DOCUMENT RESUME

ED 206 644 TM 810 533

AUTHOR Diekhoff, George M.
TITLE Relationship Judgments and Multidimensional Scaling in the Measurement of Structural Understanding.
PUB DATE Apr 81
NOTE 26p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (Houston, TX, April, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Classification; Cognitive Ability; *Concept Formation; Correlation; Essay Tests; Higher Education; Measurement Techniques; *Multidimensional Scaling; Multiple Regression Analysis; *Predictive Validity; Predictor Variables; *Testing; Test Reliability

ABSTRACT

The aim of this study is to suggest an alternative to the essay testing method of assessing structural understanding of concepts, which suffers from time constraints and lack of scoring reliability. In this method, students' relationship judgments between concept pairs are examined directly rather than by means of multidimensional scaling analysis (a common alternative to essay testing). The similarity between student and instructor judgments, and the reliability of students' judgments as predictors of structural understanding as measured by essay examinations, are investigated. Data were collected over three separate examinations in each of two experiments. In the first experiment, numerical judgments of the strength of relationships between concepts were obtained from students in three units of undergraduate general psychology. The second experiment demonstrated that restricting relationship judgments to high reliability concept pairs increases the accuracy of essay score predicitions from predictor measures derived from reliability judgments. Reliability and correlation measures are obtained by means of multiple regression analysis. Results indicated that students' relationship judgments do not need to be analyzed through multidimensional scaling in order for them to be useful in evaluating structural understanding, and that this sample correlational approach is a valid and objective alternative to essay testing. (AEP)

Relationship Judgments and Multidimensional Scaling in the Measurement of Structural Understanding

George M. Diekhoff, Ph.D.
Department of Psychology
Midwestern State University
Wichita Falls, TX 76308

Paper presented at the meeting of the Southwestern Psychological Association, Houston, April, 1981.

Abstract

Essay testing has traditionally served as a means by which students' understanding of structural interrelationships between concepts could be evaluated and promoted. However, essay test scores frequently suffer from poor reliability and validity, and the time required to evaluate essay test responses may be prohibitive in large classes. This paper discusses an alternative procedure for measuring and promoting structural understanding which overcomes many of the problems encountered when using essay tests. In Experiment I, numerical judgments of the strength of relationships existing between concepts from three units of undergraduate general psychology were obtained from between 81 and 103 students. Both the reliability with which these judgments were made and the similarity between students' judgments and those of the instructor were found to be significantly correlated to students' essay test scores over the same three units. Previous to this research, judgment reliability has served only as a criterion by which relationship judgment data was judged acceptable or unacceptable for further analysis through multidimensional scaling (MDS). Importantly, MDS analysis of relationship judgments was not necessary in order for these judgments to serve as an effective means of evaluating structural understanding in the present research. Consequently, only those relationship judgments on which experts show strong agreement need be considered. Multidimensional scaling analysis, in contrast, requires that all judgments, even those which may be unstable, be obtained and utilized in evaluations of cognitive structures. Experiment II demonstrated that restricting relationship judgments to these high-reliability concept pairs increased the accuracy of predictions of essay scores from predictor

measures derived from relationship judgments. The use of relationship judgments may provide an alternative to essay testing in situations in which essay testing is impractical.

Relationship Judgments and Multidimensional Scaling in the Measurement of Structural Understanding

The widespread use of essay examinations is based on the premise that essay examinations are capable of testing different kinds of knowledge than can be easily tested through multiple-choice or true-false examinations. Indeed, there is support for this proposition. It is well established that the free recall of information demanded by essay examinations requires considerably greater integrative rehearsal (Lindsay & Norman, 1977) or depth-of-processing (Craik & Tulving, 1975; Craik & Watkins, 1973; Woodward, Bjork, & Jongeward, 1973) during learning than is needed in order for recognition to occur. Multiple-choice and true-false examinations usually test only recognition. Since integrative rehearsal or deep processing involves processing the meaning of information, it follows that essay examinations more clearly test students' comprehension than do recognition tests such as multiple-choice and true-false. Bloom (1956) and others (e.g., Ayers, 1966; Billeh, 1974; Gall, 1970; Roberts, 1976; and Scriven, 1967) have also argued that essay testing can more easily be used in evaluating "higher levels" of understanding (specifically, the structural interrelationships and implications of a domain) than can multiple-choice or true-false tests. Yet another reason for the use of essay examinations is the observation that students who anticipate examinations which focus on higher levels of understanding seem to study and learn in a qualitatively different fashion than do students who expect multiple-choice or true-false examinations. That is, a teacher may guide learning through a careful manipulation of students' expectancies concerning examinations. Doak (1970), Ladd & Anderson (1970), and Willson (1973) have all shown that

the level of a teacher s questions and discourse influences student's level of discourse and performance on examinations designed to assess several levels of understanding.

While essay examinations are helpful in evaluating and promoting higher levels of understanding, the utility of this testing method is limited. Chase (1968), Linn, Klein, & Hart (1972), Marshall (1967, 1972), Marshall & Powers (1969), and Scannel (1966) have pointed to the influence of a variety of extraneous characteristics of essay responses that influence grades, including neatness, spelling errors, and grammatical errors. In addition, the quality of previously graded essay responses can influence the scoring of subsequent responses through a contrast effect. Positive and negative halo effects established through a scorer's prejudices also may influence essay grades. Clearly, essay exam scores are influenced by many factors other than the quality of information contained in essay responses, and thus, may suffer from poor reliability and validity. While these difficulties may be overcome to some extent through careful control of essay scoring procedures, it is not so simple a matter to eliminate another disadvantage to the use of essay examinations: the time required to adequately score essay responses from large classes.

Alternative methods of assessing higher-level, structural understanding have been investigated, presumably in the hope of finding a less time-consuming, more objective testing method. Word-association and graph-construction methods are among these alternatives (Johnson, 1967, 1969; Preece, 1976; Shavelson, 1972, 1973, 1974), but both are more time consuming to score than are essay tests.

Another approach to evaluating understanding of the structural interrelationships between concepts in a domain has been multidimensional scaling (MDS) of students' similarity or relationship judgments. Here, students use a numerical

scale in judging the strength of relatedness or similarity within all possible pairs of a selected set of concepts. These judgments are considered to be equivalent to distances between the concepts and are used in constructing a graphic array of points in space (sometimes called a "cognitive map") that best captures the pattern of relationship judgments considered as a whole. Cognitive maps formed through MDS reflect some important aspects of a student's structural understanding of the domain from which the concepts were sampled. For example, Johnson, Cox, & Curran (1970) obtained similarity ratings for all possible pairs of six concepts from mechanical physics. The similarity judgments were averaged across students and analyzed through MDS. The MDS-produced map was found to be very similar to arrays constructed through logical examination of the mathematical relationships that exist between the six concepts. Thus, MDS analysis of similarity judgment data was shown to capture mathematical relationships that exist between concepts in mechanical physics.

Wainer & Kaye (1974) also validated MDS-produced cognitive maps as a measure of structural understanding. Sixteen concepts from developmental psychology were scaled by 45 undergraduates before and after receiving instruction in that topic area. Averaged similarity judgments were analyzed through MDS and the cognitive map obtained was compared to a map based on the instructor's similarity judgments. The student map obtained following instruction was found to be significantly more similar to the instructor's map than was the map formed prior to instruction.

Fenker (1975) also reported a study in which MDS-produced cognitive maps were evaluated as a measure of structural understanding. At the beginning of an undergraduate course in statistics and research design, 27 students were given a list of 21 concepts to which they were told they should give special

attention. Following completion of the appropriate course units, students rated the relatedness of these 21 concepts. Several of the concept pairs were repeated and judgments obtained from these repeated pairs were correlated with initial judgments for those pairs, enabling assessment of each student's reliability in making the relationship judgments. Ten of the students showed reliability coefficients of less than .50 and their data was given no further consideration. Relationship judgments from the remaining 17 students were analyzed through MDS and these student maps were compared to a map based on averaged relationship judgments obtained from a panel of eight experts (faculty members and graduate students). A significant correlation was found between students' course grades and an index of overall similarity between the maps of students and the expert map. There was also a high degree of similarity between maps of the individual experts.

In sum, efforts aimed at developing MDS of relationship judgments as a method of evaluating and promoting structural understanding have been promising. Additional work, however, is needed. Relationship judgments carry all of the information which subsequently appears in MDS-produced cognitive maps, yet characteristics of these original, unanalyzed relationship judgments have not been examined as measures of structural understanding. In addition, previous work has focused exclusively on the relationship between students' test scores and the similarity between their cognitive maps and those of the instructor. Other characteristics of students' relationship judgments, such as judgment reliability, may also reflect structural understanding, but have not been examined. The purpose of the first experiment was to examine the relationship between characteristics of students' original, unanalyzed relationship judgments and their performance on essay examinations covering the content of three units of an undergraduate general psychology course. The second experiment

sought to determine whether or not the strength of this relationship could be enhanced by having students judge only concept pairs on which several experts showed high levels of agreement.

EXPERIMENT I

Method

Subjects

Students enrolled in five sections of an undergraduate general psychology course participated in Experiment I. All sections were taught by the same instructor using similar lecture notes and identical testing materials. The five sections were taught over a period of three semesters, but data from the five sections have been combined. Data from three examinations were examined in this study. The number of students who completed each of the three examinations was 103, 81, and 88 for exams 1, 2, and 3, respectively. The number of students varied as a consequence of attrition and absences.

Procedure

Students were tested on three separate occassions over information presented in three units of the general psychology course (biopsychology, developmental psychology, and social psychology). Each of the three examinations consisted of four parts: (a) a 40-minute, 50-item multiple-choice test which focussed on definitions of single concepts; (b) a 15-minute relationship judgment test in which students were instructed to assign a number between 1 and 9 to each of 45 pairs of concepts to indicate how strong a relationship they thought existed between the concepts in that pair. A rating of 9 indicated that the two concepts were viewed as very highly related, while a rating of 1 indicated that the two concepts were felt to be unrelated; (c) a second set of nine concept pairs sampled from the larger set which students were instructed to rate again; and, (d) a 20-minute essay test which called for students to discuss

the nature of the relationships within each of three pairs of concepts selected by the instructor.

Each section of each examination was collected prior to administering the subsequent section. Students were told that their course grades would be based 75% on multiple-choice performance, 25% on essay performance, and that "borderline grades" would be decided through an examination of the "reasonableness" of their relationship judgments. Since the purpose of this study was to examine the relationship between students' strength-of-relationship judgments and their performance on essay examinations of structural understanding, only relationship judgments and essay exams will be given further consideration here.

Scoring and Reliability of Essay Tests

Essay tests were scored by the course instructor using guidelines designed to improve scoring reliability. These guidelines indicated the sorts of information that should be included in each response and how many points would be given for each such piece of information. Likely errors were listed and penalties associated with these errors were noted. Each essay item was scored from 0 to 5 points, with 0 assigned in the case of no response, and 5 assigned to a perfect response. The average of each student's item scores served as that student's essay test score. An attempt was made to maintain anonymity in the scoring procedure and there were very few unavoidable exceptions to this rule. Tests were scored item-by-item, rather than student-by-student, so as to reduce the influence of halo effects operating within tests. Tests and test items were scored in a different order on a second occassion one week following the first scoring. Correlations were computed between item scores obtained on these two occassions. All items yielded scoring reliability coefficients of .80 or greater. Total scores showed standard error values of .57, .71, and .42 points for exams 1, 2, and 3, respectively. These values

represent the average absolute differences between essay test scores assigned on the two scoring occassions.

Measures Obtained from Relationship Judgments

Two measures were obtained from each student's relationship judgments. First, judgments obtained for the nine duplicated concept pairs were correlated to the original judgments for these concept pairs. These correlations served as an index of test-retest judgment reliability for each student. While this reliability coefficient is usually used in eliminating unreliable data (e.g., Fenker, 1975), it was examined in the present study as a potential measure of students' structural understanding. The logic was that students who do not understand the interrelationships between concepts in an area will base their relationship judgments on guesswork, and thus, will show less stability in judgments obtained on repeated occassions. The second measure extracted from each student's relationship judgments was a correlation between the student's judgments and those of the instructor. This correlation served as an index of similarity between relationship judgments of students and those of the instructor. Previous research has focussed on the similarity between students' MDS-produced cognitive maps and the maps of their instructors. The correlation between student and instructor judgments was examined in the present study as a simpler and quicker alternative to MDS analysis followed by complex comparisons between students' solutions and that of the instructor.

Relationship Judgments as Measures of Structural Understanding

In evaluating students' strength-of-relationship judgments as indicators of structural understanding, the two variables described above (to be referred to subsequently as "reliab" and "corr") were entered as predictor variables into stepwise multiple regression analyses in predicting students' essay test scores. Three separate analyses were completed, one for each of the three

essay tests.

Results

The results of these stepwise multiple regression analyses are summarized in tables 1 and 2. Scores on all three essay tests were predicted to a

Insert tables 1 and 2 about here

statistically significant extent by linear combinations of "reliab" and "corr." In two out of three cases (tests 2 and 3), "reliab" was a stronger predictor of essay test scores than was "corr."

In evaluating the accuracy of the predictions made through combinations of "reliab" and "corr," essay test scores predicted for each subject were compared with students' obtained essay test scores. Correlations between these two sets of scores (multiple correlations) were .49 (F (2, 100) = 15.61 $p < .001$), .44 (F (2, 78) = 9.20, $p < .001$), and .46 (F (2, 85) = 11.15 $p < .001$) for tests 1, 2, and 3, respectively. The average absolute differences between predicted and obtained essay scores (the standard errors of estimate) were .85, .96, and .83 points on tests 1, 2, and 3. It may be recalled that essay test scores could vary from 0 to 5 points.

In considering the sizes of these errors in prediction, they may be compared to differences found between essay test scores obtained on the two scoring occassions (.57, .71, and .42 points for tests 1, 2, and 3, respectively). While measures derived from relationship judgments do not perfectly predict essay test scores, essay test scores assigned on one occassion also fail to perfectly predict essay test scores assigned one week later.

It is possible to identify several factors which can account for some of the error made in predicting essay scores from "reliab" and "corr." First,

the lack of perfect reliability (i.e., unpredictability) in essay test scores limits the degree to which those scores can be predicted. Second, while essay test items required students to discuss the nature of the relationships between three concept pairs, the relationship judgment tests required students to consider relationships existing in 45 pairs of concepts. Therefore, essay test scores were based on a different, considerably narrower domain than were measures derived from relationship judgments. Since the two types of tests covered different material, the correlation between essay test scores and measures based on relationship judgments cannot be expected to be perfect. Third, because students were told that relationship judgments would only be used in deciding borderline grades, there was little incentive for them to expend much effort in completing the judgments. This undoubtedly introduced some error into relationship judgments which in turn weakened the correlation between relationship judgments and essay test scores. Finally, relationship judgments were obtained for all possible pairs of concepts in each unit. While a complete set of judgments such as this is necessary for MDS analysis, the simpler correlational approach followed in this research does not require that all pairs of concepts be judged. It may be assumed that in Experiment I some of the concept pairs judged for relationship were judged less reliably, even by the instructor, than were other pairs. Thus, students' judgments were compared for similarity to a set of instructor-generated judgments which were somewhat unstable and thus provided a less than ideal standard for comparison.

EXPERIMENT II

The purpose of the second experiment was to further examine the utility of testing structural understanding through relationship judgments. Several of the situational characteristics of Experiment I which may have weakened the relationship between measures derived from relationship judgments and

essay test scores were eliminated in Experiment II: (a) performance on relationship judgments contributed 15% towards each exam grade (multiple choice contributed 60% and essay scores contributed 25%): and, (b) relationship judgments were obtained only for concept pairs on which the top 10 students of previous semesters (based on overall course performance) showed relatively good agreement (specifically, only concept pairs with standard deviations of 2.0 or less were used in Experiment II) and the median of their judgments served as the "expert" judgment pattern against which the judgments of students participating in Experiment II were compared for similarity. (It should be noted that a very high degree of similarity was observed between these median expert judgments and the judgments of the instructor). This change in the selection of concept pairs reduced the number of pairs to be judged from 45 per exam in Experiment I to 15, 20, and 14 in Experiment II on exams 1, 2, and 3, respectively. Because the number of pairs was so reduced, it was possible to administer the entire set of judgments on two occassions in evaluating judgment reliability. It will be recalled that in Experiment I, only nine concept pairs were repeated in evaluating judgmen' -bility.

Subjects

Between 36 and 38 students enrolled in an undergraduate general psychology course participated in Experiment II. The course was taught by the same instructor involved in Experiment I using the same lecture materials and text, and testing procedures were the same as described in Experiment I, except as noted above. The number of students who completed each of the three examinations was 38, 38, and 36 for exams 1, 2, and 3, respectively. The number of students completing the exams varied as a consequence of attrition.

Procedure

Procedures followed during Experiment II were essentially the same as those of Experiment I, with the modifications associated with relationship judgments that have been noted previously. Measures obtained in Experiment II were identical to those of Experiment I, except that judgment reliability ("reliab") was based on completion of the full set of judgments twice and the similarity between each student's judgments and those of the panel of top students ("corr") was determined by computing a correlation between each student's averaged judgments and the median judgments obtained from the 10 top students.

Stepwise multiple regression analyses were again used in evaluating the ability of "reliab" and "corr" to predict students' essay scores.

Results

The results of these stepwise multiple regression analyses are summarized in tables 3 and 4. Essay scores from Exam 2 and 3 were significantly predicted

---

Insert tables 3 & 4 here

---

by linear combinations of "reliab" and "corr," although essay scores from exam 1 were not. The multiple correlations between "reliab" and "corr" and essay scores were .28 (F (2, 35) = 1.44, p = n.s.), .61 (F (2, 35) = 10.17, $p < .005$) and .76 (F (2, 33) = 22.32, $p < .001$) for exams 1, 2, and 3, respectively. These multiple correlations may be compared to the values obtained in Experiment I: .49, .44, and .46, for exams 1, 2, and 3.

The average absolute differences between obtained and predicted essay scores (the standard errors of estimate) were .87, .85, and .70 for exams 1, 2, and 3,

compared to the standard error values from Experiment I: .85, .96, and .83 points. It should again be remembered in interpreting the sizes of these errors that essay scores could vary from 0 to 5 points and that the essay scores were themselves not totally reliable.

In comparing the predictive accuracy of Experiment II to that found in Experiment I, a general tendency towards greater predictability of essay scores in Experiment II is seen. In Experiment I, the average percentage of essay score variance predicted across the three exams was 22% compared to 34% in Experiment II. The average standard error of the estimate across the three exams of Experiment I was .88 points, while the average standard error of the estimate for Experiment II was .81 points.

In further comparing results from the two experiments, the predictive contribution of "corr" increased from Experiment I to Experiment II, perhaps because the "expert" judgments were more stable, being based on only highly reliable concept pairs. The predictive contribution from "reliab" remained approximately the same as in Experiment I.

Discussion

Measuring and promoting students' understanding of the structural interrelationships between concepts calls for the use of testing procedures which focus on these interrelationships. Essay testing has traditionally been used for this purpose, but suffers from several weaknesses, not the least of which is the great time required to reliably score essay exams administered in large classes. Testing knowledge of interrelationships by having students rate the perceived strength-of-relationship between concepts presented in pairs was examined in the present study as an alternative to essay testing for use in those situations in which essay testing is impractical.

In contrast to previous research in this area, the present study examined students' original relationship judgments, rather than MDS-analyzed judgments; both the similarity between students' and the instructor's judgments and the reliability of students' judgments were examined as predictors of structural understanding as measured by essay examinations; and the study used data collected from a large number of students replicated over three separate examinations.

It was found from Experiment I that the reliability of students' relationship judgments were significantly related to essay test scores. This finding is important because judgment reliability has previously been used as a criterion for accepting or discarding relationship judgment data (e.g., Fenker, 1975), whereas it was found in Experiment I to be an important indicator of structural understanding. The present study also showed that students' relationship judgments do not need to be analyzed through multidimensional scaling prior to comparison to expert judgments in order for them to be useful in evaluating structural understanding. Pearson product-moment correlations between each student's judgments and those of the instructor of panel of "experts" provided statistically significant prediction of students' essay scores. The importance of this finding is twofold. First, previous research which has attempted to measure structural understanding through relationship judgments has always involved MDS analysis of relationship judgments followed by comparisons of MDS solutions for similarity. These procedures are far more difficult and time consuming than the simple correlational approach taken in the present research. Second, MDS analysis of relationship judgments requires a complete matrix of judgments, i.e., all concept pairs must be judged. Even experts may have difficulty generating reliable judgments for some of these pairs, and the "ideal" MDS solutions produced from their judgments will necessar-

ily reflect this poor reliability, thus providing a standard of excellence which is less than perfect. In contrast, if students' judgments are compared correlationally to those of the instructor or expert, it is not necessary to obtain judgments on all possible concept pairs. Only concept pairs which experts can reliably judge for relatedness need be considered, thus providing a superior standard against which student judgments are compared for similarity. Experiment II which utilized only these high-reliability concept pairs demonstrated an increase in the predictive importance of similarity between students' and experts' judgments. Overall, the use of these high-reliability pairs resulted in enhanced prediction of essay performance, relative to that found in Experiment I.

In conclusion, in situations in which essay testing is impractical, the use of testing through relationship judgments may be a reasonable alternative method of assessing and promoting students' structural understanding. Characteristics of students' relationship judgments are significantly correlated to traditional essay measures of structural understanding; evaluation of students' numerical relationship judgments is more objective than is the evaluation of essay responses; a wider range of knowledge can be assessed through relationship judgments in a limited amount of time; and the computerized analysis of students' relationship judgments creates an enormous time saving over the scoring of essay responses.

References

Ayers, J. D. Justification of Bloom's taxonomy by factor analysis. Paper presented at the meeting of the American Educational Research Association, Chicago, February, 1966.

Billeh, V. Y. An analysis of teacher-made science test items in light of the taxonomic objectives of education. Science Education, 1974, 58, 313-319.

Bloom, B. S. The taxonomy of educational objectives: Cognitive domain. New York: David McKay, 1956.

Chase, C.I. The impact of some obvious variables on essay test scores. Journal of Educational Measurement, 1968, 5, 315-318.

Craik, F.W.M., & Tulving, E. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 1975, 104, 268-294.

Craik, F.I.M., & Watkins, M. J. The role of rehearsal in short-term memory. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 599-607.

Doak, E. D. Evaluating levels of thinking. School and Society, 1970, 98, 177-178.

Fenker, R. M. The organization of conceptual materials: A methodology for measuring ideal and actual cognitive structures. Instructional Science, 1975, 4, 33-57.

Gall, M. D. The use of questions in teaching. Review of Educational Research, 1970, 40, 707-721.

Johnson, P. E. Some psychological aspects of subject-matter structure. Journal of Educational Psychology, 1967, 58, 75-83.

Johnson, P. E. On the communication of concepts in science. Journal of Educational Psychology, 1969, 60, 32-40.

Johnson, P. E., Cox, D. L., & Curran, T. E. Psychological reality of physical concepts. Psychonomic Science, 1970, 19, 245-246.

Ladd, G. T., & Anderson, H. O. Determining the level of inquiry in teachers' questions. Journal of Research in Science Teaching, 1970, 7, 395-400.

Lindsay, P. H. & Norman, D. A. Human information processing. New York: Academic Press, 1977.

Linn, R. L., Klein, S. P., & Hart, F. M. The nature and correlates of law school essay grades. Educational and Psychological Measurement, 1972, 32, 267-279.

Marshall, J. C. Composition errors and essay examination grades reexamined. American Educational Research Journal, 1967, 4, 375-385.

Marshall, J. C. Writing neatness, composition errors, and essay grades reexamined. Journal of Educational Research, 1972, 65, 213-215.

Marshall, J. C. & Powers, J. M. Writing neatness, composition errors, and essay grades. Journal of Educational Measurement, 1969, 6, 97-101.

Preece, P. F. Mapping cognitive structure: A comparison of methods. Journal of Educational Psychology, 1976, 68, 1-8.

Roberts, N. Further verification of Bloom's taxonomy. Journal of Experimental Education, 1976, 45, 16-19.

Scannel, D. P. & Marshall, J. C. The effect of selected composition errors on the grades assigned to essay examinations. American Educational Research Journal, 1966, 3, 125-130.

Scriven, M. The methodology of evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1, Chicago, 1967.

Shavelson, R. J. Some aspects of the correspondence between content structure and cognitive structure in physics instruction. Journal of Educational Psychology, 1972, 63, 225-234.

Shavelson, R. J. Learning from physics instruction. Journal of Research in Science Teaching, 1973, 10, 101-111.

Shavelson, R. J. Methods for examining representations of a subject matter structure in a student's memory. Journal of Research in Science Teaching, 1974, 11, 231-249.

Weiner, H. & Kaye, K. Multidimensional scaling of concept learning in an introductory course. Journal of Educational Psychology, 1974, 66, 591-598.

Willson, I. A. Changes in mean levels of thinking in grades 7-8 through use of an interaction analysis system based on Bloom's taxonomy. Journal of Educational Research, 1973, 66, 423-429.

Woodward, A. E., Bjork, R. A., & Jongeward, R. H. Recall and recognition as a function of primary rehearsal. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 608-617.

Table 1

Correlations between essay test scores (essay), relationship judgment reliability coefficients (reliab), and similarity between students' and instructor's relationship judgments (corr).

EXPERIMENT I

Exam 1 (N = 103)			Exam 2 (N = 81)			Exam 3 (N = 88)		
Essay	Reliab	Corr	Essay	Reliab	Corr	Essay	Reliab	Corr
1.00	.16	.49***	1.00	.39***	.31**	1.00	.44***	.21*
	1.00	.36***		1.00	.30**		1.00	.22*
		1.00			1.00			1.00

*** $p < .001$

** $p < .01$

* $p < .05$

Table 2

Stepwise multiple regression summary tables: Prediction of essay scores from relationship judgment reliability (reliab) and similarity between students' and instructor's relationship judgments (corr).

EXPERIMENT I

Exam 1

(N = 103)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Corr	.49	1, 101	31.49***	.85
2	Reliab	.49	2, 100	15.61***	.85

Exam 2

(N = 81)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Reliab	.39	1, 79	13.77***	.98
2	Corr	.44	2, 78	9.20***	.96

Exam 3

(N = 83)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Reliab	.44	1, 85	20.56***	.84
2	Corr	.46	2, 85	11.15***	.83

*** $p < .001$

Table 3

Correlations between essay test scores (essay), relationship judgment reliability coefficients (reliab), and similarity between students' and experts' relationship judgments (corr).

EXPERIMENT II

	Exam 1 (N = 38)			Exam 2 (N = 38)			Exam 3 (N = 36)		
	Essay	Reliab	Corr	Essay	Reliab	Corr	Essay	Reliab	Corr
Essay	1.00	.24	.16	1.00	.53***	.48**	1.00	.28	.76***
Reliab		1.00	13		1.00	.37*		1.00	.43**
Corr			1.00			1.00			1.00

*** $p < .001$

** $p < .01$

* $p < .05$

Table 4

Stepwise multiple regression summary tables: Prediction of essay scores from relationship judgment reliability (reliab) and similarity between students' and experts' relationship judgments (corr).

EXPERIMENT II

Exam 1 (N = 38)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Reliab	.24	1, 36	2.26	.88
2	Corr	.28	2, 35	1.14	.87

Exam 2 (N = 38)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Reliab	.53	1, 36	13.81***	.91
2	Corr	.61	2, 35	10.17**	.85

Exam 3 (N = 36)

Step	Variable Selected	Multiple R	df	F	Std. Error
1	Corr	.76	1, 34	45.63***	.70
2	Reliab	.76	2, 33	22.32***	.70

*** $p < .001$

** $p < .005$