LA-SUB-95-193

ASSESSING MIDDLE SCHOOL STUDENTS' UNDERSTANDING OF SCIENCE RELATIONSHIPS AND PROCESSES:
Year 2 - Instrument Validation

Final Report

by

Candace Schau, Professor, University of New Mexico;
Kirk Minnick, Minnick & Associates, Inc., Albuquerque;
Nancy Mattern, Doctoral Candidate, University of New Mexico;
Robert Weber, Research Professor, University of New Mexico

ASSESSING MIDDLE SCHOOL STUDENTS' UNDERSTANDING OF SCIENCE RELATIONSHIPS AND PROCESSES: Year 2 - Instrument Validation

Executive Summary

PURPOSE AND GOALS

- Our purpose for this multi-year project was to develop an alternative assessment format measuring rural middle school students' structural knowledge of science.
- During year 1, we identified and developed further a select-and-fill-in concept map format as a measure of structural knowledge.
- Our goal for year 2 was to examine the validity of our mapping format.
- During this project, we engaged in an essential collaboration with the TOPS Program.

BACKGROUND INFORMATION

- In our work, we utilized Messick's framework of seven categories describing possible sources of test validity evidence.
- As reported in the literature, many science assessments scores have shown consistent mean gender and ethnic group differences:
    - male students have shown small but statistically significant higher mean scores in comparison to female students beginning by the middle school years, and
    - White students have shown large and statistically significant higher mean scores in comparison to Black and Hispanic students beginning in elementary school.

INSTRUMENT DEVELOPMENT

- We developed a draft version of our select-and-fill-in concept measure assessing understanding in four major areas often covered in middle school science classes (life sciences, earth sciences, physical sciences, and scientific inquiry).
- We developed three comparison measures. These included:
    - a multiple-choice achievement measure consisting of items selected from the most recent National Assessment of Educational Progress,
    - a short attitude survey, and

DISCLAIMER

- questions asking students to compare the mapping and multiple-choice formats.

- We field-tested our draft achievement measures in two ways.
  - We validated the adequacy and coverage of the achievement measures using TOPS middle school teachers and University faculty.
  - We tested our measures on a pilot sample of seventh and eighth grade students in science classes taught by TOPS teachers.

INSTRUMENT VALIDATION METHODS

- We revised our achievement measures based on the field test results.
  - The final version of the map measure consisted of 37 select-and-fill-in nodes.
  - The final version of the multiple-choice measure contained 27 items.
- We collected data from 251 seventh grade and 117 eighth grade science students with the assistance of seven teachers (five of whom were in TOPS).
- For the seventh grade, we had large enough numbers of students to include three ethnic groups (Hispanic, Native American, and Anglo); we could not include Native Americans in our eighth grade analyses due to the small number in our sample.
- We balanced the order of administration of the achievement measures; half of the students in each class received the map measure first while half received the multiple-choice measure first.

GENERAL PERFORMANCE LEVELS: Achievement, Attitudes Toward Science, and Format Comparisons

- Seventh grade students answered correctly about half of the items on each achievement measure; eighth grade students answered correctly about two-thirds of them.
- On the multiple-choice items, seventh and eighth grade students scored above average as compared to the scores from a national norm group.
- Seventh grade students' attitudes were slightly positive and were more positive than those of a regional norm group; eighth grade students' attitudes were neutral but also were more positive than the norm group.
- Seventh grade students favored neither achievement format over the other, while eighth grade students slightly preferred the mapping format.

SUMMARY OF VALIDITY EVIDENCE

- The internal consistency of the map measure was excellent at both grade levels.
- Scores on the map measure showed both convergent validity (that is, the map measure assessed the same general science achievement construct - conceptual understanding - as the multiple-choice measure) and discriminant validity (the map measure also assessed unique aspects of achievement not assessed by the multiple-choice measure).
- Attitude scores were not related to map scores, indicating that the two measures assessed different constructs; attitude scores were related to multiple-choice scores.
- As expected from a valid achievement assessment, eighth grade students scored higher on the map measure than seventh grade students.
- We found patterns of mean gender and ethnic group differences on the map measure that were similar to those found on the multiple-choice measure, again supporting the validity of the map measure.
    - There were no gender differences on either measure in the seventh grade; males scored higher than females on both measures in the eighth grade.
    - Means scores on both achievement measures differed by ethnic group depending on task order administration.
        - In seventh grade, Anglo students scored higher than Native American and Hispanic students on both measures.
        - In eighth grade, there were no differences in mean map scores between Anglo and Hispanic students; Anglo students scored higher than Hispanic students on the multiple-choice measure.
        - In seventh grade, Hispanic students scored higher than Native American students on both measures when the multiple-choice measure was administered first; when the map measure was administered first, their scores were equal.
        - In the seventh grade, Native American students scored higher on both achievement measures when they took the map measure first.
        - Hispanic and Anglo students scored higher on the map measure when they took the multiple-choice measure first; order did not affect their multiple-choice scores.

CONCLUSIONS

- Our findings provide initial evidence that the select-and-fill-in concept map assessment format can be a valid measure of science achievement when used with rural ethnically-diverse middle school science students.
    - The internal consistency of the map measure was excellent
    - Scores on the map measure showed both convergent and discriminant validity.
    - Ethnic and gender group scoring patterns on the map measure tended to mirror those found on a multiple-choice measure, a format that has been used extensively to assess science achievement.
- The most interesting finding was the positive effect of taking the map measure first on seventh grade Native Americans.
    - This group scored higher on both measures when they had been exposed to the map measure first.
    - The difference between mean scores on both measures from Native Americans and Hispanic students disappeared when Native Americans took the map measure first.
- The TOPS program is in a unique position to contribute to work on assessment.
    - New Mexico has a large number of rural ethnically-diverse students.
    - The TOPS program is especially important because it targets rural middle-school science (and mathematics) teachers and their students.
- The project benefitted the TOPS program and participants in several ways.
    - TOPS may the only teacher enhancement program offered by the National Laboratories that contained an applied educational research component.
    - The project included TOPS participants in research based in a scientific discipline distinct from those usually represented at the Laboratories.
    - As contributors, they had critical input into the project.
    - The project gave TOPS teachers and students experience with additional types of alternative assessments.
- The map format provides another important type of measure to use in program and student evaluations and assessments and in educational research; however, additional validity work needs to be done with this assessment format.

ASSESSING MIDDLE SCHOOL STUDENTS' UNDERSTANDING OF SCIENCE RELATIONSHIPS AND PROCESSES: Year 2 - Instrument Validation

Our overall purpose for this multi-year project was to develop an alternative assessment format measuring rural middle school students' understanding of science concepts and processes and the interrelationships among them. This kind of understanding is called structural knowledge. We had 3 major interrelated goals:

1. Synthesize the existing literature and critically evaluate the actual and potential use of measures of structural knowledge in science education.

2. Develop a structural knowledge alternative assessment format.

3. Examine the validity of our structural knowledge format.

We accomplished the first two goals during year 1. The structural knowledge assessment we identified and developed further was a select-and-fill-in concept map format. The goal for our year 2 work was to begin to validate this assessment approach. This final report summarizes our year 2 work.

During the project, we engaged in an essential collaboration with the staff of the Los Alamos National Laboratory's TOPS Program and the Program's teachers and their students. We also worked with selected middle school science teachers from the TOPS program at Sandia National Laboratories.

The text of this report is organized into six sections. The first presents a brief summary of background information about test validity and gender and ethnic differences in science achievement. The second summarizes the development, field testing, and revision of our select-and-fill-in concept map measure, as well as the other measures used in our validity research. The third section describes the methods used in our research, including the final versions of our measures, the subjects, and data collection procedures. The fourth section presents a description and evaluation of the subjects' general levels of performance in terms of achievement and attitudes. The fifth section includes a summary of the validity evidence that resulted from our research. The sixth section presents our conclusions and suggestions for further work. Appendix A contains copies of the final versions of our assessment measures. Appendix B contains a description of the kinds of analysis techniques we used and more detailed statistical results.

Background Information

In addition to the literature reviewed in our Year 1 report, two important sets of literature contributed to our Year 2 work. First, we attended carefully to current developments in test validity theory and research. Second, we also examined theory and research concerning patterns of mean gender and ethnic group score differences from

traditional and alternative assessment achievement formats.

Test Validity

Traditional ideas about test validity are expanding with the advent of increased emphasis on, and corresponding use of, alternative assessments for educational measurement. Part of this expansion includes inextricably connecting the test with the students who take it. This connection demands increased scrutiny of what test scores mean and the consequences of their use with various learner groups.

Messick (1992), one of the acknowledged theoretical experts in test validity, recently wrote about validity as "empirical evaluation of the meaning and consequences of measurement" (p. 1487). He presented a framework of seven categories that described possible sources of validity evidence (p. 1488). As he said, there are other ways to categorize these evidence sources, but we found his framework useful. We have applied variations of these categories to our validation work; underlined phrases are used to refer to the source throughout the rest of this report.

1. "relevance and representativeness" of the map measure's science content,

2. internal structure of the map measure (relationships among the responses to the fill-in nodes),

3. external structure of the map measure (relationships of map scores to scores on other types of assessments such as those measuring achievement and science attitudes),

4. cognitive processes underlying performance on the map measure,

5. generalizability of the use of our map measure across gender and ethnic groups,

6. expected patterns of variations in map test scores due to increased exposure to science instruction and in comparison to patterns associated with multiple-choice assessments,

7. "value implications and social consequences" associated with using the map measure.

We designed our Year 2 validation research to yield evidence that would fit into six of these categories. Our Year 1 development work with groups of students and with individual students provided evidence regarding some of the cognitive processes underlying performance on the map measure (source #4). Our work is the beginning of a validation process of the use of the fill-in map format with a variety of learners in varied educational contexts.

Gender and Ethnic Differences in Science Achievement

Many standardized achievement tests and alternative assessments show consistent mean gender and ethnic group differences. The National Assessment of Educational Progress (NAEP) results are typical of the differences found on many standardized measures of science achievement. A NAEP science assessment was last administered in 1990 between January and mid-May to stratified national samples of over 6500 students in the fourth, eighth, and twelfth grades. The eighth grade sample was 71% White, 15% Black, 10% Hispanic, 3% Asian, and 1% Native American. Male students showed small but statistically significant higher mean scores in comparison to female students in grades 8 and 12 within each ethnic group; gender differences were larger at the twelfth grade. White students, on the average, scored higher than Black students and than Hispanic students at all three grade levels and for each gender. These achievement gaps by ethnic group increased with increasing grade level. Even with these mean differences, however, students from both genders and from each ethnic group were among the highest and lowest achievers on NAEP. Although Native American students were included in the assessment, their number were too low to yield reliable results. See Jones, Mullis, Raizen, Weiss, and Weston (1992) for more information.

Instrument Development

This section summarizes our work in development of the draft versions of the measures and their field testing.

Draft Formats

As part of our validity research, we wanted to compare students' performance on a traditional measure of science achievement with their performance on our fill-in concept maps. We selected multiple-choice items to comprise the traditional measure since this format has been used for decades to measure science achievement in large-scale standardized assessments. Good multiple-choice items measure understanding of, although not usually interrelationships among, science concepts. We initially designed drafts of both of our achievement measures such that each assessed understanding in the same four major areas often covered in middle school science classes (life sciences, earth sciences, physical sciences, and scientific inquiry).

The first draft of the multiple-choice measure consisted of 40 items selected from the items released to the public that were used in the 1990 National Assessment of Educational Progress for students in the eighth grade. Students chose the best answer from a set of four options given for each item. NAEP items were used because these items were carefully constructed with input from relevant groups and materials from across the U.S. and because difficulty levels were available based on the national sample to whom these items were administered.

We constructed four concept maps, one to cover aspects of these same four areas. The life sciences were represented by a map about plants, the physical sciences by a map about energy, the earth sciences by a map about the earth, and scientific inquiry by a map about the nature of scientific knowledge.

We created the draft select-and-fill-in concept maps by removing 38 nonconsecutive concepts from these maps; these words or phrases were placed in a corner box of the appropriate map. Students completed this assessment format by selecting a response from the box and writing it in the blank. About half of these fill-in nodes assessed connected understanding of the same concepts assessed in half of the multiple-choice items.

We asked students to compare the two assessment formats in terms of five characteristics. These included effort, interest, level of performance, accuracy as measures of achievement, and liking. Students responded on a 3-point scale by selecting either maps or multiple-choice or both.

We included ten Likert scale items designed to measure selected aspects of students' attitudes toward science including, for example, self-efficacy, affect, and value. The 5-point response scale ranged from "Strongly Disagree" through "Neither" to "Strongly Agree." Most of these items were selected from the Longitudinal Study of American Youth. They had been used as part of the UCAN-RSI survey designed to identify the science attitudes of eighth grade students in Arizona, New Mexico, and Colorado. These responses provided a norm group to use for comparison purposes in our study. We could not include more attitude questions due to the need to complete data collection from each class within one class period.

Pilot Field Tests

We engaged in two kinds of pilot field testing. First, we validated the adequacy and coverage of the four concept maps using both teachers and discipline experts. Five TOPS middle school teachers evaluated all four maps. Four University faculty evaluated the maps that assessed disciplines in which they were expert; each map was evaluated by two different faculty members, with some faculty evaluating more than one map.

Second, and concurrently, we pilot tested the two measures on 63 seventh and eighth grade students in four science classes taught by two TOPS teachers, one from the LANL program and one from the Sandia program. About half of the students were female and half were male; similarly, Native American, Hispanic American, and Anglo American students were about equally represented. Although almost all of the students seemed to understand the tasks and appeared capable of responding to them, it was difficult for them to complete all measures in some of the classes which had shorter class periods.

Instrument Validation Methods

This section describes the final version of our measures, the teacher and student sample, and data collection procedures. We also report general levels of performance on the achievement measures and on the attitude measure.

Final Measures

The draft versions of the two achievement measures were revised a number of times

based on the results of the two kinds of field testing. These revisions primarily involved redrawing parts of the concept maps and eliminating multiple-choice items. The final version of the map measure included 37 select-and-fill-in nodes; half (19) of these assessed structural knowledge of concepts found on the multiple-choice assessment. The final version of the multiple-choice measure consisted of 27 items; 70% (19) of these tested concepts that were also assessed in the map measure. Using responses from our pilot sample of students, we analyzed the internal consistency of these scores using Cronbach's alpha, the traditional reliability index used for this purpose. When each item in a measure contributes appropriately to the total score, the alpha value will be large and positive (with a perfect, and unattainable, value of 1.0). Values of Cronbach's alpha were .94 for the map measure and .85 for the multiple-choice measure. We were unable to evaluate the internal consistency of the draft attitude measure because many of the pilot test students were unable to complete it due to time constraints. The final version of our total assessment included the two kinds of achievement formats as well as the ten-item attitude measure and the five comparison questions. See Appendix A.

Each set of achievement items was contained in a test packet. For the multiple-choice items, students read each item and marked their response on a machine readable answer sheet. Students completed the map measure by writing directly in the map test packet. We later coded their responses and added them to the student's answer sheet. The attitude and format comparison items and space for responses were contained on the answer sheet.

Student responses to the multiple-choice items and to the fill-in map nodes were scored as correct or incorrect, with omitted responses scored as incorrect. Students then were given a total correct score on each measure. We formed an attitude score for each student by reversing responses to negatively-worded items and averaging across all responses. Higher scores meant more positive attitudes.

Subjects

We collected data with the assistance of four LANL TOPS teachers (from El Pueblo, Clayton, Cuba, and Gallup) and one Sandia TOPS teacher (from Las Lunas) who taught middle school science. Two of the LANL TOPS teachers (from Clayton and Gallup) each arranged data collection with a colleague who also taught middle school science in that school. We collected data from these 7 teachers and their 251 seventh and 117 eighth grade students.

About half of the seventh grade students were female and half male. Half of these students reported that they were Hispanic, with about one-quarter Native American and one-quarter Anglo. One student was Black, and one chose "Other."

About two-thirds of the eighth grade students were female and one-third male. About 45% of these students reported that they were Hispanic and another 45% Anglo. Nine percent were Native American, and two students chose "Other."

Procedures

Two of us collected data in each classroom. We balanced the order of administration of the achievement measures. We randomly split each class in half; about half of the students completed the multiple-choice measure first followed by the fill-in concept map measure while the other half received the measures in the reverse order. The multiple-choice measure was shorter than the fill-in concept map measure; its format was also more familiar to the students. Thus, most students completed it more quickly than the fill-in concept map measure. To balance the time requirements, students completed the attitude items after finishing the multiple-choice items. They finished the format comparison items last. The response form also asked them to report grade level, class period, gender, and ethnicity (Anglo/White, Hispanic, American Indian, African American, Other).

General Performance Levels

We examined the general performance levels of all students who filled in at least 50% of the responses within a measure. Since most (but not all) of the students involved in this work were in TOPS teachers' classrooms, this section provides information that could be used as preliminary evidence of the impact of this Program. In addition, this information presents a global picture of these students' achievement levels and attitudes toward science.

We used two methods whenever possible, First, we looked at mean scores from each measure. Second, we compared the multiple-choice and attitude responses to norm group data. Because we developed the map measure and the comparison items, we had no norm data for them. For the multiple-choice items, we estimated the percentage of NAEP students who would have correctly answered each multiple-choice item if the ethnic breakdown of NAEP's sample had matched ours. Because our ethnic proportions varied in seventh and eighth grade, we estimated these percentages twice, once for each grade level in our study. We then compared the percentage of students in our study to the percentage of students in the weighted NAEP norm group who correctly answered each item. See Table 1. For the attitude items, we compared the percentage of students in the study to the percentage of UCAN-RSI students who expressed positive attitudes by responding "Agree" or "Strongly Agree" to positively-worded items and "Disagree" or "Strongly Disagree" to negatively-worded items. See Table 2. In all percentage comparisons, the study sample percentages were considered to be different from the norm group percentages if they differed by more than 2%.

Achievement

The seventh grade students received an average total map score of 53% correct (mean of 19.5 points with a standard deviation of 8.1 points); the eighth grade students' mean score was higher at 64% correct (24 points with a standard deviation of 8). Map 4 (on energy) was the most difficult map for both the seventh and eighth graders. The other three maps were of about equal difficulty within each grade level. See Table 3.

Table 1. Percentage of students selecting correct response on multiple-choice items

	Weighted NAEP - 7th Grade Comparison	7th Graders	Weighted NAEP - 8th Grade Comparison	8th Graders
Animals must breathe oxygen (d)	93	97	94	94
Liquid expands when heated (c)	47	51	53	61
Light raises temperature of objects (b)	52	44	56	58
NE wind blows toward SW (c)	26	25	30	37
Space taken by rocks in container (c)	35	43	37	32
Experiments reliable with controls (c)	38	41	40	58
Opinion that flowers are beautiful (b)	75	54	79	74
Cells evidence of life (a)	79	78	80	89
Flowers attract insects & make seeds (d)	26	38	31	56
Leaf makes plant food (b)	20	21	22	31
Dust caused extinction of dinosaurs (a)	44	45	47	66
Running water wears earth surface (a)	47	50	51	39
# of plants to use in exp. design (d)	20	30	24	37
Mineral not major rock type (d)	57	44	59	54
Statement described hypothesis (a)	31	45	34	35
Hypotheses ideas that can be tested (a)	52	66	58	82
Force is a push or a pull (d)	74	81	77	92
Water freezes at 0 degrees Celsius (a)	51	46	52	69
Folded mountains as in rockies (a)	38	37	42	43
Sci. know. based on exp. & obser. (a)	66	67	70	75
Earth temp decreases core to crust (d)	32	36	37	35
Seismograph measures earthquakes (c)	77	74	80	71
Different ways to classify fossils (d)	67	72	71	73
Oceans contain most earth's water (b)	87	72	87	89
Knowledge changes with fossils (b)	73	60	77	69
Sun main energy source for photo.(c)	74	67	77	70
Grav. attraction increases with mass (b)	33	35	35	41
Seasons change by tilt of earth axis (a)	57	44	64	65

Table 2. Percentage of students responding with positive attitudes on each item

	Norm	Seventh	Eighth
I enjoy science+	43	71	62
Doing science makes me nervous-	49	70	49
I am good at science+	38	51	41
I usually understand what we are doing+	53	79	66
Learning science is memorizing-	38	35	27
I worry about how well I do on science test-	24	18	17
Almost all people use science in their jobs+	30	58	54
I will use science in many ways as an adult+	47	61	49
I am encouraged to ask questions in science+	52	58	57
Other people learn science better than I do-	NI	32	26

+ = positively worded item: % "Agree" and "Strongly Agree" reported
- = negatively worded item: % "Disagree" and "Strongly Disagree" reported
NI = not included on UCAN-RSI survey

Table 3. Mean percent correct map scores

	Map 1 (Plants)	Map 2 (Earth)	Map 3 (Science)	Map 4 (Energy)	Total
Seventh Grade	60	66	57	32	53
Eighth Grade	67	75	63	51	64

The multiple-choice measure was of the same difficulty as the total fill-in map measure. For seventh grade students, the average total score was 53% correct (13.71 points with a standard deviation of 4). For eighth grade students, the average total score was 63% percent correct (16.3 points with a standard deviation of 5.0).

The item by item responses of the seventh and eighth grade students were compared to the appropriately weighted NAEP norm (see Table 1). The percentages of seventh grade students correctly answering items were greater than the norm (by more than 2%) on 43% of the items. The percentages were equal to the norm on 25% of the items and were below the norm on 32% of the items. On the average multiple-choice item, the percentage of seventh grade students correctly answering was equal to the norm percentage. Overall, the seventh grade students scored slightly above average in comparison to the eighth grade norm group students from comparable ethnic groups.

The percentages of eighth grade students correctly answering items were greater than the weighted NAEP norm on 46% of the items, were equal to the norm on 29% of the items, and were less than the norm on 25%. On the average multiple-choice item, the percentage of eighth grade students correctly answering was over 5% higher than the norm percentage. Overall, the eighth grade students were above average in comparison to the scores of the NAEP norm group students from comparable ethnic groups.

Attitudes Toward Science

Seventh grade students expressed attitudes toward science that were slightly positive overall as evidenced in the total attitude score ($\underline{M}$ = 3.4, $\underline{SD}$ = .5, where "3" was neutral). The majority of seventh grade students responded with positive attitudes to 70% of the attitude items (see Table 2). They were more positive than the norm group (by at least 2%) on 78% (7/9) and more negative on 22% (2) of the items.

The eighth grade students ($\underline{M}$ = 3.1, $\underline{SD}$ = .7) were less positive than the seventh grade students; their overall attitudes were neutral. The majority of eighth grade students responded positively to less than half of the items (40%). However, they were more positive than the norm group on 56% (5/9), equal to the norm group on 22% (2), and more negative than the norm on 22% (2) of the items.

Format Comparisons

We examined average student scores summed across the five format comparison items. The seventh grade students tended to be neutral ($\underline{M}$ = 2.0, $\underline{SD}$ = .5, where "2" represented a choice of both formats, "1" the map measure, and "3" the multiple-choice measure); that is, these students favored neither format over the other. Eighth grade students slightly preferred the mapping format ($\underline{M}$ = 1.8, $\underline{SD}$ = .5).

Summary of Validity Evidence

In this section, we have grouped our findings into the most appropriate of Messick's categories of validity evidence. We collected initial evidence concerning the cognitive processes underlying performance on the map measure (source #4) during our Year 1 development work; those results were summarized in our Year 1 final report. See Appendix B for more information about the analysis techniques used to generate our findings and for more detailed statistical results.

Relevance and Representativeness (Source #1):

As described in the "Draft Format" section, we initially developed our maps to assess aspects of four major discipline areas often covered in middle school science classes. About half of the fill-in concepts included in the maps matched concepts assessed in the multiple-choice questions that served as the comparison assessment measure. Through their development and testing process, those multiple-choice questions represented a national view of important concepts in middle school science.

In addition, as described in the "Pilot Field Test" section, we validated the adequacy and coverage of the four concept maps using both middle school teachers and University faculty who were discipline experts. In general, their feedback suggested that few revisions to the maps were needed. We redrew those parts of the maps to match their suggestions, whenever possible.

Internal Structure (Source #2): Internal Consistency

The internal structure of the map measure was assessed using Cronbach's alpha. The internal consistency of the map measure was excellent at both grade levels (seventh: alpha = .90; eighth: alpha = .92).

Because we related map scores to scores on the multiple-choice and the attitude measures, we also examined their internal consistencies. The internal consistency of the multiple-choice measure at both grade levels was acceptable (seventh: alpha = .74; eighth: alpha = .82). For the attitude measure, the internal consistency was too low for the seventh grade (alpha = .59) but acceptable at the eighth grade (alpha = .76). It was not clear why the seventh grade value was lower.

We did not examine the interval consistency of the comparison total score. Internal consistency values for scales with only five items often are poor. Because we were able to ask what we needed to know in five items, we did not try to enlarge that scale. We decided to look at relationships using a total score, but we realized that we might have to look at individual item results.

External Structure (Source #3): Relationship Between Concept Map Scores and Multiple-Choice, Attitude, and Format Comparison Scores

We wanted to show that the map measure when used with rural middle school science

students possessed both convergent and discriminant validity. The map measure was designed to assess understanding of science concepts and their interrelationships. If it possessed convergent validity, students' score distribution on the map measure should have shown a positive and at least moderately-sized relationship with their score distribution on a good multiple-choice measure, a format that has been used extensively to assess conceptual understanding of science. However, multiple-choice questions usually do not assess structural knowledge. If the map measure also possessed discriminant validity, the size of this relationship should not have been too large, indicating that the map measure also assessed unique aspects of achievement not tested by the multiple-choice measure.

This relationship pattern was found. The relationship between total correct scores on the two achievement measures was significant, quite strong, and positive for students at both grade levels. For seventh grade students, the two score distributions shared about half (53%) of their variability. For eighth grade students, the distributions shared about 60% of their variability.

Because the map measure was designed to assess science achievement and not attitudes toward science, the relationship between score distributions on these two measures should be, and in fact was, very small. Attitude scores and map scores were not related at either grade level. Interestingly, however, students in both grade levels with more positive attitude scores also tended to have higher multiple-choice scores. For seventh grade, this relationship was small, with only about 3% of their score variability shared. For eighth grade, this relationship was moderate, with 10% of the variability shared. These multiple-choice achievement questions also seemed to assess attitudes, at least to a small degree.

For both grade levels, students with higher scores on both achievement measures favored the mapping format. These relationships were moderate in size, with shared variance percentages ranging from 5% to 14%.

Generalizability (Source #5) and Value Implications and Social Consequences (Source #7): Gender and Ethnic Relationships

If we are to use the map measure with both female and male students from multiple ethnic groups, we need to understand whether mean score differences exist among these groups. Using Analysis of Variance techniques, we tested for possible gender and ethnic relationships. For the seventh grade, we had large enough numbers of subjects to examine three ethnic groups (Hispanic, Native American, and Anglo) and to test all possible interactions. Our eighth grade sample was smaller. We could not include Native Americans due to the small number in our sample nor could we test the third-order interaction. Unless otherwise noted, all gender and ethnic effects presented are statistically significant.

For seventh grade, mean map scores varied significantly and differentially by ethnic group depending on whether the students completed the map measure or the multiple-choice measure first. See Table 4 for these means and Figure 1 for a plot of the interaction. We first examined mean differences associated with task order for each ethnic group of students.

Table 4. Unweighted percent (and number) correct cell means for significant seventh grade ethnic by order interactions on both achievement measures

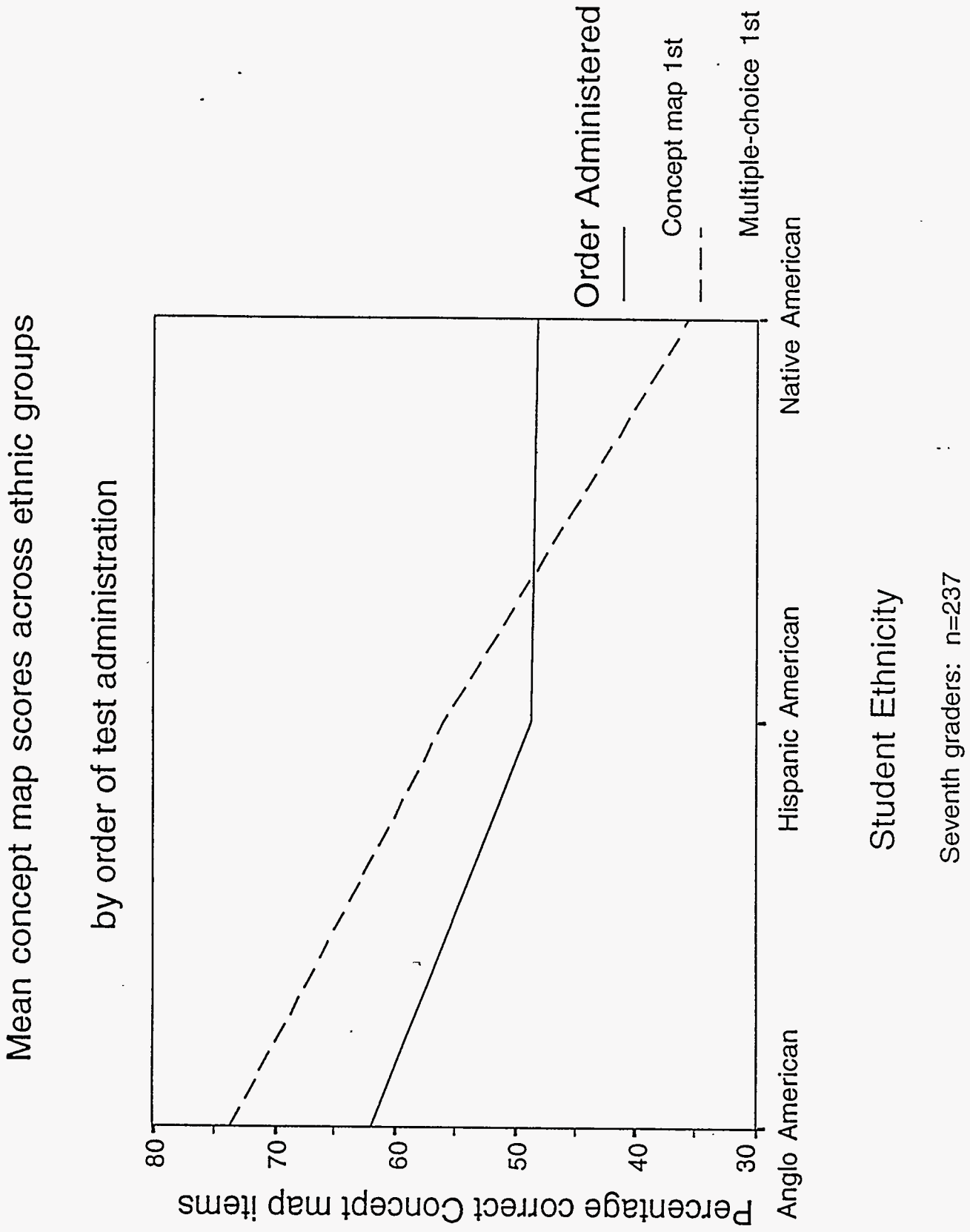
Score - Maps (maximum score = 37 points):

	Anglo	Hispanic	Native American
Maps first	62% 22.92 n=31	49% 18.06 n=54	48% 17.84 n=32
Multiple-choice first	74% 27.28 n=21	56% 20.67 n=65	35% 13.09 n=34

Score - Multiple-Choice (maximum = 26 points; cell sizes same as above):

	Anglo	Hispanic	Native American
Maps first	64% 16.66	51% 13.20	51% 13.25
Multiple-choice first	60% 15.63	53% 13.85	41% 10.75

Figure 1. Plot of significant ethnic group by task order interaction of concept map scores



When students completed the map measure first, Anglo students' average percent correct map score was 12% lower than when they completed the multiple-choice questions first. Similarly, Hispanic students' mean score also was lower (by 7%). However, Native American students' average map score was 13% higher when they completed the map measure first. That is, taking the map measure before the multiple-choice measure helped Native American students score higher on the map measure while it resulted in lower mean map scores for Hispanic and Anglo students. The latter two groups, especially the Anglo students, learned science knowledge as they completed the multiple-choice measure that helped them boost their performance on the subsequent map measure.

We then looked at ethnic group differences within each task order. Mean map scores for Anglo students were significantly higher than those for Hispanic students (map first - by 13%; multiple-choice first - by 18%) and Native American students (map first - by 14%; multiple-choice first - by 39%) regardless of which measure was completed first. However, mean map scores for Hispanic students were greater than those of Native American students (by 21%) only when they completed the multiple-choice measure first; their means were about equal when they completed the map measure first.

On the map measure, gender was not a significant effect for the seventh grade students.

For eighth grade students, we found no ethnic nor order effects (although they were close to significance) on the map measure. Gender, however, was significant and showed the traditional pattern of higher male performance, male mean = 74% correct (27.5 points) and female mean = 64% correct (23.7).

Expected Patterns of Variations (Source #6): Grade Level Effects and Comparison of Effects Between Achievement Measures

If the map measure was a valid assessment of science achievement, we expected to find certain patterns of variations in mean scores. One of these concerned grade level. Eighth grade students have been exposed to science instruction for one year more than have seventh grade students. We expected that eighth grade students would score higher, on the average, than seventh grade students on the map measure. We also expected the map scores to exhibit grade effects and gender and ethnic group patterns similar to those found in the multiple-choice scores.

Grade Level Effects

Eighth grade students scored higher than seventh grade students on the map measure at the total score level, at the map level, and at the individual item level. The total percent correct score of eighth grade students (64%) was over 10% higher than that of seventh grade students (53%). As Table 3 shows, eighth grade students also scored higher on each map. In addition, a greater percentage of eighth grade students answered correctly most of the nodes (92%, 34 of 37). See Table 5.

Table 5. Percentage of students selecting correct node response on maps measure

	7th Grade	8th Grade
MAP 1:		
Node 1 - Sunlight (8)	82	89
Node 2 - Insect (9)	75	77
Node 3 - Seed (7)	68	73
Node 4 - Leaf (3)	42	47
Node 5 - Carbon Dioxide (1)	54	59
Node 6 Oxygen (6)	69	73
Node 7 - Stem (2)	62	75
Node 8 - Anchor (4)	38	47
Node 9 - Water (5)	67	75
Map 2:		
Node 1 - Milky Way (8)	75	76
Node 2 - Sun (4)	76	80
Node 3 - Water Vapor (2)	60	72
Node 4 - Granite (3)	43	63
Node 5 - Oceans (6)	83	91
Node 6 - Land Surface (5)	61	78
Node 7 - Core (7)	68	77
Node 8 - Metamorphic (10)	42	65
Node 9 - Mountain Folding (1)	76	72
Node 10 - Volcanoes (9)	71	73
Map 3:		
Node 1 - Seismographs (6)	75	77
Node 2 - Thermometers (2)	77	74
Node 3 - New Evidence (1)	57	66
Node 4 - Smell (7)	79	84
Node 5 - Opinions (8)	42	60
Node 6 - Controls (3)	48	58
Node 7 - Repeatable (4)	38	34
Node 8 - Tested (5)	37	52
Map 4:		
Node 1 - Conserved (11)	30	48
Node 2 - Force (3)	68	74
Node 3 - Radioactivity (1)	18	26
Node 4 - Chemical (4)	32	47
Node 5 - Mechanical (8)	13	29
Node 6 - Falling Object (9)	27	43
Node 7 - Mass (7)	25	48
Node 8 - Light (5)	31	43
Node 9 - Expansion (6)	30	55
Node 10 - Solid (2)	43	79
Node 11 - Gas (10)	39	69

Comparison of Effects Between Achievement Measures

This same grade-level pattern occurred on the multiple-choice measure. The eighth grade students' total percent correct score (63%) was 10% higher than that of the seventh grade students (53%). A higher percentage of eighth grade than seventh grade students answered correctly the majority of multiple-choice items (71% or 20/28 items). See Table 1.

As occurred for seventh grade students on the map measure, their mean multiple-choice scores also varied differentially by ethnic group depending on task administration order. See Table 4 for these means and Figure 2 for a plot of the interaction. Unlike map scores, task order did not affect the mean multiple-choice scores of Anglo and Hispanic students. For Native American students, taking the map measure first again had a positive effect on their multiple-choice scores; their mean multiple-choice score was 10% higher when they took the map measure first.

The pattern of ethnic group differences in mean map scores for seventh grade students mirrored that found for multiple-choice scores. On the multiple-choice measure, Anglo students scored higher than Hispanic students (maps first - by 13%; multiple-choice first - by 7%) and than Native American students (maps first - by 13%; multiple-choice first - by 19%). Hispanic students scored higher than Native American students (by 12%) only when the multiple-choice measure was administered first.

The ethnic group pattern for eighth grade students varied between the map and multiple-choice measures. Ethnicity was not a significant effect for the map measure, while it was significant for the multiple-choice measure. For the eighth grade, Anglo students scored higher than Hispanic students on the multiple-choice measure, Anglo mean = 71% (18.5) and Hispanic mean = 62% (16.0).

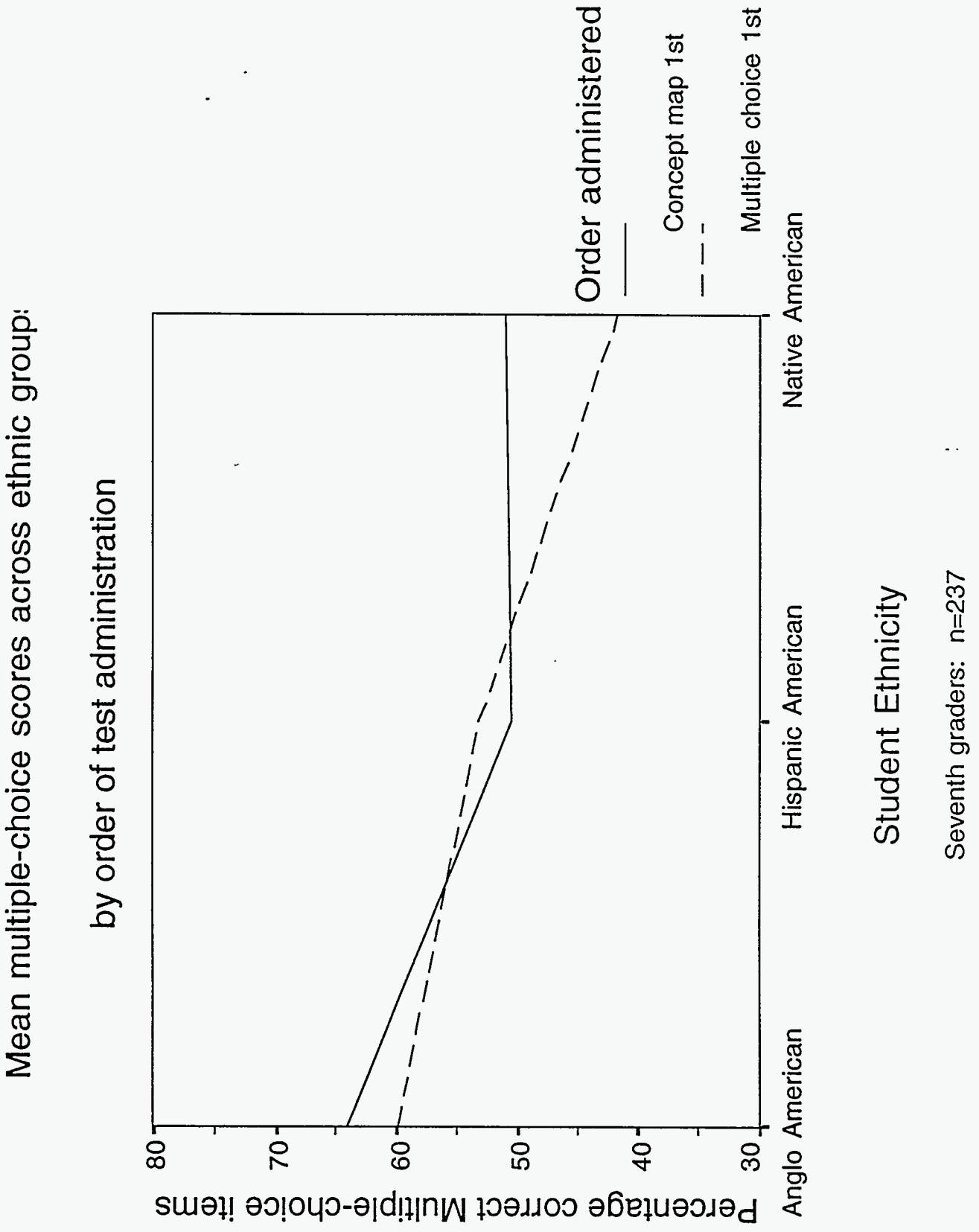
The gender similarities and differences with the map scores were consistent with those found for the multiple-choice scores. Gender was not a significant effect on either measure for seventh grade students. For the eighth grade students, gender was a significant effect on each achievement measure. On the average, males again scored higher than females on the multiple-choice measure, male mean = 72% (18.6 points), female mean = 61% (15.9).

Conclusions

Our findings provide initial evidence that the select-and-fill-in concept map assessment format can be a valid measure of science achievement when used with rural ethnically-diverse middle school science students. Our development and field testing processes yielded evidence of the relevance and representativeness of our map measure. The internal structure of the map measure as indicated by its internal consistency was excellent at both grade levels, indicating that the fill-in nodes work well together in producing a total map score.

Evidence related to the external structure of the map measure suggested that it possessed both convergent validity and discriminant validity for these students. Scores on the

Figure 2. Plot of significant ethnic group by task order interaction of multiple-choice scores



map measure shared significant amounts of variability with scores on the multiple-choice measure at both grade levels (evidence of convergent validity). These relationships were higher than expected, probably due to the explicit overlap in concepts assessed in the two achievement measures. Even so, the score distributions also did not share important amounts of variability (evidence of discriminant validity). Also as desired, map scores were not related to attitude scores.

We found scoring patterns on the map measure that indicated its validity as a science achievement assessment when used with these students. Eighth grade students scored higher than seventh grade students, as would be expected in a good measure of science achievement. We found a pattern of ethnic group mean differences (mediated by task order administration in seventh grade) on the map measure similar to that found on the multiple-choice measure. We found the identical pattern of mean gender similarities (in the seventh grade) and differences (in the eighth grade), again supporting the validity of the map measure.

Since previous research using these same multiple-choice items has shown ethnic group differences, the ethnic group differences we found for the seventh grade students on those items were not surprising. However, the seventh grade ethnic group differences found on the map measure were unwanted. It is surprising that we found ethnic group differences on the map measure at only one grade level; we expected to find these differences either at both grade levels (as occurred with the multiple-choice measure) or at neither. We had thought that the visual aspect of the maps would lessen the high verbal cognitive load associated with reading items on multiple-choice tests and so would help the ethnic groups that score traditionally lower. Unfortunately, this occurred only in eighth grade. This research only emphasizes that the ethnic group differences in achievement in our country and our state must be addressed and minimized.

The most interesting finding, however, was the positive effect of taking the map measure first for Native American students. This group scored higher on both measures when they had been exposed to the map measure first. In fact, the difference between mean scores from this ethnic group and Hispanic students disappeared on both measures when Native Americans took the map measure first. This finding clearly needs replication. However, if it can be replicated, it suggests an interesting approach for teaching and test preparation for Native American students. Teachers can prepare fill-in concept maps over the same discipline areas covered on the standardized test (not the exact same content, of course) and use those to prepare Native American students. It was unfortunate that we could not include this ethnic group in the eighth grade analyses; it would have been interesting and important to determine if this order effect occurred for eighth grade Native Americans also.

Other researchers have not attended to possible order effects in their research. In some of this work, large time periods occurred between administrations of the various kinds of tasks; in others, the time periods are shorter or are not clear. Our work showed clear and large immediate order effects that differentially affected ethnic groups. Other researchers need to attend to this possible effect in their work.

Overall, seventh grade students favored neither assessment format over the other, while eighth grade students favored slightly the mapping format. This finding is important in indicating that students at least considered the mapping format as no worse to take than the traditional measure. However, students who scored higher on either achievement measure preferred the mapping assessment. This relationship may indicate that the higher achieving students welcomed the challenge of a new assessment format while lower achieving students did not.

In New Mexico, we are in a unique position to contribute to work on assessment. We have a large number of rural ethnically-diverse students. The TOPS program is especially important because it targets rural middle-school science (and mathematics) teachers and their students. In this project, we collaborated with the staff of the Los Alamos National Laboratory's TOPS Program and the Program's teachers and their students. We also worked with selected middle school science teachers from the TOPS program at Sandia National Laboratories. This collaboration was essential to the success of this project.

Our project benefitted the TOPS program and participants in several ways. TOPS may be the only teacher enhancement program offered by the National Laboratories that contained an applied educational research component. Our project included TOPS teachers and their students in research that is based in a scientific discipline distinct from disciplines usually represented at the Laboratories. As research contributors, they had critical input into the design, implementation, and evaluation of our project. In addition, our project emphasized assessment. Because of this emphasis, teachers and students also gained experience with additional types of alternative assessments.

The map format provides another important type of measure to use in program and student evaluations and assessments and in educational research. However, additional validation work needs to be done with this assessment format. We would like to demonstrate that scores on the mapping measure for rural middle school students are at least moderately related to scores on existing research measures of structural knowledge. Unfortunately, the most accepted measure of structural knowledge (relatedness ratings) has not been used below college level and does not appear to be appropriate for younger students. We have demonstrated such a relationship between our mapping measure and related ratings using astronomy concepts with postsecondary students enrolled in an introductory undergraduate astronomy course. We should continue to investigate the order effect by replicating the outcomes associated with completing the mapping measure prior to a multiple-choice measure with other seventh grade students. We should extend our work to other grade levels. We need to explore further the ethnic group differences on the mapping measure and to develop educational approaches that begin to equalize these differences. We also need to show improved performance on the mapping measure for the same students across time.

References

Jones, L. R., Mullis, I. V. S., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). The 1990 Science Report Card: NAEP's Assessment of Fourth, Eighth, and Twelfth Graders. U.S. Department of Education.

Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (ed.), Encyclopedia of Educational Research, 6th ed. (pp. 1487-1495). NY: Macmillan.

Appendix A

Assessments

NAME _____

PLEASE USE NO. 2 PENCIL

RIGHT	WRONG
● ■ ■	✓ ⊘ ⊘ ⊘

What class period in this class?
① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨

What grade are you in? ⑩ ⑪

ETHNICITY
Anglo/White ○
Hispanic ○
American Indian ○
African American ○
Other ○

GENDER
Female ○
Male ○

How much time do you spend on science homework each school day?
Homework not usually assigned ○
Have homework but don't do it ○
1/2 hour or less ○
1 hour ○
2 or more hours ○

What grade do you expect to receive in this class? Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ

SCIENCE QUESTIONS

Example Ⓐ Ⓑ Ⓒ Ⓓ

	A	B	C	D		A	B	C	D		A	B	C	D		A	B	C	D
1.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	8.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	15.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	22.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
2.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	9.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	16.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	23.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
3.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	10.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	17.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	24.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
4.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	11.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	18.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	25.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
5.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	12.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	19.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	26.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
6.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	13.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	20.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	27.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
7.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	14.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	21.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	28.	Ⓐ	Ⓑ	Ⓒ	Ⓓ

STUDENT SURVEY

Please answer the following by filling in only one oval for each question.

	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
I enjoy science.	○	○	○	○	○
Doing science often makes me nervous or upset.	○	○	○	○	○
I am good at science.	○	○	○	○	○
I usually understand what we are doing in science.	○	○	○	○	○
Learning science is mostly memorizing.	○	○	○	○	○
I worry about how well I will do on science tests.	○	○	○	○	○
Almost all people use science in their jobs.	○	○	○	○	○
I will use science in many ways as an adult.	○	○	○	○	○
I am encouraged to ask questions in science class.	○	○	○	○	○
Other people can learn science better than I can.	○	○	○	○	○

SCIENCE CONCEPT MAPS

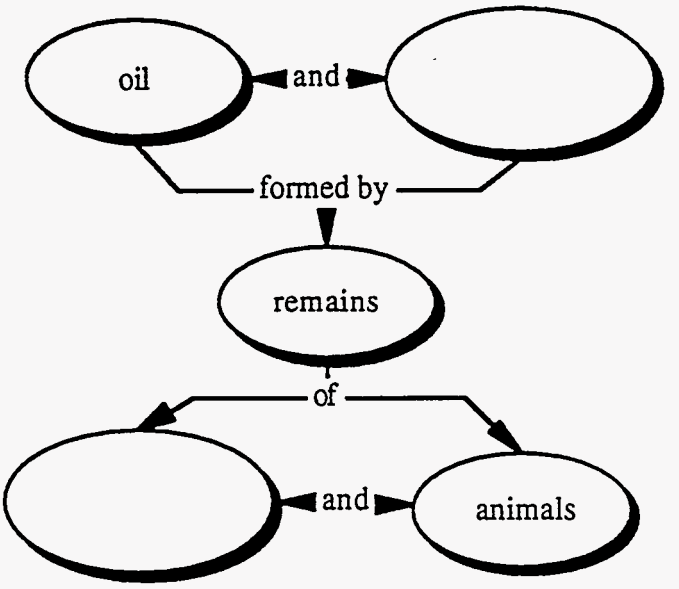
NAME:_____ PERIOD:_____

DIRECTIONS: PRINT YOUR NAME AND CLASS PERIOD ON THE TOP OF THIS BOOKLET.

THERE ARE FOUR CONCEPT MAPS ABOUT SCIENCE IN THIS BOOKLET. HERE IS AN EXAMPLE:

Example Map

Answer List	
plants	coal



PRINT YOUR ANSWERS IN THE BLANK OVALS ON THE EXAMPLE MAP.

WHEN TOLD TO OPEN THE BOOKLET PLEASE READ EACH MAP CAREFULLY BEFORE CHOOSING YOUR ANSWERS. FOR EACH BLANK OVAL SELECT THE ANSWER YOU THINK IS BEST FROM THE ANSWER LIST AT THE TOP OF THAT PAGE. YOU CAN USE THESE ANSWERS MORE THAN ONCE. PRINT YOUR ANSWER CLEARLY IN EACH BLANK OVAL. IF YOU GET STUCK ON AN OVAL, SKIP IT AND GO ON TO ANOTHER ONE.

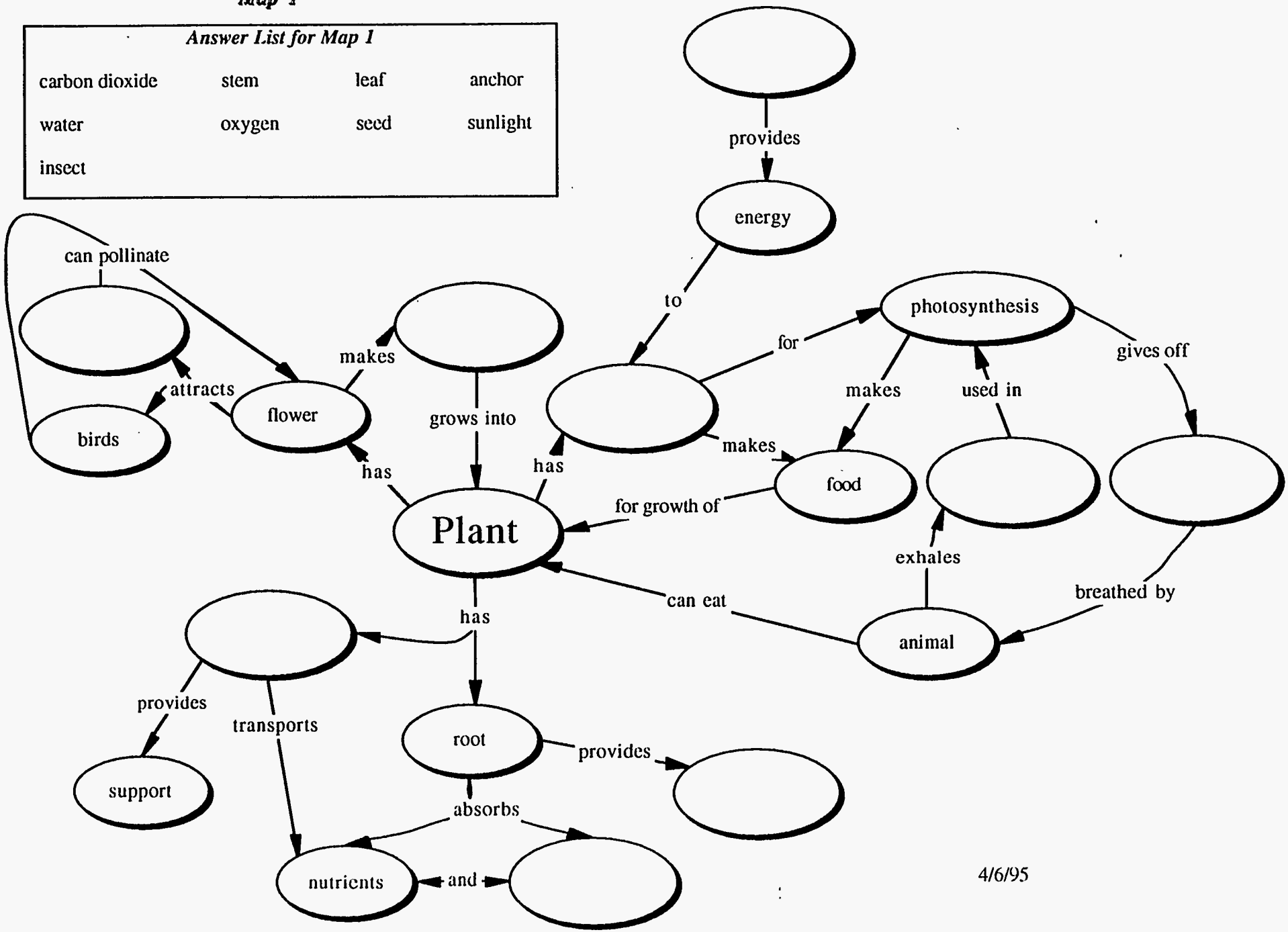
YOU WILL HAVE 20 MINUTES TO FINISH. IF YOU GET DONE BEFORE THE TIME IS UP, GO BACK AND CHECK YOUR WORK.

DO NOT OPEN BOOKLET YET!

Map 1

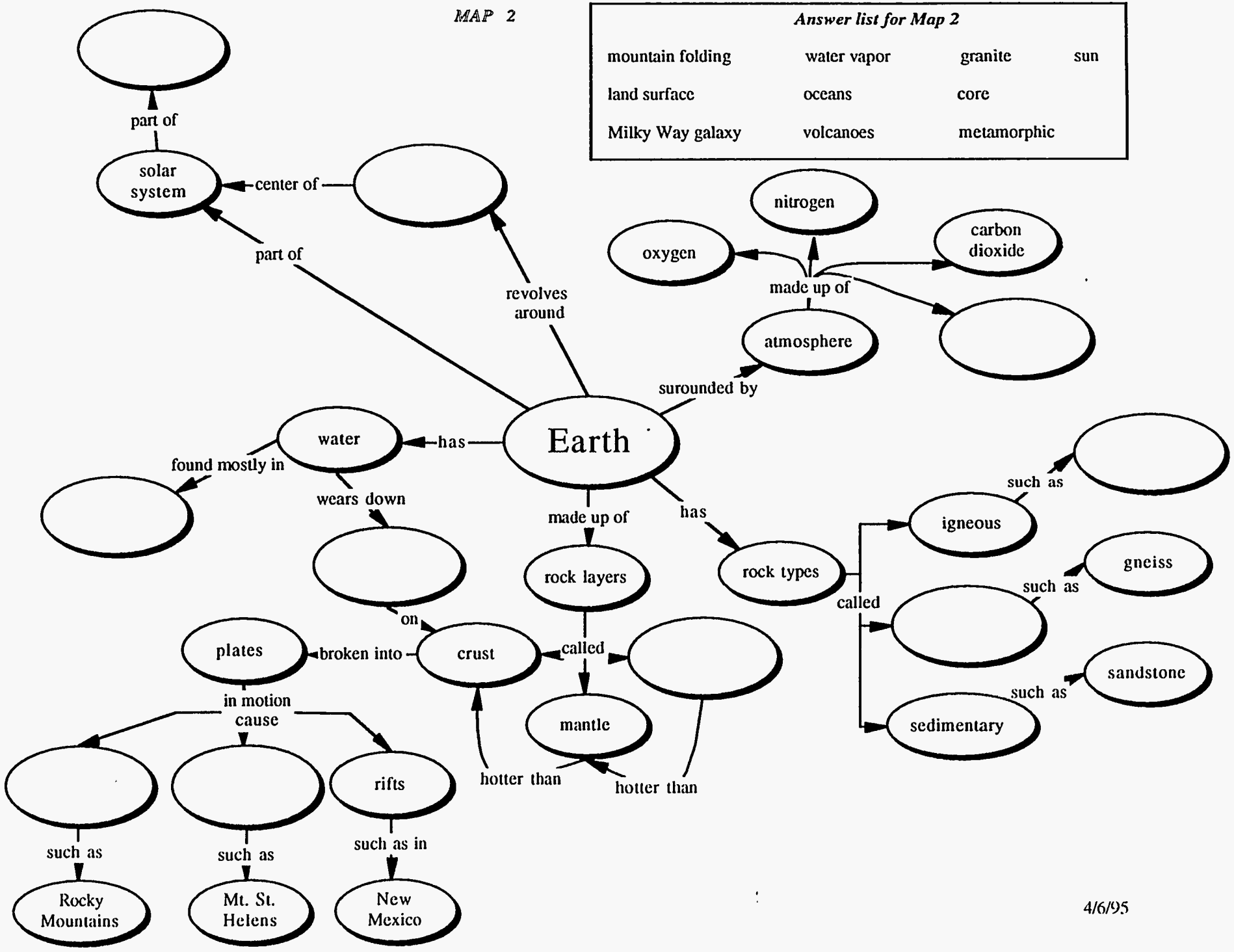
Answer List for Map 1

carbon dioxide	stem	leaf	anchor
water	oxygen	seed	sunlight
insect			



4/6/95

MAP 2



Answer list for Map 2

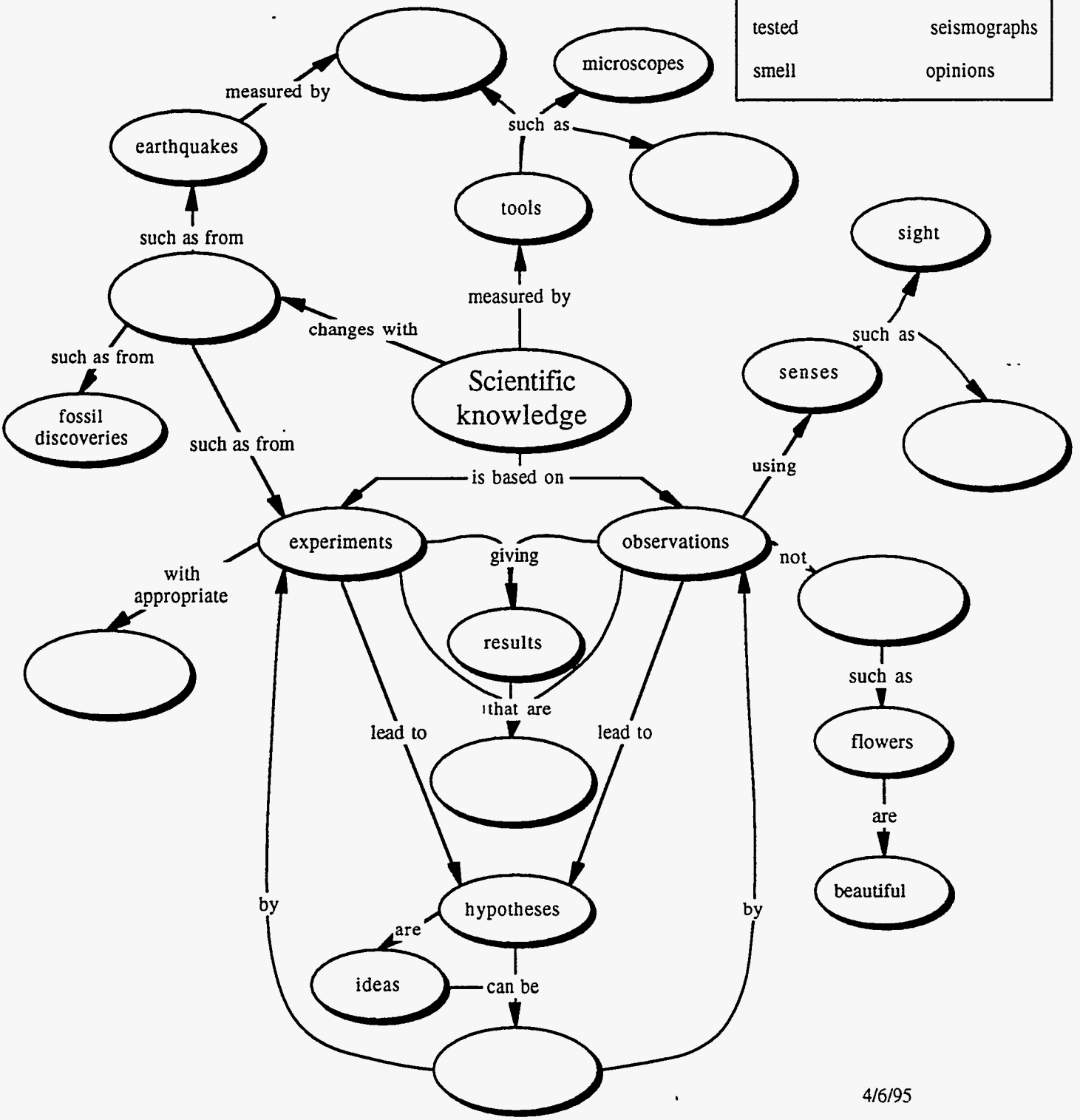
mountain folding	water vapor	granite	sun
land surface	oceans	core	
Milky Way galaxy	volcanoes	metamorphic	

4/6/95

MAP 3

Answer List for Map 3

new evidence	thermometers
controls	repeatable
tested	seismographs
smell	opinions

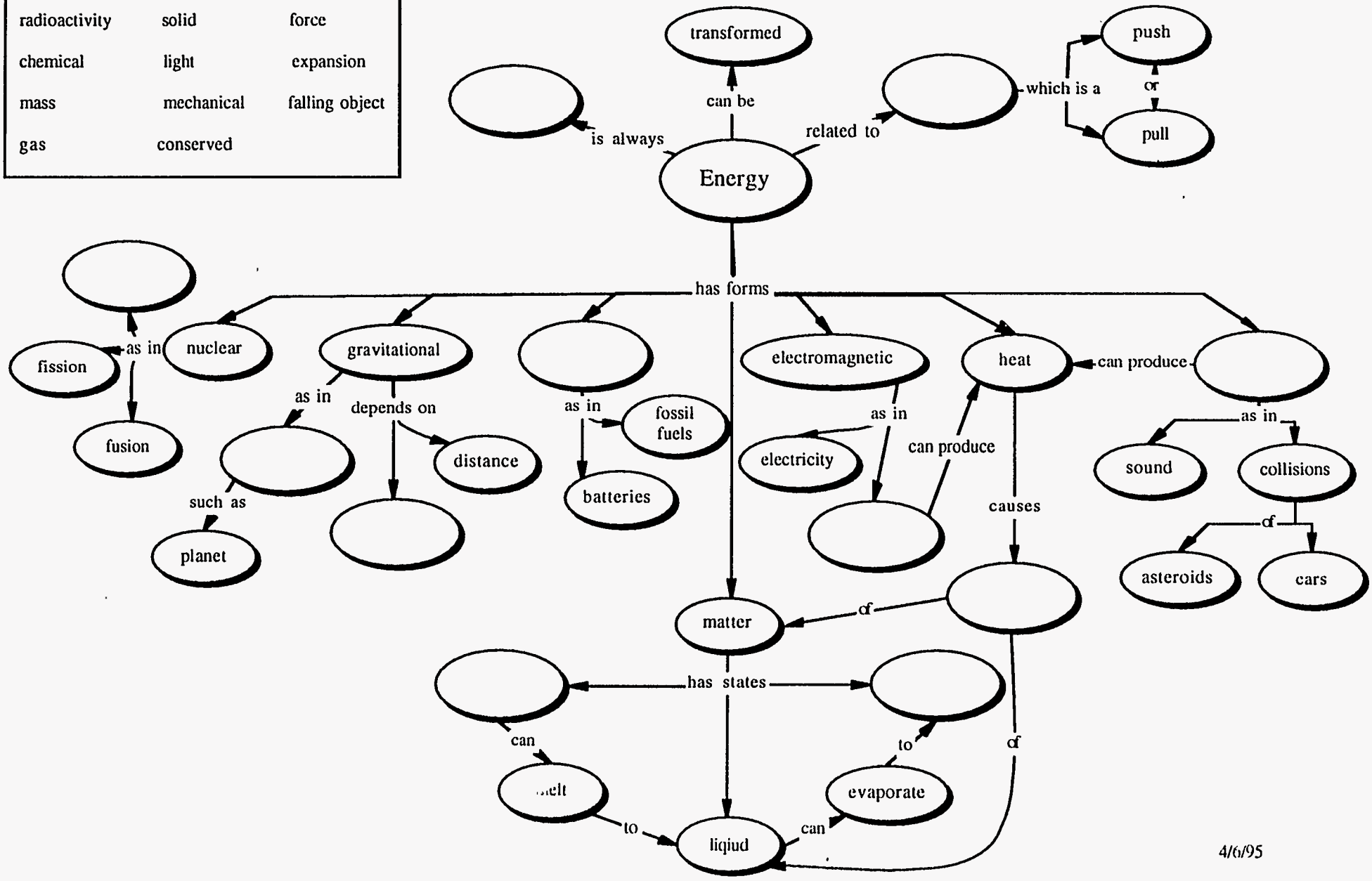


4/6/95

MAP 4

Answer List for Map 4

radioactivity	solid	force
chemical	light	expansion
mass	mechanical	falling object
gas	conserved	



SCIENCE QUESTIONS

DIRECTIONS: PRINT YOUR NAME AND ANSWER THE QUESTIONS ON THE TOP OF THE ANSWER SHEET. PLEASE STOP AT THE SECTION MARKED SCIENCE QUESTIONS.

THERE ARE TWENTY-EIGHT QUESTIONS ABOUT SCIENCE IN THIS BOOKLET.

HERE IS AN EXAMPLE:

Example: Oil and coal are formed

A from the remains of plants and animals

B from minerals

C by earthquakes

D by cooling lava

ON YOUR ANSWER SHEET FIND THE "EXAMPLE" IN THE SCIENCE QUESTIONS SECTION. FILL IN THE OVAL ON YOUR ANSWER SHEET BESIDE THE WORD "EXAMPLE."

WHEN TOLD TO OPEN THE BOOKLET PLEASE READ EACH QUESTION CAREFULLY BEFORE CHOOSING YOUR ANSWER. NOTICE THAT QUESTIONS ARE ON BOTH SIDES OF THE PAGES. MARK YOUR ANSWER TO EACH QUESTION ON THE SEPARATE ANSWER SHEET. BE SURE TO COMPLETELY DARKEN THE OVAL ON THE ANSWER SHEET THAT MATCHES THE LETTER OF YOUR ANSWER. IF YOU GET STUCK ON A QUESTION, SKIP IT AND GO ON TO THE NEXT ONE.

YOU WILL HAVE 20 MINUTES TO FINISH MARKING YOUR ANSWERS. IF YOU GET DONE BEFORE THE TIME IS UP, GO BACK AND CHECK YOUR WORK.

DO NOT OPEN BOOKLET YET!

1. Which of the following gases must an animal breathe in order to remain alive?

   A Helium

   B Hydrogen

   C Nitrogen

   D Oxygen

2. The bulb of a thermometer is placed in your mouth. Which of the following explains why the level of the liquid rises in the thermometer?

   A Hot air rises inside the thermometer.

   B Heat energy changes into light energy.

   C A liquid expands when heated.

   D Heat can change a solid into a liquid.

3. Which of the following provides the best evidence that light is a form of energy?

   A Light reflects from a smooth surface like glass.

   B Light raises the temperature of an object on which it falls.

   C Light usually travels in straight lines.

   D Light diffracts when it passes through a narrow opening.

4. A northeast wind is one that is blowing toward the

   A northeast

   B northwest

   C southwest

   D southeast

5. Which of the following methods could be used to find out how much space is taken up by rocks placed in a 1-liter container?

   A Weigh the rocks, and then weigh the container, then subtract the weight of the container from the rocks.

   B Empty the container and count the number of rocks.

   C Fill the container holding the rocks with water, pour the water into a measuring cup, then subtract the amount of water from the entire capacity of the container.

   D Weigh the container empty, fill it with the rocks and water, then weigh it again.

6. The evidence gained from an experiment will always be more reliable if the experiment is conducted

   A with expensive equipment

   B by a team of scientists

   C with appropriate controls

   D in a laboratory

7. Which of the following is an opinion rather than an observation?

   A Many plants are green.

   B Many flowers are beautiful.

   C Plants require sunlight.

   D Plants can grow in different places.

8. Which of the following would be the best evidence that something is alive?

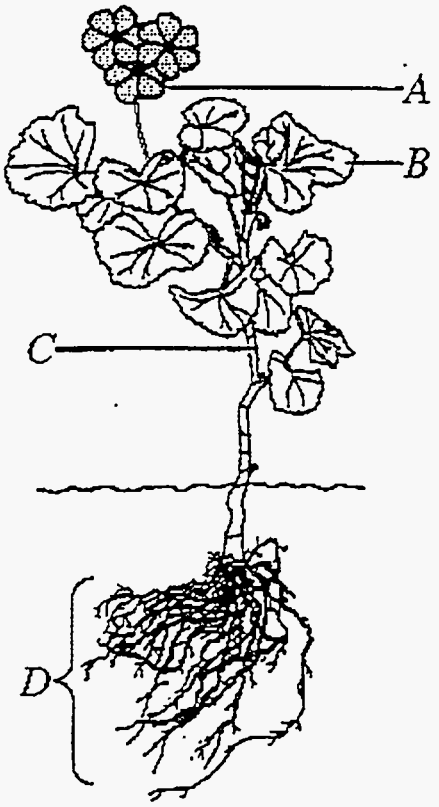
   A It is made of cells.

   B It has a regular shape.

   C It changes color.

   D It is large.

Questions 9-10 refer to the plant shown below.



9. The main function of part A of the plant is to

   A make oxygen

   B store oxygen and water

   C catch sunlight and store energy

   D attract insects and make seeds

10. Which plant part makes most of the plant's food?

   A A

   B B

   C C

   D D

11. Some scientists think that dinosaurs became extinct shortly after a huge meteor crashed into Earth about 60 million years ago. They think this extinction is related to a thick cloud of dust thrown into the air by the meteor.

How might this event have caused the extinction of dinosaurs?

A The dust blocked the sunlight changing the climate.

B The dust made it impossible for dinosaurs to find mates.

C The dust buried the dinosaurs.

D The dust covered the dinosaurs' food.

12. Which of the following wears down the Earth's land surface the most?

A Running water

B Earthquakes

C Volcanoes

D Wind

13. Of the numbers below, which would be the number of plants the scientist could use to obtain the most reliable data?

A 1

B 2

C 20

D 200

14. Which of the following is NOT considered to be a major type of rock?

A Igneous

B Sedimentary

C Metamorphic

D Mineral

15. Several students notice that the river running through a park is shallower than it was one week before. One student says that this change must be due to a lack of rain. This statement is best described as

A a hypothesis that can be investigated

B an observation based on theory

C a conclusion based on firm evidence

D a theory based on experimentation

16. Hypotheses are

A ideas that can be tested

B facts about science

C observations of nature

D results of experiments

17. Which of the following is an example of force?

A A wedge

B A heavy load

C A strong muscle

D A push or pull on something

18. At which of the following temperatures would water begin to freeze?

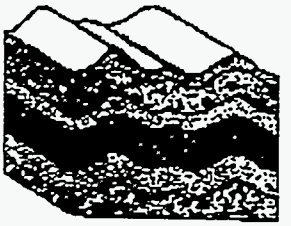
A 0° C

B 32° C

C 100° C

D 212° C

19. The diagram of the Rocky Mountains below best represents the effects of which of the following geologic processes?



A Folding

B Faulting

C Weathering

D Earthquakes

20. Which of the following statements about scientific knowledge is correct?

A It is based on observations and experiments that can be repeated by scientists.

B It cannot be tested.

C It is based on laws that never change.

D It is based on beliefs and faith.

21. Which of the following best describes how the temperature in the Earth's layers changes between the core and the crust?

A It increases from the core through the mantle and then decreases toward the crust.

B It increases from the core toward the crust.

C It remains constant from layer to layer.

D It decreases from the core toward the crust.

22. Which of the following instruments is used to measure earthquakes?

A Microscope

B Barometer

C Seismograph

D Telescope

23. A group of students is told to put 30 fossils into several groups. The students find it difficult to agree on one solution. The main reason that scientists often have the same problem is that

A fossils have not been studied very much

B they work in groups

C fossils are very old

D there are many different ways to classify fossils

24. Which of the following contains the greatest amount of Earth's water?

A Rivers, lakes, and ponds

B Oceans

C Atmosphere

D Glaciers

25. Knowledge of Earth's past continues to change as scientists find additional fossils. This is because

A scientific knowledge cannot be trusted

B scientists change their ideas as new evidence is found

C scientists do not accurately report what they observe

D fossil study is not a true science

26. The main energy source for photosynthesis is

A soil

B heat

C sunlight

D radioactivity

27. Which of the following statements about gravitational attraction between two objects is true?

A It does not depend upon the mass of the objects.

B It increases as the masses of the objects increases.

C It increases as the distance between the objects increases.

D It does not depend upon the distance between the objects.

28. Which of the following causes the seasons to change as the Earth revolves around the Sun?

A The tilt of Earth's axis

B The gravitational pull of the Moon

C The intensity of light emitted by the Sun

D The frequency of sunspot occurrences

Please fill in the Student Survey section of your answer sheet.

If you finish before time is called please check your answers to the Science Questions.

Appendix B

Statistics Information

Analysis Techniques

Internal Structure of Measures

We first examined the internal structure of each measure using traditional reliability indices, including Cronbach's alpha for the measure as a whole and the squared multiple correlation coefficient and deleted alpha for each item. Cronbach's alpha is a global measure of how well the items together represent the total score. It usually varies from about 0 (a very poor measure whose score means nothing) to 1 (a perfect measure); neither extreme is usually attained. Researchers hope for scores that are as high as possible; values in the .90's often are considered excellent, in the .70's and .80's acceptable, and below .70 not acceptable. It is harder to obtain statistical significance from tests done on scores from a measure with an unacceptably low internal consistency. Each item can be evaluated in several ways. We utilized two common ones: alpha if the item is deleted and the item-total correlation coefficient. For each item, the global alpha is recalculated omitting that item from the total score. If the item works well in the measure, the deleted alpha will be lower than the alpha that includes the item. A bad item has a deleted alpha that is higher than the alpha with the item included, indicating that the total score is better without the item. The item-total correlation coefficient indicates how well each item relates to the total score (excluding that item). Again, researchers want higher positive values; values that are negative are unacceptable. Items that did not work well in our total scores were eliminated as outlying (unusual) items.

Outlying Scores

We then examined our score distributions for outlying students. Most students appeared to participate fully, to understand the tasks, and to try to select a reasonable response for each item. Some, however, did not; these students were outliers. We anticipated two possible kinds of outliers. The first did not fully participate; that is, they did not select responses to at least half of each measure. In this report, general performance levels do not include these non-participators.

Of the 249 Anglo, Hispanic, and Native American seventh grade students, we identified and eliminated five non-participators on the achievement measures (two on the multiple-choice measure, two on the fill-in concept map measure, and one on both measures), leaving 244 in the analysis sample. These five students were from the same school and were called out of their classrooms during data collection. Ten students were non-participators on the attitude survey, leaving 239 students for those analyses; nine were non-participants on the comparison items, leaving 240 students.

Of the 105 Anglo and Hispanic eighth grade students, we identified and eliminated one non-participator on the achievement measures (leaving 104 students), one on the attitude measure (leaving 104), and four on the comparison items (leaving 101).

The second type of outlying student scored either much higher or much lower than the

rest of the students in their analysis cell. These statistical outliers were eliminated from all statistical analyses due to their extreme influence on results. We examined all score distributions in each most complex interaction cell: order by gender by ethnicity cells for seventh grade; order by gender, order by ethnicity, and gender by ethnicity cells for eighth grade. Students whose responses placed them more than 20% of the possible scale from their next closest neighbor(s) on that measure or more than 3 standard deviations from the mean and separated from their next closest neighbor(s) were identified and eliminated. For example, the maximum possible score on the mapping measure was 37. If students at either the high or low end of the cell distribution scored eight or more points (20% of 37) from the next closest score(s), they were identified and eliminated from statistical analyses.

Of the remaining 244 seventh grade students in the achievement sample, seven were extreme scorers. They were about equally distributed by sex, ethnicity, and order of administration. Two were eliminated due to exceptionally high scores (1 on the map measure and the other on both) while five obtained exceptionally low scores (all on the map measure). This process left 237 seventh graders. One student was a low outlier on the attitude survey.

Of the remaining 104 eighth grade students, six were extreme scorers on the achievement measures, leaving 98 students in the analysis sample. Four scored exceptionally low on the map measure, one exceptionally low on the multiple-choice measure, and one exceptionally high on the map measure. One student scored exceptionally low on the attitude survey and was eliminated.

Relationships Among Score Distributions

Person correlation coefficients were used to examine the magnitude and direction of the relationships among the score distributions of the measures. Two-tailed alpha values were set at the traditional values of .05 and .01. When relationships were statistically significant, the correlation coefficient was squared to yield percent of variance shared by the distributions.

Gender and Ethnic Relationships

Gender and ethnic group relationships were examined using analysis of variance techniques, with alpha values of .05 and .01. Order of administration of the measures was included as a third independent variable. For the seventh grade, we had large enough numbers of subjects to examine three ethnic groups (Hispanic, Native American, and Anglo) and to test all possible interactions. Our eighth grade sample was smaller. We could not include Native Americans due to the small number in our sample nor could we test the third-order interaction; we pooled that interaction into the error term. Significant interactions were analyzed using simple effects with a family-wise adjustment to alpha values. The unique approach was used. In all analyses, variances were homogeneous, and examination of residuals gave no indication of model outliers. Significant results were evaluated for meaningfulness by examining mean percent differences.

ANOVA Source Tables for Map Measure

SEVENTH GRADE:

	SS	df	MS	F
Gender	38.76	1	38.76	<1
Ethnicity	2638.36	2	1319.18	28.28**
Maps first	555.07	2	277.53	5.95*
M-C first	2742.63	2	1371.31	29.40**
Order	28.10	1	28.10	<1
for Anglos	237.40	1	237.40	5.09**
for Hispanics	198.87	1	198.87	4.26*
for Native Am.	371.31	1	371.31	7.96**
Gender by Ethnicity	20.36	2	10.18	<1
Gender by Order	45.72	1	45.72	<1
Ethnicity by Order	760.16	2	380.08	8.15**
Gender by Ethnicity by Order	263.30	2	131.65	2.82
Error	10495.75	225	46.65	
Total	14107.25	236	59.78	

EIGHTH GRADE:

	SS	df	MS	F
Gender	219.48	1	219.48	4.34*
Ethnicity	170.71	1	170.71	3.37
Order	151.48	1	151.48	2.99
Gender by Ethnicity	8.26	1	8.26	<1
Gender by Order	108.28	1	108.28	2.14
Ethnicity by Order	28.41	1	28.41	<1
Error	4604.14	91	50.59	
Total	5660.34	97	58.35	

*$p \leq .05$ experiment-wise (<.025 family-wise for the simple effects analyses),
**$p \leq .01$ experiment-wise (<.005 family-wise)

ANOVA Source Tables for Multiple-choice Measure

SEVENTH GRADE:

	SS	df	MS	F
Gender	19.11	1	19.11	1.24
Ethnicity	491.61	2	245.81	15.98**
Maps first	269.61	2	134.80	8.76**
M-C first	352.72	2	176.36	11.46**
Order	47.25	1	47.25	3.07
for Anglos	13.34	1	13.34	<1
for Hispanics	12.50	1	12.50	<1
for Native Am.	102.57	1	102.57	6.66**
Gender by Ethnicity	37.41	2	18.70	1.22
Gender by Order	4.55	1	4.55	<1
Ethnicity by Order	107.49	2	53.75	3.49*
Gender by Ethnicity by Order	52.87	2	26.43	1.72
Error	3461.91	225	15.39	
Total	4239.98	236	17.97	

EIGHTH GRADE:

	SS	df	MS	F
Gender	135.56	1	135.56	6.55*
Ethnicity	118.30	1	118.30	5.71*
Order	3.59	1	3.59	<1
Gender by Ethnicity	41.03	1	41.03	1.98
Gender by Order	16.43	1	16.43	<1
Ethnicity by Order	50.41	1	50.41	2.43
Error	1884.69	91	20.71	
Total	2240.20	97	23.09	

*$p \leq .05$ experiment-wise (<.025 family-wise for the simple effects analyses),
**$p \leq .01$ experiment-wise (<.005 family-wise)