Driver, R., & Erickson, G. (1983). Theories-in Action: Some theoretical and empirical issues in the study of students' conceptual frameworks in science, Studies in science education, 10, 37-60.

Fleener, M., & Marek, E. (1992). Testing in the learning cycle. Science scope, 15(6), 58-49.

Fisher, Kathleen M. (1990). Semantic networking: The new kid on the block. Journal of research in science teaching, 27(10), 1001-18.

Fraser, K., & Edwards, J. (1985). The effects of training in concept mapping on student achievement in traditional tests. Research in science education, 15, 158-165.

Gaffney, K. E. (1992). Multiple assessment for multiple learning styles. Science scope, 15(6), 54-55.

Hambleton, R. K. (1989). Principles and selected application of item response theory. In Robert Linn (Ed.), Educational measurement. New York: American Council on Education and Macmillan publishing company.

Hambleton, R. K., Swaminathan, H.,& Rogers, H. Jane (1991). Fundamentals of item response theory. Newbury Park: Sage publications.

Liu, X. (1993). The validity and reliability of concept mapping as an alternative science assessment. Paper presented at the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics. Ithaca: Cornell University

Messick, Samuel (1989). Validity. In Robert Linn (Ed.). Educational measurement (3rd. edition). American Council on Education and Macmillan Publishing Company. 13-103.

Moreira, Marco (1985). An alternative strategy for evaluation. Assessment and evaluation in higher education, 10(2), 159-68.

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

Novak, J. D.,& Gowin, D. B., & Johansen, G. T. (1983). The use of concept mapping and knowledge Vee mapping with junior high school science students. Science education, 67(5), 625-645.

O'Loughlin, Michael (1992). Rethinking science education: Beyond Piagetian constructivism toward a sociocultural model of teaching and learning. Journal of research in science teaching. 29(8), 791-820.

Roth, W. M. (1992). Dynamic evaluation. Science scope. 15(6), 37-40.

Ross, Bertram, & Mundy, Hugh (1991). Concept mapping and misconceptions: A study of high-school students' understandings of acids and bases. International journal of science education, 13(1), 11-23.

Schreiber, Deborah A., & Abegg, Gerold L. (1991). Scoring student-generated concept maps in introductory college chemistry. Paper presented at the annual meeting of the National Association for Research in Science Teaching. Lake Geneva, WI, April 7-10, 1991.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement, 4 (Part 2), Whole #17.

Thissen, D. (1991). MULTILOG User's guide. Chicago: Scientific Software, Inc.

Tippins, D., & Dana, F. (1992). Culturally relevant assessment. Science scope, 15(6), p50-53.

Trigwell, K., & Sleet, R. (1990). Improving the relationship between assessment results and student understanding. Assessment and evaluation in higher education, 15(3), 190-197.

Vargas, Elena Maldonado, & Alvarez, Hector Joel (1992). Mapping out Students' abilities. Science scope, 15(6), 41-43.

Wallace, Josephine D., & Mintzes, Joel J. (1990). The concept map as a research tool: Exploring conceptual change in biology. Journal of research in science teaching, 27(10), 1033-1052.

Wilson, J. M. (1993). The predictive validity of concept-mapping: Relationships to measures of achievement. Paper presented at the Third International Seminar on

Misconceptions and Educational Strategies in Science and Mathematics. Ithaca: Cornell University. August, 1993.

Woodcock, R. W. (1978). Development and standardization of the Woodcock-Johnson Psycho-Educational Battery. Hingham, MA: Teaching Resources Corporation.

Wright, B. D. (1977). Solving me: .ement problems with the Rasch model. Journals of educational measurement, 14, 97-116.

- ENVIRONMENT
- INTERACTION
- BIOTIC
- ABIOTIC
- NICHE
- HABITAT
- TEMPERATURE
- COMMENSALISM
- PARASITISM
- PRODUCERS
- CONSUMERS
- PHOTOSYNTHESIS
- HERBIVORE
- CARNIVORE
- PREDATOR
- PREY
- SCAVENGER
- DECOMPOSER
- BACTERIA
- FUNGUS
- TOLERANCE
- MOISTURE
- LIGHT
- MUTUALISM

Figure 1

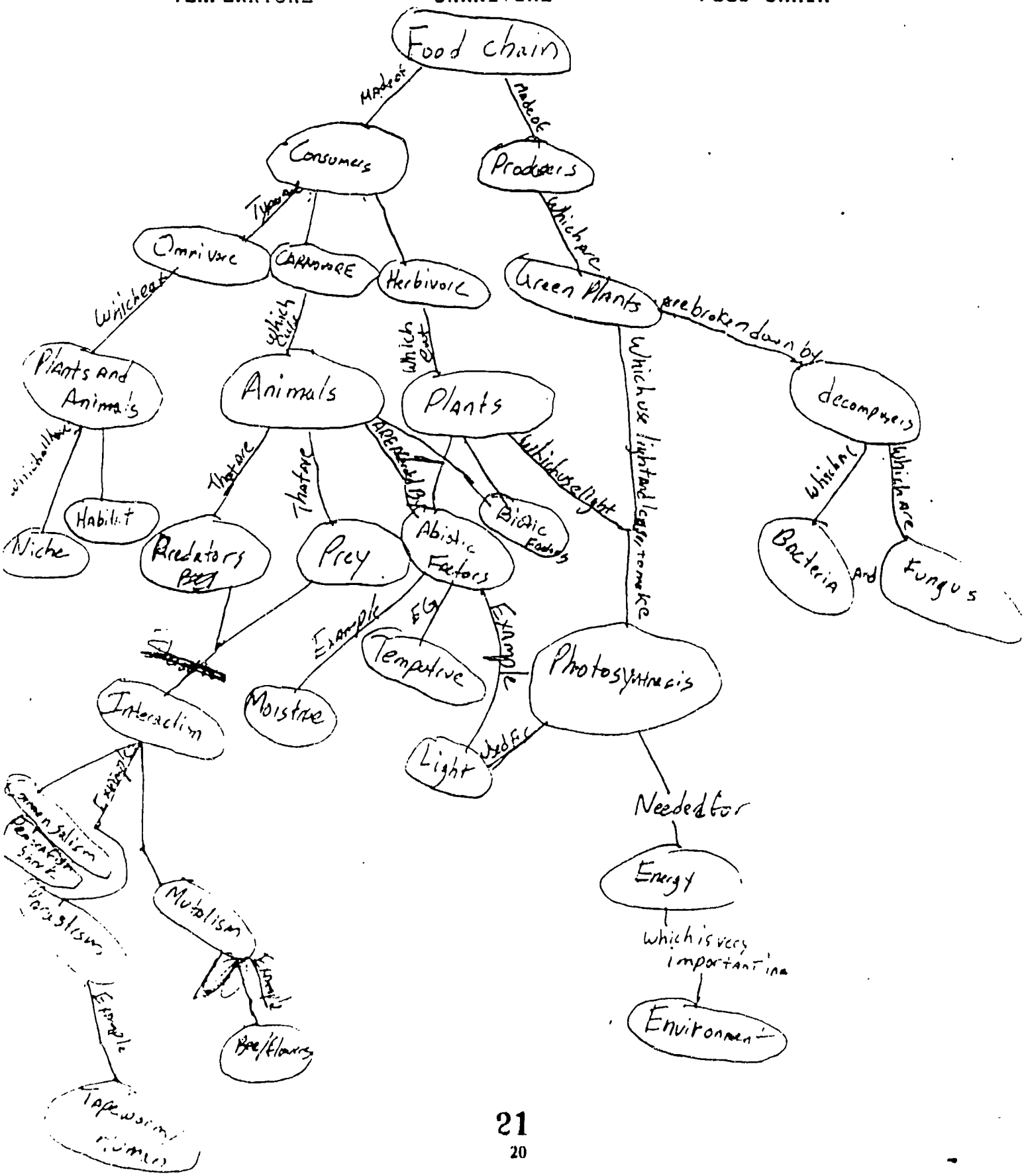

Table 1

Structural characteristics of students' concept maps: number of links/valid links (c1/c5), number of hierarchies/valid hierarchies (c2/c6), number of cross-links/valid cross-links (c3/c7), number of examples/valid examples (c4/c8), IRT ability estimates (c9), conventional test scores (c10), and standard errors of estimation (c11)

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
1	25	6	3	0	25	6	1	0	- .362	87	.467
2	22	7	1	4	22	7	1	2	- .155	88	.512
3	23	9	1	0	21	7	1	0	.655	95	.582
4	25	4	0	2	25	3	0	0	- .461	90	.549
5	26	6	4	6	26	6	4	5	.775	95	.468
6	39	13	0	4	38	13	0	4	3.550	91	1.409
7	36	9	3	0	32	8	3	0	1.716	88	.461
8	35	14	5	0	35	4	5	0	2.405	84	.545
9	27	8	1	12	27	8	1	12	.974	93	.471
10	15	6	0	0	15	6	0	0	- .680	86	.516
11	32	10	4	4	29	8	0	1	1.713	93	.484
12	26	5	0	1	26	5	0	1	.264	95	.571
13	39	9	4	0	38	8	4	0	2.352	91	.542
14	20	7	1	5	16	5	2	0	.111	95	.520
15	25	7	2	0	25	7	2	0	.270	94	.486
16	17	6	1	3	17	6	1	3	.708	91	.512
17	29	7	0	10	29	9	0	0	.778	100	.470
18	23	9	0	0	22	8	0	0	.573	88	.578
19	13	5	1	4	13	5	1	4	-1.237	97	.513

20	36	7	2	2	36	7	2	2	1.122	98	.567
21	16	5	0	10	15	4	0	9	- .877	96	.479
22	22	4	2	0	22	4	2	0	- .460	85	.532
23	25	6	0	0	21	5	4	2	.073	98	.491
24	36	7	4	2	36	7	4	2	1.438	95	.516
25	27	5	2	5	27	5	2	5	.316	94	.533
26	24	8	0	5	18	7	0	5	.599	88	.527
27	30	11	0	3	29	10	0	3	1.651	89	.510
28	23	8	0	0	19	7	0	0	.527	84	.532
29	33	6	8	0	33	8	2	2	.944	92	.540
30	23	5	5	0	14	8	2	2	.131	95	.517
31	23	5	5	0	17	4	0	5	.131	98	.517
32	17	5	0	5	14	3	0	5	- .803	95	.479
33	16	3	0	0	16	7	2	2	1.336	92	.503
34	18	7	2	2	18	6	2	1	.346	80	.481
35	21	4	0	0	14	4	0	0	.744	88	.470
36	26	7	1	0	16	5	1	0	.766	89	.473
37	10	5	1	2	5	3	1	0	-1.365	98	.505
38	14	4	1	0	10	3	0	0	-1.104	99	.471
39	24	6	4	3	22	6	4	3	.216	98	.470
40	24	7	0	0	20	7	0	0	.305	97	.502
41	14	4	0	0	10	4	0	0	-1.092	88	.468
42	25	4	0	0	22	4	0	0	- .364	84	.555
43	21	5	1	1	19	4	1	1	- .645	92	.458
44	21	5	0	3	15	2	0	3	- .558	78	.449
45	35	7	3	0	33	7	3	0	1.377	96	.521

46	22	5	1	1	22	5	1	1	- .275	95	.512
47	28	5	0	1	20	5	0	1	.264	86	.571
48	25	8	2	0	24	8	2	0	.366	84	.514
49	22	5	1	0	18	5	0	0	- .161	90	.512
50	26	5	1	0	24	5	1	0	.387	94	.560
51	31	7	2	0	31	7	1	3	.948	98	.482
52	22	7	1	3	18	6	1	3	.257	97	.502
53	15	8	2	4	14	7	2	4	- .464	93	.605
54	10	3	0	0	7	2	0	0	-1.724	84	.471
55	2	7	0	10	16	6	0	0	.533	94	.880
56	26	7	1	3	23	7	1	3	.727	91	.470
57	8	4	1	0	8	4	1	0	-2.000	82	.611
58	14	6	1	3	4	3	0	0	- .708	84	.512
59	23	6	0	1	20	6	0	1	- .033	82	.492
60	5	2	0	0	5	2	0	0	-6.119	71	.641
61	5	2	0	0	5	2	0	0	-6.119	71	.641
62	28	6	1	0	25	6	1	0	.557	62	.500
63	8	2	0	0	8	2	0	0	-4.168	69	1.855
64	8	2	0	0	8	2	0	0	-4.168	60	1.855
65	8	2	0	0	8	2	0	0	-4.168	62	1.855
66	4	2	0	0	4	2	0	0	-6.119	50	.641
67	17	12	1	0	17	12	1	0	- .330	84	.759
68	21	6	0	10	21	6	0	10	- .679	51	.516
69	24	3	0	0	22	3	0	0	- .544	70	.630
70	9	4	0	0	9	4	0	0	-1.982	89	.601
71	33	8	0	0	33	8	0	0	1.313	89	.453

72	38	7	0	4	29	7	0	4	1.497	31	.548
73	10	3	0	8	10	3	0	8	-1.628	81	.471
74	28	10	4	7	23	8	4	7	1.521	69	.553
75	20	7	1	11	18	6	1	11	- .188	86	.514
76	25	3	0	0	25	3	0	0	- .544	73	.630
77	25	13	0	0	25	13	0	0	.825	63	.667
78	25	8	0	0	25	8	0	0	.527	77	.532
79	33	7	3	0	28	6	3	0	1.072	86	.454
80	26	9	0	2	25	9	0	2	1.054	56	.487
81	22	4	2	0	22	4	2	0	- .460	61	.532
82	15	4	1	0	15	4	1	0	-1.104	70	.471
83	22	5	0	12	22	5	0	12	-.147	77	.508
84	24	2	0	0	24	2	0	0	- .695	64	.751
85	27	8	0	0	26	8	0	0	.985	78	.468
86	24	13	0	0	16	16	0	0	.825	40	.667
87	22	6	0	0	21	6	0	0	.073	50	.491
88	10	6	0	4	6	3	0	0	-1.047	79	.585
89	27	5	0	0	27	5	0	0	.403	63	.553
90	20	6	1	0	20	6	1	0	- .362	76	.467
91	23	7	1	0	23	7	1	0	.292	89	.506
92	18	3	0	21	18	3	0	21	- .943	81	.534
M	22.0	6.2	1.1	2.2	20.3	5.8	92	2.0	-.256	83.3	.576
SD	8.3	2.6	1.6	3.7	8.2	2.6	1.3	3.6	1.653	14.3	.269

Table 2

Correlation between the number of links (c1), number of hierarchies (c2), number of cross-links (c3), number of examples (c4), IRT ability estimates (c5), and conventional test scores (c6) (n=92)

	c1	c2	c3	c4	c5
c2	0.562*				
c3	0.413*	0.243*			
c4	-0.064	0.025	-0.096		
c5	0.839*	0.733*	0.356*	0.097	
c6	0.130	0.077	0.315*	0.086	0.276*

*p<.05 (two tails)

Table 3

Item characteristics

	Links	Hierarchies	Cross-links	Examples
a	2.64	1.88	1.13	0.44
b(1)	-5.86	-6.18	- 14.03	- 12.91
b(2)	-1.89	-2.40	0.37	- 2.68
b(3)	-1.37	-1.45	1.02	0.03
b(4)	-0.73	-0.89	2.50	2.58
b(5)	-0.34	-0.24	18.07	4.85
b(6)	0.57	0.27	1.10	5.83
b(7)	1.17	0.95	- 9.98	7.35
b(8)	1.58	1.44	----	13.14
b(9)	2.39	1.84	----	- 10.72

Table 4

Correlation between students' ability estimates (c1), number of valid links (c2), number of valid hierarchies (c3), number of valid cross-links (c4), number of valid examples (c5), and total concept mapping scores (c6) (n=92)

	c1	c2	c3	c4	c5
c2	.800*				
c3	.648*	.538*			
c4	.336*	.412*	.147		
c5	.096	.029	.001	-.036	
c6	.764*	.800*	.753*	.696*	.131

*p < .05 (two tails)

Table 5

Correlation between students' conventional test scores (c1), number of valid links (c2), number of valid hierarchies (c3), number of valid cross-links (c4), number of valid examples (c5), and total concept mapping scores (c6) (n=92)

	c1	c2	c3	c4	c5
c2	.153				
c3	.009	.538*			
c4	.298*	.412*	.147		
c5	.067	.029	.001	-.036	
c6	.210*	.800*	.753*	.696*	.131

*p<.05 (two tails)

Table 6

Correlation between IRT ability estimates (c5), conventional test scores (c6), number of links (c1), number of hierarchies (c2), number of cross-links (c3), and number of examples (c4) for the enriched classes (n=59)

	c1	c2	c3	c4	c5
c2	.555*				
c3	.433*	.220			
c4	-.141	.138	-.165		
c5	.872*	.825*	.449*	-.001	
c6	.114	-.039	.187	.266*	.084

*p<.05 (two tails)

Table 7

Correlation between IRT ability estimates (c5), conventional test scores (c6), number of links (c1), number of hierarchies (c2), number of cross-links (c3), and number of examples (c4) for the regular class (n=26)

	c1	c2	c3	c4	c5
c2	.298				
c3	.188	.178			
c4	-.211	-.219	-.026		
c5	.791*	.732*	.301	-.232	
c6	-.368	-.186	.187	.095	-.344

*p<.05 (two tails)

Table 8

Correlation matrix between IRT ability estimates (c5), conventional test scores (c6), number of links (c1), number of hierarchies (c2), number of cross-links (c3), and number of examples (c4) for the adjusted class (n=7)

	c1	c2	c3	c4	c5
c2	.979*				
c3	.979*	1.0			
c4	.0	.0	.0		
c5	.915*	.817*	.817*	.0	
c6	-.057	-.92*	-.092	.0	-.088

*p<.05 (two tails)