DOCUMENT RESUME

ED 231 646 SE 042 097

AUTHOR Blosser, Patricia E., Ed.; Mayer, Victor J., Ed.
TITLE Investigations in Science Education. Volume 9, Number 1.
INSTITUTION ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.; Ohio State Univ., Columbus. Center for Science and Mathematics Education.
PUB DATE 83
NOTE 91p.
AVAILABLE FROM Information Reference Center (ERIC/IRC), The Ohio State Univ., 1200 Chambers Rd., 3rd Floor, Columbus, OH 43212 (subscription $8.00 per year, $2.25 single copy).
PUB TYPE Collected Works - Serials (022) -- Information Analyses - ERIC Information Analysis Products (071) -- Guides - Non-Classroom Use (055)
JOURNAL CIT Investigations in Science Education, V9 n1 1983

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Academic Achievement; College Science; Earth Science; Elementary School Science; Elementary Secondary Education; Higher Education; *Inservice Teacher Education; Preservice Teacher Education; Questioning Techniques; Science Education; *Science Instruction; Science Teachers; *Secondary School Science; Student Characteristics; Teacher Characteristics; *Teaching Methods; Test Construction; *Testing
IDENTIFIERS Meta Analysis; *Science Education Research

ABSTRACT

Abstractor's analyses of 12 science education research studies are presented. Nine analyses in the first section, focusing on various aspects of science instruction, include: a comparison of different approaches to helping students understand metric units of volume; use of specific questions to cue elementary school students in obtaining information from graphical materials; a meta-analysis of research results on instruction; effects of participation in an inservice program on earth science teachers' attitudes/creativity; use of two different teaching strategies in an earth science course for elementary education majors to determine if contrasting teaching environments would influence students' concept of science instruction; examination of whether teachers who advocated use of living organisms to teach science practiced what they espoused; and an analysis of geology teaching assistant reaction to a training program utilizing video-taped teaching episodes. Three analyses of research on testing are presented in the next section. Research analyzed focused on the reliability/content validity of the Science Curriculum Improvement Study (SCIS) Organism Unit test, comparison of multiple choice/essay tests, and development of an instrument to measure understanding of science. An analysis of a paper on inservice teachers' needs and the author's response to the analysis are provided in the final section. (JN)

INVESTIGATIONS IN SCIENCE EDUCATION

INVESTIGATIONS IN SCIENCE EDUCATION

Volume 9, Number 1, 1983

Published Quarterly by

The Center for Science and Mathematics Education
College of Education
The Ohio State University
1945 North High Street
Columbus, OH 43210

Subscription Price: $8.00 per year. Single Copy Price: $2.25
Add $1.00 for Canadian mailings and $3.00 for foreign mailings.

Volume 9, Number 1, 1983

NOTES FROM THE EDITORS

Twelve research-based articles, previously published in refereed journals, have been analyzed for publication in this issue of ISE. Nine of these articles are reports of research focused on some aspect of instruction. Bilbo and Milkent compared different approaches to helping students understand metric units of volume. Kauchak and colleagues investigated the use of specific questions to cue elementary school students in obtaining information from graphical material. Boulanger used the technique of meta-analysis to synthesize the results of research on instruction. Sallam and Krockover studied the effects of participation an inservice program on earth science teachers' attitudes and creativity. Stallings and others used two different teaching strategies in an earth science course for elementary education majors to determine if contrasting teaching environments would influence students' concept of the teaching of science. Dumpert gathered information to see if teachers who advocated the use of living organisms to teach science practiced what they espoused. Schade and Bartholemew attempted to help graduate teaching assistants gain some knowledge of what constitutes competent instruction.

In the section containing articles on testing, Perry and Merkle assessed the reliability and content validity of the SCIS test for the Organism unit, Gordon compared essay and multiple choice tests, and Fraser reported on the development of an instrument to measure understanding of science.

Another article by Fraser is reviewed in the section on curriculum. The final section contains an analysis of Rubba's article on inservice teachers' needs and Rubba's response to this analysis.

Patricia E. Blosser
Editor

Victor J. Mayer
Associate Editor

INSTRUCTION

Bilbo, Thomas E. and Marlene M. Milkent. "A Comparison of Two Different Approaches for Teaching Volume Units of the Metric System." Journal of Research in Science Teaching 15(1): 53-57, 1978.

Descriptors--College Science; Educational Research; Higher Education; Instruction; *Mathematics; Mathematics Education; *Measurement; *Metric System; Science Education; *Teaching Methods

Expanded abstract and analysis prepared especially for I.S.E. by George G. Mallinson, Western Michigan University, Kalamazoo, Michigan.

Purpose

The purpose of the study was to determine if student understanding of metric units of volume could be enhanced by incorporating direct measuring activities and eliminating a discussion of metric units of area. The investigators state that the literature indicates that the concept of volume is an aspect of metric measurement that poses particular difficulty for students. Although no hypotheses are stated directly, the investigators indicate that the common approach is to teach linear measurement, then area and finally volume. However logical the approach appears to be mathematically, references are cited to suggest that such an approach may complicate the concept of volume and that the step dealing with the measurement of area could be eliminated. The plan for investigating the relative merits of the approaches is evidence that the null hypothesis is being tested. The independent variable was the method of instruction using slide-tape activity programs. The dependent variable was student achievement as measured by the Test of Volume Units of the Metric System, developed specifically for the study.

Rationale

The rationale for the study is implicit in the statements related to the purpose. The common approach to teaching the concept of volume is to sequence the learning experiences beginning with linear measurement, moving to area measurement and then to volume. The investigators indicate that such a sequence has been challenged in the literature based on the belief "that volume is easier to teach than area because it is further removed from length." The exact meaning of this latter statement is a

matter of supposition. At any rate, the divergent views for teaching the concept of volume in the metric system obviously appear to be the motivation for testing both approaches.

Research Design and Procedure

The research design was the randomized control-group posttest-only design described by Van Dalen and Meyer, 1962 [sic]. The subjects were 173 students enrolled in Physical Science I (FS 104) at the University of Southern Mississippi during the fall quarter 1975. The randomization involved the assignment of the subjects to three groups referred to as Approach A (N = 59), Approach B (N = 51), and Approach C (N = 63). Those in Approach A studied length, then area and finally volume with an emphasis on computation. Those in Approach B studied volume but not area. Those in Approach C did not receive any instruction related to metric units of volume. The instructional methodology for the experimental approaches (A and B) consisted of slide-tape programs developed especially for the study. These included exercises involving computations, "hands-on" measuring experiences, and practice in estimating. The students in the control group (C) studied a unit on heat and temperature that included films, activities, and problem worksheets.

In the first unit the students in Approach A studied metric units of length and had exercises in measuring these units. This was followed by similar activities with metric units of area including practice in measuring and computing these units. The next step involved metric units of volume including descriptions of these units, practice in calculating volume in cubic units, exercises involving equating cubic centimeters with milliliters, and measuring with metric units of volume. The final unit dealt with estimating length, area and volume.

In Approach B the students dealt with volume but not area and with little stress on length. The activities with volume were similar to those in Approach B but, without emphasizing area and length, more time was spent with direct measuring experiences than with Approach A. All three approaches (A, B, and C) involved three and one-half hours of instructional time and one hour of testing.

Student achievement was measured with an instrument developed by the principal investigator entitled Test of Volume Units of the Metric System. Some of the items were those taken from other tests with permission of the authors and some were prepared by the principal investigator. The final test consisted of 52 multiple-choice items on three subscales, 20 designed to test the student's ability to estimate volumes of various solids and containers; 20 related to computation; and 12 designed to measure the student's ability to make measurements involving metric units of volume. The validity of the 52 items was established by having 12 "selected authorities" evaluate the items in the original pool for agreement with the answer, clarity and phrasing. At least six authorities had to approve each item to include it on the final test. Reliability was assessed by administering the test to 76 students in an introductory physical science class for nonscience majors. The alpha coefficient for the entire test was found to be 0.90 and reliability coefficients for the three subscales were 0.76, 0.86 and 0.73, respectively.

At the end of the treatment, analyses of variance were applied to the scores for the total test, and for each of the three subscales. The analyses indicated that significant differences existed at the 0.01 level, and consequently the Scheffé Test of Multiple Comparisons was applied to "locate the differences."

Findings

With respect to total scores on the Test of Volume Units of the Metric System, the Scheffé Test revealed significant differences at the 0.01 level among the three groups of students. Students in Approach B scored significantly higher than did students in Approach A and both groups scored significantly higher than did those in Approach C. On the computation section of the achievement test, significant differences were found between the experimental Groups A and B and the control Group C. However, significant differences were not found between the experimental groups with respect to estimating ability. A significant difference was found at the 0.01 level between the experimental groups in favor of

Group B. Group A students scored significantly higher than the control group as did Group B students, at the 0.05 and 0.01 levels, respectively.

On the measurement section, significant differences were found at the 0.01 level between Groups A and B and between Groups B and C, in both cases in favor of Group B. However, a significant difference was not found between Groups A and C.

Interpretations

The investigators concluded that the method of instruction eliminating area was more effective in increasing student understanding of metric measurement of volume than was the method that stressed length, area and volume. Also, it was concluded that the volume-only approach was more effective in teaching estimation of volume than was the computation approach, although both were better than having no related instruction in the measurement of volume. Likewise, it was indicated that both experimental approaches were more effective in improving students' ability to work problems involving computations with metric measures of volume than was the approach with no related instruction.

Finally, it was indicated that the volume-only approach was more effective in increasing students' ability to make measurements of metric measures of volume than was either the computational approach or the approach involving no related instruction.

In summary, it appears that it is not necessary to proceed sequentially from length, through square and cubic units in teaching metric units of volume. However, the study did not indicate that the teaching of area necessarily interferes with the teaching of volume. It was suggested that the similarities between area and volume might cause difficulties in discriminating between them, a form of proactive inhibition. Thus, proceeding directly to volume, neglecting the concept of area, at least in terms of teaching the concept of volume, does not seem unwarranted. Achievement, using this technique, may be enhanced by offering opportunities for greater amounts of "hands-on" measurement.

ABSTRACTOR'S ANALYSIS

It would seem reasonable that the sequential approach to teaching concepts, particularly those in mathematics, is certainly consistent with logic and can be defended by intuition. Yet the investigators cite references that suggest that there may be some disagreement. As indicated in this study, there are those who believe that concepts dealing with three-dimensional measurement (volume) in the metric system need not necessarily be preceded with learning experiences dealing with one dimension (length) and two dimensions (area). Whether or not this belief applied also to teaching three-dimensional measurements in the English system is not mentioned by the investigators. However, since the belief in the logic of the sequential concept with respect to teaching volume units of the metric system is questioned, apparently based on proactive inhibition, there is sufficient reason to investigate the matter. One can, of course, question as to when it becomes sensible to teach area units of the metric system since it can hardly be suggested they be ignored.

The experimental design used--the randomized control-group posttest-only design described by VanDalen and Meyer--is standard for studies such as this. However, in the section of the article entitled "Experimental Design," the reference to VanDalen and Meyer is dated 1962, whereas in the list of references at the end, the reference is dated 1972. This can best be described as sloppy editing, either on the part of the investigators or those who publish the journal. Also in the same section there are two statements that are subject to challenge. The first statement dealing with the absence of a pretest is, "A pretest was not used since randomization techniques allow the researcher to assume that the groups are equal at the time of assignment." Such a statement is rather categorical. Randomization certainly helps to produce equality but it will not assure equality. For example, the alpha coefficient, that will be discussed later and which was used to assess the reliability of the total Test of Volume Units of the Metric System, is designed to compensate for the lack of equality that may occur with one randomization in a split-half technique, specifically, odds versus evens. This abstractor would have been much more comfortable with a pretest and a comparison of gains particularly since the analysis involved, in addition to total scores on 52 items,

subscores based on 20 items dealing with estimation, 20 with computation and 12 with measuring ability. These are not many items and there needs to be a better assurance of equality of groups when so few test items are involved.

The second statement is "Since the posttest-only control group design allows for extension with other groups and other treatments (Campbell and Stanley, 1969), it was particularly suitable to this study." The abstractor does not understand what this statement implies and neither do two of his colleagues who teach statistics and with whom he consulted.

There is some question about the way in which the control group was used. The abstractor and the two colleagues were of the opinion that it was a trivial use of a control group--hardly more than a placebo. The testing of the control group (C) that received no metric instruction amounted to little more than a pretest, and it may be noted that, while the accomplishments were less than those of the two experimental groups, the students were not at "ground zero." In effect, the scores of the experimental groups were those from posttests that were compared with the scores of the control group that were those from a pretest. It might have been more appropriate to have divided the 173 subjects into two groups and used a typical two-tailed design.

The investigators indicated that "each of the approaches involved three and one-half hours of instructional time and one additional hour for testing." However, nothing was said about the total time elapsed between the initiation of the instruction and the end of the testing period. Neither is there information concerning the identity of the instructors nor are specific examples given of the activities in the slide-tape programs. This does not suggest that the slide-tape programs were of low quality, but their merits must be taken on faith.

The validity of the Test of Volume Units of the Metric System, the instrument used in the study, is open to some question. The article states that "the validity of the test was established by having 12 selected authorities evaluate the individual items on agreement with the answer, approval of the item, clarity of the item, and suitability of

phrasing. Any of the items that were not approved by more than six of the judges were eliminated from the test." No evidence is provided concerning the criteria for being a "selected authority." Also an item could be retained if seven approved and five did not approve. One could look with askance at an item that five "authorities" rejected. It might have been helpful if some data had been provided concerning the extent of approval, or lack thereof, of the individual items.

The reliability of the Test was assessed using the alpha coefficient that is determined with a Fortran program that enables one to compute all possible split-half coefficients of correlation for a test and, in a sense, "average" them. The alpha coefficient thus computed for the total test of 52 multiple-choice items was 0.90 and is within the realm of respectability. It is stated that "Parts I [computation], II [estimation], and III [measurement] had reliability coefficients of 0.76, 0.86 and 0.73, respectively." There is no indication of the technique used to compute the reliability coefficients on these subscales so it is a matter of conjecture as to whether these are alpha coefficients. It may be noted that the reliability coefficients for Parts I (0.76) and II (0.73) are marginal.

Although there are innumerable research studies in the fields of mathematics and science education in which various teaching methodologies are compared, there is a dearth of studies dealing with the specific topic addressed here. Thus, within the limitations that have been mentioned in this abstract, this study certainly makes a contribution. It does point out that one apparently does not have to teach or study units of area in the metric system before teaching or studying units of volume. Whether this conclusion is a consequential consideration in mathematics education can only be ascertained in the classroom. Certainly, the techniques in any future studies related to this topic should take into account the reservations that have been mentioned. Also, since one can hardly ignore "units of area in the metric system," research should be undertaken to determine when the topic may best be dealt with. And, as indicated by the authors, "a logical extension of this research would involve an investigation of the interaction of area concepts and volume concepts" and also "a reconsideration of 'logical' approaches to teaching concepts of measurement."

Kauchak, D.; P. Eggen, and S. Kirk. "The Effect of Cue Specificity on Learning from Graphical Materials in Science." Journal of Research in Science Teaching, 15(6): 499-503, 1978.

Descriptors--Achievement; Ability; *Cues; Elementary Education; Elementary School Science; *Graphs; *Instruction; *Learning; Science Education; *Stimulus Devices; Visual Aids

Expanded abstract and analysis prepared especially for I.S.E. by David R. Stevenson, Truro, Nova Scotia.

Purpose

The research investigated the effectiveness of using specific questions to direct elementary school students to seek out information from graphs that were used to record results of experiments. No hypothesis is stated.

Rationale

Science texts present information in various forms for student attention. The extent to which the data are processed by the reader is open to study, especially for younger subjects. In particular, the use of graphs within or adjacent to print is questionable as an effective format.

The researchers cite studies that have explored the topic and they point out the variations in findings. The proportion of negative conclusions from past investigations is not overlooked. On the other hand, the usefulness of direct questions, or expressions, that cue the reader to data characteristics within graphs has not been completely explored. Positive results are reported for previous studies in which specific questions triggered searching for answers in prose material. The work on mathemagenic cues has not produced results that are as clear.

Research Design and Procedures

The objective of the research was to investigate the effects of cue specificity, grade level and ability level (independent variables) on acquisition of science content. Subjects all read passages describing experiments on plant growth, with the results of the experiments presented after the text in the form of graphs. No written commentary tied the experimental procedures to the results.

Subjects were randomly assigned to three treatment groups. The first group (Specific Cues) received a specific question about the graphed data. The second group (General Cues) was given a more general request about the data. The third group (Control) was asked to give notice to the graph but was not given a request for general or specific information. All groups received the instruction between the written text and the graph.

The researchers describe the subjects as 143 students of whom 45 are fourth graders, 40 are fifth graders, and 38 are in sixth grade. All are said to reflect a middle-class socioeconomic background. The subjects were divided into two groups by ability, using reading achievement scores.

Subjects read the passages and then completed a 20-item instrument containing three kinds of questions. Cued Questions measured ability to recall information that was directly linked to a specific text cue; Non-Cued Questions measured recall of incidental data; and, Generalizing Questions measured ability to identify generalizations based on graphed data. The scores were kept separately for each kind of question and a total score was computed. Total test reliability, using the Kuder-Richardson 21, was 0.73.

The scores from the 20-item instrument were subjected to three-way analysis of variance with treatment (i.e., reading passages with cues), grade levels and ability levels in a $3 \times 3 \times 2$ factorial design.

Findings

The results are summarized:

Factor	Scores Showing Significant Differences	Highest Group
Treatment	Total Scores	Specific Cues
	Cued Questions	Specific Cues
Grade Level	Total Scores	Grades 6, 5
Ability Level	Total Scores	High Ability
	Cued Questions	High Ability
	Non-Cued Questions	High Ability

No two- or three-way interactions produced significant results.

The analysis of differences within categories was done with Tukey's HDS using q. All results reported as significant were less than 0.01 with the exception of Grade Level differences (0.05).

Interpretations

The researchers feel that the study showed that textual cues can significantly increase the amount of information to be gained from graphical material. That the only significant treatment subtest was Cued Questions suggests to the researchers that the textual cues only aided learning of direct information as compared to a general scanning for incidental learning. They feel the results cast doubt on the use of broad, non-specific questions.

The researchers suggest two explanations for the findings. The placement of the cue before the graph causes subjects to search for specific information requested by the cue, rather than to peruse the data for incidental information. Also, it is harder to control scanning motions with graphs as compared to print, and subjects could skip over uncued data.

Differences noted due to grade and ability levels suggest development of conceptualization with age, and attention to graph materials seems intuitively associated with ability.

The authors indicate that teaching of upper elementary science may be improved by offering students specific cues so that learning from graphs may be more effective. They feel the findings are significant in light of earlier research that showed the limited value of graphs as learning tools.

ABSTRACTOR'S ANALYSIS

Science education has gathered a large and respectable following of researchers who explore to the edges of the subject. From time to time research pushes beyond the boundaries of the subject and studies are reported within journals of science education even though the topic is of peripheral interest at best.

Kauchak, Eggen and Kirk present an investigation within science education of a topic that may be of limited use to the science educator. Even if the concerns raised by the research are resolved and a clear pattern emerges for use of graphs and for cues to informational retrieval, science education might well ignore the issue. For science education has been stressing, and (to give a value judgment) should continue to stress, exploration of events, recording of results and discussion of possible reasons for the patterns discovered. An argument in favor of teaching for interpretation of graphed information must be lost based on past research and the results from the present study.

The study itself raises several questions for the reader. No hypothesis is stated, and it should be wondered whether or not the researchers were clear about the results they expected. The research quoted in support of the study suggests conflicts that make the topic more open to personal interpretations than a researcher may wish. A statement of expectation would clarify that point.

The subjects in the study are sketchily described and more information may be helpful. Even so, the categorization of the students should present few concerns for a study of this kind. Any school may be a suitable source for subjects and the number used may be of marginal interest, provided small numbers are avoided. That the subjects reflected a middle-class background and were separated into ability groups by unknown measures limits replication possibilities.

Procedures followed in the study are unclear. The reader may understand that three treatment groups were created, each with subjects in the three grades and of both sexes and both abilities. The time span over which the treatment was given and the nature of the 20-item instrument that the subjects completed are not told to the reader. One may conclude that one setting, of limited time, was used. If so, the results may be criticized as a one-time occurrence rather than part of a pattern recognized as part of an ongoing study of student characteristics. No teaching about graphs is assumed or stated. Nor are students described as having experience with the type of research pattern in one grade or another. One might wonder if there were, in fact, adequate treatments to warrant a report.

The researchers state that a "Table of Specifications" was used to design the 20-item instrument. It is not clear what the table is or how it was used. The number of questions in each category, the wording used, the type of answer sheet, the total score possible, the range of possible correct answers, the readability of questions, and other concerns could be described.

The researchers indicate that follow-up studies are being conducted. The limited extent of the research being analyzed here does not auger well for future findings from equally limited samples and research settings.

Boulanger, F. David. "Instruction and Science Learning: A Quantitative Synthesis." Journal of Research in Science Teaching, 18(4):311-327, 1981.

Descriptors: *Academic Achievement, Elementary School Science, Elementary-Secondary Education, *Instructional Innovation, *Learning, Science Education, *Science Instruction, *Secondary School Science.

Expanded abstract and analysis prepared especially for I.S.E. by Gerald G. Neufeld, Brandon University, Canada

Purpose

The purpose of this review was to use the techniques of meta-analysis to synthesize the results of published research on the quality and quantity of instruction. Only studies dealing with instruction in science in grades 6-12 that were published during the 1963-78 period were considered.

Rationale

Research on the quality of instruction is extensive, diverse, and the results of individual studies are often inconclusive. Previous reviews of the research in science instruction have tended to be long narratives that provide little basis for objective comparisons and accumulation of results. The technique of meta-analysis developed by Glass (1978) provides a more quantitative and objective way of reviewing the research in an area.

Research Design and Procedures

The studies included in this quantitative synthesis were located by a literature search. The primary source of citations was the collection of ERIC science education bibliographies and annual reviews. In addition, the appropriate volumes of the Journal of Research in Science Teaching and Science Education were scanned. A total of 137 published studies relating to the quality of instruction and 3 relating to the quantity of instruction (2 published and 1 dissertation) were found. Of these, the 95 that involved the experimental manipulation of an instructional situation were analyzed further.

The multi-dimensional concept, quaity of instruction, was not defined prior to the literature search. The component variables were determined by categorizing and counting the independent variables used in the 95 experimental studies. A total of 43 independent variables, such as advance organizer, group size, inductive vs deductive, questioning level, and teacher background, were identified. Only those six independent variables, or clusters of closely related independent variables, that were used in at least five experimental studies, were included in the quantitative synthesis of finding. The six clusters chosen had been used in a total of 51 studies. The clusters and their component independent variables were:

Cluster	Component Variables	Studies
Preinstructional Strategies	Advance organizers	4
	Behavioral Objectives	5
	Set Induction	2
Directness of Instruction	Direct vs Nondirect	7
	Indirect/Direct Ratio	2
Inductive/Deductive Strategies	Same as cluster	9
Training in Scientific Thinking	Training in Logical Operations	7
	Training in Science Processes	2
Structure in the Verbal Verbal Content of Materials	Same as cluster	5
Realism or Concreteness in Adjunct Materials	Same as cluster	9

The study variables (characteristics) were coded using a scheme developed prior to the selection of the studies. The scheme was refined as coding progressed. Each comparison of treatment means was coded according to about

40 study variables: dependent measure type, origin, and reliability; subject, grade, sex, SES, etc. A small sample of the studies was coded independently by two raters. The inter-rater agreement on the ratings of the 40 study variables was 90 percent. Many studies did not report these variables or the values were constant across studies. As a result, only those study variables that were adequately reported and had nonconstant values were considered in the analysis.

The results of the studies were standardized using the techniques proposed by Glass (1976, 1978). This involved standardizing the differences between the treatment and control group means by calculating the effect size (difference between means divided by the standard deviation of the control group).

Each dependent variable in each study was placed in one of four categories:

1. Factual learning (retention test).
2. Conceptual learning (concept, process, logical operations, critical thinking, or standardized achievement test).
3. Attitudinal learning.
4. Laboratory performance test.

Because there was a great deal of overlap in content between the factual learning category and the conceptual learning category and the size and directionality of the observed differences were similar, these two categories were combined into a single category named cognitive outcome.

Methodological flaws were examined and coded as either "a potential threat" or "adequately minimized." The flaws examined were: treatment reliability, statistical power, error rate, maturation, history, selection bias, compensating or differential incentives, generalizability, and mortality. A simple sum of these ratings gave an overall index of the quality of the research design.

Due to the range in the number of comparisons in different studies (1 to 11), and the limited number of studies in any one cluster, the median effect size from each study was used in the outcome category. The 51 quality of instruction studies yielded 160 comparisons which reduced to 69 median comparisons.

Findings

Preinstructional Strategies: The eight studies in this category involved 1024 subjects. The mean cognitive effect size (1.03) was significant (p .05) and favorable to the use of a preinstructional strategy.

Indirectness of instruction: The eight studies in this category involved 1135 subjects. The mean cognitive effect size (0.11) was not statistically significant but was favorable to the use of an indirect approach. The two studies in this category that reported attitudinal findings almost exactly cancelled each other (mean effect size of 0.002).

Inductive vs deductive strategies: There were nine studies in this category. The mean cognitive effect size (-.22) was not statistically significant but was favorable to the use of a deductive strategy. Only one study reported attitudinal outcomes. It favored the deductive approach and was not significant.

Training in scientific thinking: The eight studies in this category involved 716 subjects. The mean cognitive effect size (0.89) was significant (p .05) and favorable to training students in logical operations or science processes.

Structure in the verbal content of materials: There were five studies in this category. The mean cognitive effect size (0.74) was significant (p .05) and favorable to the higher structure treatment. One study reported the results of a lab performance test. The results were significant (mean effect size of 1.364) and favored the higher structure treatment.

Realism or concreteness of adjunct materials: The nine studies in this category involved 512 subjects. The mean cognitive effect size (0.58) was significant (p .05) and favorable to the use of realistic and concrete adjunct instructional materials. One study reported the results of a laboratory performance test. The results were significant (mean effect size of 1.540) and favored more realism or concreteness. This study also reported an attitudinal outcome. The results were significant (mean effect size of -0.848) and favored an expository rather than a lab approach.

Interpretations

On the basis of his analysis the author concluded that:

1. Preinstructional strategies, especially the use of behavioral objectives and set induction, can improve student conceptual learning when used with other instructional activities by classroom teachers.

2. There was no difference in the general effectiveness of nondirect or indirect instruction and direct instruction in regard to cognitive outcomes.

3. Although the cognitive outcome results slightly favored a deductive rather than an inductive approach, no firm general conclusion could be drawn regarding their relative effectiveness.

4. Deductive or direct instruction tends to be more effective in terms of cognitive outcomes with students in required courses in grades 6-8, while indirect, nondirect, or inductive instruction was more effective with students in elective courses in grades 10-12.

5. Training in scientific thinking, especially in the use of logical operations, is effective in terms of cognitive outcomes when conducted on an individual basis by a special teacher.

6. More highly structured verbal context in printed or audio materials is more effective in promoting cognitive learning than less structured content.

7. Greater realism or concreteness in adjunct materials resulted in greater cognitive learning.

There were too few studies that reported attitudinal or laboratory outcomes to draw any general conclusions about what aspects of the quality of instruction have favorable or unfavorable effects.

When all the studies were considered as a whole, several trends were evident:

1. Most of the studies showed a result favorable to the experimental treatment. It appears that systematic instructional innovation in instruction resulted in significantly positive improvements over the norm or "traditional" practice.

2. Studies that used published tests to measure instructional outcomes tended to yield larger effect sizes than those using teacher or experimenter-made measures.

3. As the number of design flaws in a study diminishes, the difference between the experimental and control group means increases.

The author concludes his synthesis with a number of recommendations for researchers regarding: the need for planned variation when replicating research studies, the need for measuring and reporting study variables, and the need for improved research design and analysis.

ABSTRACTOR'S ANALYSIS

Reviews of the research in a field serve several useful functions. For the non-expert they provide a quick overview of the field. For the researcher they provide a quick overview of the work of others, indicate gaps in the research, and serve as the basis for hypotheses and theory generation. For graduate students they provide a quick introduction to an area, a source of potential research topics, and a list of relevant citations.

Reviewing educational research is particularly difficult because the research is so extensive and diverse and the results are often inconclusive or conflicting. In addition, the methodology and procedures for conducting such reviews is not well developed (Jackson, 1980). As a result, reviewing educational research has been more of an art than a science and the reviews produced have been rather qualitative and subjective.

Meta-analysis as proposed by Glass (1976, 1978), provides a more quantitative and objective way of conducting an integrative review. It has generated a great deal of interest in the research community, and many papers have been published on the theory and practice of meta-analysis (Glass, et al., 1980; Glass, 1982; Haladyna, 1981; Hattie and Hansford, 1982; Hedges, 1981(b); Jackson, 1980; Kulik, J., 1981; McGaw and Glass, 1980; Stock et al., 1982). Numerous recent reviews of general educational research have made use of this methodology (Cohen, 1981(a); Cohen, 1981(b); Hattie and Hansford, 1982; Hetzel et al., 1980; Iverson and Levy, 1982; Kozlow, 1978; Kulik, C., 1981; Kulik, J., et al., 1980(a); Kulik, J., et al., 1980(b); Luiten et al., 1979; Readence and Moore, 1981; Redfield and Rousseau, 1981; Smith and Glass, 1980; Strube, 1981).

This review was one of the first to apply this new methodology to research in science education. It appears that meta-analysis is rapidly gaining favor in the science education community because most recent research reviews have used this technique (Anderson, et al., 1981; Bredderman, 1982; Eng., et al., 1982; Haladyna and Shaughnessy, 1982; Kahl et al., 1982; Sweitzer, 1982; Weinstein et al., 1982).

Jackson (1980) has conceptualized the methodology of an integrative research review as involving six basic tasks: (1) selecting the questions or hypotheses for the review, (2) sampling the research studies that are to be reviewed, (3) representing the characteristics of the studies and their findings, (4) analyzing the findings, (5) interpreting the results, and (6) reporting the review. This analysis of the Boulanger review will consider each of these points.

(1) Selecting the questions or hypotheses for the review

The author chose rather broad questions as the basis for his review: what factors relating to the quality and quantity of instruction effect the cognitive, affective, and psychomotor (lab skills) outcomes of instruction? Since a review of all the research related to these broad questions would have been unmanageable, the author had to narrow the scope of his review. He chose to restrict it to research in science education.

In view of the broad questions addressed by the review and the study clusters actually considered (preinstructional strategies, directness of instruction, inductive vs deductive strategies, realism in adjunct materials, etc.), this way of restricting the scope of the review seems unfortunate. Instruction in science has some unique features, but one would hope that science education researchers are not so ingrown and self-centered that they would ignore good research in other areas of education - especially when considering broad educational questions.

By restricting his review to science education research the author had relatively few studies with which to work. For example, his study cluster named preinstructional strategies included four studies involving advance organizers, five involving behavioral objectives, and two involving set induction. Even by clustering these studies, the author had insufficient numbers to be able to tease out any meaningful relationships between the outcomes and the 40 study variables. In contrast, two other reviewers interested in this area chose a different way of restricting the scope of their meta-analytic reviews - they focused only on the effects of advance organizers. Kozlow (1978) located a total of 77 relevant studies and Luiten et al (1979) located 135. Because these reviewers had not restricted the scope of their reviews in an arbitrary and artificial way, they had a larger data base to work with and were able to find meaningful relationships between their study variables and the learning outcomes.

(2) Sampling the research studies that are to be reviewed

The author is to be commended for reporting the indexes and journals he searched to locate the studies included in his review. This necessary detail is often omitted and the reader is left wondering about how thorough the search was. As the volume of research continues to grow and reviewers become more dependent on indexes, bibliographies, and computer searches, reviewers should be encouraged to report not only the data bases searched, but also the actual search terms used to conduct the literature search.

For this review the author searched the ERIC science education bibliographies and annual reviews and scanned the appropriate volumes of the Journal of

Research in Science Teaching and Science Education. In view of the broad issues concerned and the relatively few studies located, it seems surprising that unpublished dissertations were largely ignored. The reader is left wondering whether the studies reviewed were a representative sampling of even the full set of existing science education research on these topics.

(3) Representing the characteristics of the primary studies

The author did a good job of describing the findings of the studies he reviewed. Table III was very effective in summarizing the results of the studies. It indicated the number of studies in each category, the number of positive effects, the number of significant positive and negative effects, and the combined effect sizes and the relevant confidences intervals.

In his narrative description of each cluster, the author briefly indicated some of the characteristics of the studies. This section of the review would have been more useful to the reader if the individual studies were described more fully.

(4) Analyzing the primary studies

The analysis of the results of the primary studies using Glass's meta-analytic techniques appears to have been done competently. The author's technique of rating the strengths of the research designs appears somewhat crude. However, a statistically-defensible and objective method for weighting studies to be included in an analysis on the basis of the strength of their research designs or their sample sizes has not yet been developed.

The author's narrative analysis of each study could have been expanded and more detail provided. However, the fact that the project final report (ERIC ED 197939) contains an abstract of each study, a code book, a code sheet, and a table of coded values, means that an interested reader has ready access to more detailed information.

(5) Interpreting the results

Although the author did draw some conclusions that can be interpreted as suggestions for improved instructional practice, he did not relate his finding to any theoretical framework or model. The review would have been much more valuable if the author had included a test of some existing educational theory or had used the results as the basis for proposing a new theory.

The author's conclusion regarding the general effectiveness of systematic innovation in instruction is suspect. The observation that more studies showed a statistically significant positive effect than showed a negative effect (23 vs [illegible] that the mean cognitive effect size was significantly positive effect than showed a negative effect (23 vs 3) and that the mean cognitive effect size was significantly positive (0.55) does not necessarily mean that any deviation from the norm or "traditional" practice will have a positive effect on learning. These results may be readily explained in terms of the biases of researchers and school administrators. Almost any experimental treatment would be aborted, either by the researcher or by the school administration, as soon as there is any evidence that it is having a negative effect on the students' learning. In view of these biases it is amazing that even three research studies showing negative results reached the journals.

(6) Reporting the review

In general the review was well written and presented. As previously indicated, additional detail regarding the search techniques and the characteristics and analysis of the primary studies would have been helpful.

The writing of research reviews is an important task. Potential reviewers should keep a number of points in mind:

a) The review should be carefully focused so that it taps all the relevant studies. Arbitrarily restricting the search to a sub-field such as science education is not appropriate when addressing broad educational issues.

b) The reviewer should carefully detail the indexes and search terms used to locate the primary studies. This enables another reviewer to expand the search without unnecessary duplication of effort.

c) Techniques such as Glass's meta-analysis are useful for combining the results of different studies but they are not a panacea. A blend of objective and subjective methods is still required.

d) Whenever possible a review should serve as the basis for theory testing or generation or for identifying guidelines for the educational practitioner.

e) It appears that the social science community has finally become interested in the methodology of conducting integrative reviews. Those articles dealing with meta-analysis have been mentioned. Those dealing with alternative approaches include Cooper (1981), Light and Smith (1971), Rosenthal (1978), Schmidt et al (1979), and Yager (1983). Potential reviewers should keep abreast of the literature in this rapidly developing field.

REFERENCES

Anderson, Ronald D. and Others. "The Major Questions Addressed by the Extant Science Education Research: A Map for Meta-Analysis." Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Ellenville, NY, 6p, ED 202 739, 1981.

Bredderman, Ted. "The Effects of Activity-based Elementary Science Programs on Student Outcomes and Classroom Practices: A Meta-Analysis of Controlled Studies." New York State University System, Albany, 86p, ED 216 870, 1982.

Cohen, Peter A. "Educational Outcomes of Tutoring: A Research Synthesis." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, 17p, ED 204 416, 1981(a).

Cohen, Peter A. "Using Student Ratings to Improve Instruction: A Synthesis of Research Findings." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, 36p, ED 200 647, 1981(b).

Cooper, Harris M. "Scientific Guidelines for Conducting Integrative Research Reviews." Review of Educational Research, 52(2):291-302, 1982.

Eng, Judith and Others. "Review and Analysis of Reports of Science Inservice Projects: Recommendations for the Future." Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Chicago, IL, 16p, ED 216 883, 1982.

Glass, Gene V. "Primary, Secondary, and Meta-Analysis of Research." Educational Researcher, 5(10):3-8, Nov 1976.

Glass, Gene V. "Integrating Findings: The Meta-Analysis of Research." in Shulman, L. (Ed.) Review of Research in Education, Itasca, Illinois: Peacock Publishing, 1978.

Glass, Gene V. and Others. "Integration of Research Studies: Meta-analysis of Research. Methods of Integrative Analysis: Final Report." Colorado University, Boulder, Laboratory of Educational Research, 340p, ED 208 003, 1980.

Glass, Gene V. "Meta-Analysis: An Approach to the Synthesis of Research Results." Journal of Research in Science Teaching, 19(2):93-112, 1982.

Haladyna, Tom. "A Common Metric for Integrating Research Findings." Research report, 21 p, ED 202 873, 1981.

Haladyna, Tom and Joan Shaughnessy. "Attitudes toward Science: A Quantitative Synthesis." Science Education, 66(4):547-63, 1982.

Hattie, J.A. and B.C. Hansford. "Self Measures and Achievement: Comparing a Traditional Review of Literature with Meta-Analysis." Australian Journal of Education, 26(1):71-75, 1982.

Hedges, Larry V. "Statistical Aspects of Effect Size Estimation." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, 40p, ED 208 024, 1981(a).

Hedges, Larry V. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." Journal of Educational Statistics, 6(2):107-28, 1981(b).

Hetzel, Donna C. and Others. "A Quantitative Synthesis of the Effects of Open Education." Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA, 39p, ED 191 902, 1980.

Iverson, Barbera K. and Susan R. Levy. "Using Meta-Analysis in Health Education Research." Journal of School Health, 52(4):234-39, 1982.

Jackson, Gregg B. "Methods for Reviewing and Integrating Research in the Social Sciences." Final Report to the National Science Foundation, ED 197 939, 1978.

Jackson, Gregg B. "Methods for Integrative Reviews." Review of Educational Research, 50(3):438-60, 1980.

Kahl, Stuart R. and Others. "Sex-Related Differences in Pre-college Science: Findings of the Science Meta-Analysis Project." Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY, 18p, ED 216 909, 1982.

Kozlow, Michael James. "A Meta-Analysis of Selected Advance Organizer Research Reports from 1960-1977." Ph.D. Dissertation, Ohio State University, 306p, ED 161 755, 1978.

Kulik, Chan-Lin C. "Effects of Ability Grouping on Secondary School Students." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, 15p, ED 204 417, 1981.

Kulik, James A. "Integrating Findings from Different Levels of Instruction." Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, 18p, ED 208 040, 1981.

Kulik, James A. and Others. "Effectiveness of Computer-based College Teaching: A Meta-analysis of Findings." Review of Educational Research, 50(4):525-44, 1980(a).

Kulik, James A. and Others. "Effectiveness of Programmed Instruction in Higher Education: A Meta-Analysis of Findings." Educational Evaluation and Policy Analysis, 2(6):51-64, 1980(b).

Light, R.J. "Capitalizing on Variation: How Conflicting Research Findings can be Helpful for Policy." Educational Researcher, 8(9):7-14, 1979.

Light, R.J. and P.V. Smith. "Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies." Harvard Educational Review, 41:429-471, 1971.

Luiten, John and Others. "The Advance Organizer: A Review of Research using Glass's Technique of Meta-Analysis." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, 16p, ED 171 803, 1979.

McGaw, Barry and Gene V. Glass. "Choice of the Metric for Effect Size in Meta-Analysis." American Educational Research Journal, 17(3):325-37, 1980.

Readence, John E. and David W. Moore. "A Meta-analytic Review of the Effects of Adjunct Pictures on Reading Comprehension." Psychology in the Schools, 18(2):218-24, 1981.

Redfield, Doris L. and Elaine Waldman Rousseau. "A Meta-Analysis of Experimental Research on Teacher Questioning Behavior." Review of Educational Research, 51(2):237-45, 1981.

Rosenthal, R. "Combining Results of Independent Studies." Psychological Bulletin, 85: 185-193, 1978.

Schmidt, F.L. and Others. "Further Tests of the Schmidt-Hunter Bayesian Validity Generalization Procedure." Personnel Psychology, 32:257-276, 1979.

Smith, Mary Lee and Gene V. Glass. "Meta-Analysis of Research on Class Size and its Relationship to Attitudes and Instruction." American Educational Research Journal, 17(4):419-33, 1980.

Stock, William A. and Others. "Rigor in Data Analysis: A Case Study of Reliability in Meta-Analysis." Educational Researcher, 11(6):10-14, 1982.

Strube, Michael J. "Meta-Analysis and Cross-Cultural Comparisons: Sex Differences in Child Competitiveness." Journal of Cross-Cultural Psychology, 12(a):3-20, 1981.

Sweitzer, Gary L. "A Meta-Analysis of Research on Preservice and Inservice Science Teacher Education Practices Designed to Produce Outcomes Associated with Inquiry Strategy." Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Chicago, IL, 17p, ED 219 231, 1982.

Weinstein, Thomas and Others. "Science Curriculum Effects in High School: A Quantitative Synthesis." Journal of Research in Science Teaching, 19(6):511-22, 1982.

Yager, Robert E. "Factors Involved in Qualitative Synthesis: A New Focus for Research in Science Education." Journal of Research in Science Teaching, 19(5):337-350, 1982.

Sallam, Sallam and Gerald Krockover, "The Effect of Inquiry Geoscience Instruction on the Attitudes and Creativity of Junior High/Middle School Science Teachers," School Science and Mathematics, 82(4):279-283, 1982.

Descriptors--*Earth Science, *Educational Research; Elementary Secondary Education; *Inservice Teacher Education; Instruction; Learning Theories; *Science Education; *Science Instruction; *Teacher Education

Expanded abstract and analysis prepared especially for I.S.E. by William R. Brown, Old Dominion University

Purpose

How did participants' attitudes and creativity change as a result of participation in The Geosciences Today program? The question to be investigated was not further delineated by hypotheses. Inferred hypotheses were that participants would: (1) increase their content background concerning selected topics, (2) enhance their inductive thinking, and (3) improve their attitude toward science and science teaching.

Rationale

Six articles were cited that dealt with the use of inquiry procedures to improve the preparation of earth science teachers. An assumption was made that if teachers-in-training were taught using certain methods, that those methods would be used in their instruction of students in grades 5 through 9. The Geoscience Today (TGT) program was sponsored by the National Science Foundation. Evidently this report was part of the grant evaluation scheme.

Research Design and Procedure

The sample (and population) for the study was the twenty-seven selected teacher participants in the 32-week program.

The variate was the methodology used by the instructor. This methodology was described as inquiry using questioning techniques, field trips, and open-end activities. The instructor listened, reacted, provided materials, suggested, and coordinated discussions. Only one "method" was reported.

The participants' attitudes towards teaching science and creativity (problem-solving) were the criterion variables. Attitude was measured using the Bratt Attitude Test (BAT). The investigators report an "acceptable construct validity" and a test-retest reliability of 0.87. The New Uses Creativity Test (NUCT) had a construct validity of 0.57 and an interclass correlation reliability index of 0.65. Both tests were administered as pretests and as posttests (after 32 weeks). The BAT was also given after 16 weeks.

The design of the study may be represented as: R O_1 X O_2 O_3 when R suggests no randomination, O_1 is the two pretests, X is the single treatment, O_2 is the 16-week administration of the BAT, and O_3 is the two posttests given at the end of the 32-week program.

One-way analysis of variance (ANOVA) was used to indicate differences between pretest and posttest scores.

Findings

The O_1 to O_2 change for the BAT intellectual subscale was significant with an F value of +10.93 (df = 65). The O_1 to O_3 change for the NUCT was F = +63.80 (df = 42) which was significant.

Interpretations

The investigators implied that TGT program was effective. Positive attitudes toward science and teaching of science were achieved within 16 weeks for intellectual (knowledge-based) attitudes and by 32 weeks for humanistic (interaction of student and teacher in a learning environment) attitudes. Creativity was also promoted. The investigators concluded that TGT program was successful in fostering positive attitudes and creativity only if the program was 16-32 weeks in length.

ABSTRACTOR'S ANALYSIS

The title of the report includes both variables and the target audience. However, the use of "The Effect of" is inappropriate. This terminology should be reserved for experimental studies. No randomization or control was used. Cause and effect can only be inferred from a strong experimental design. A more appropriate title would be "The Relationship of..." Clear, concise, and precise titles aid in a literature search by providing the type of study, variables, and audience with minimum words (Baker, 1972). Since the variate and the criterion variables were included in the title, it would also have been helpful to have these specified in a design section of the written report.

No formal hypotheses were stated. In the "purpose of the study" section of the report it is stated that one of the main objectives of TGT program was to provide background and experience with recent geoscience topics. What happened to this component of the program?

The terms inductive thinking, creativity, and problem-solving are used interchangeably. A precise definition of the "creativity" variable would be helpful.

The six studies cited provide little substantive support for the contextual framework of the study. How will enhancing the teachers' creativity and improving their attitude toward science and science teaching affect children? Can we assume that teachers of grades 5-9 will teach children in the same style that they have experienced in TGT? Should they adopt the college instructor's methods? Additional support for links between creativity, attitude and teacher effectiveness are necessary to support this study.

The TGT was supported by an NSF grant. It should be explicitly stated whether or not this article is part of the NSF report.

Little mention is made of how the 27 participants were selected. Since the average number of earth science content hours previously completed by the participants was eight semester hours, it can be inferred that TGT program was really a retraining program. Were the participants volunteers?

The variate of the study was "inquiry" instruction. Since this term covers a multitude of strategies, a precise definition of this variable is desirable. For example, what kinds of questioning techniques were used? What happened on the field trips? How were participants evaluated? What exactly occurred during the 32 weeks of instruction? Was instruction daily or once a week? Were teacher participants fulltime students or part-time students and fulltime teachers? Why 32 weeks?

The investigators state that factor analysis revealed that the BAT had an "acceptable" construct validity. How is acceptable defined?

The BAT was given at 16- and 32-week points. Why was the NUCT only given at 32 weeks? A rationale for the design would help to clarify this decision.

ANOVA is a suitable technique to assess pre-to-posttest changes. Were acceptable levels of significance stated in advance of the analysis? If so, what were they?

The investigators conclude their discussion with the statement "For if we truly want to have an impact upon the junior high/middle school science teaching in our schools, we must foster attitudinal and creative development among the teachers who teach these programs." Is this a generalization based on this study? Perhaps an entirely different instructional mode for retraining teachers would foster greater attitudinal and creative development!

Because of the numerous design errors, little positive contribution results from this report. The only conclusion that can be drawn is that for 27 people, who experienced an undefined instructional mode, TGT program changed attitudes and creativity. Why or how these variables changed cannot be inferred from the report. The effect of teacher changed attitudes and creativity on teacher effectiveness, as defined by student performance, is not dealt with although it is the "real" question.

REFERENCE

Baker, Robert and Richard Schutz, Instructional Product Research. Cincinnati: Van Nostrand Co., 1972, pp. 3-9.

Stallings, Everett S., Philip M. Astwood, John R. Carpenter, and Henry B. Fitzpatrick. "Effects of Two Contrasting Teaching Strategies in an Investigative Earth Science Course for Elementary-Education Majors." Journal of Geological Education, 29(2):76-82, 1981.

Descriptors--College Students; Earth Science; *Educational Methods; Educational Research; Geology; Higher Education; *Preservice Teacher Education; Science Education; Science Instruction

Expanded abstract and analysis prepared especially for I.S.E. by Gerald H. Krockover, Purdue University

Purpose

The purpose of this study was to investigate the effect of student-structured versus teacher-structured approaches to the teaching of earth science for elementary education majors (p.76).

Rationale

The rationale for this study is based upon evidence presented by a National Science Foundation study that indicates that, "pre-college science education has changed little over the past twenty years in spite of many, and often expensive, efforts to provide new teaching materials and curricula." Furthermore, additional studies indicate that students are not receiving sufficient instruction in science and that they are being exposed to, and affected by, teachers' negative attitudes toward science (p.76).

Research and Design Procedure

The study employed strategies similar to those reported by James A. Shymansky. However, two major differences should be noted between this study and Shymansky's. First, the age of the students was different and secondly, this study was unable to assign subjects randomly to either of the two treatment groups. Thus, this study utilized preformed groups leading to an arrangement referred to by Campbell and Stanley as a quasi-experimental design. Forty-three

subjects participated in the study with 24 assigned to one teaching approach and 19 assigned to the other. T-tests were performed on the age and grade point ratio and the results indicated that the two groups were reasonably similar in background. A coin toss decided which section would receive which approach.

Data collected for the subjects included: the Learning Condition Index (LCI) and the Science Curriculum Assessment System (SCAS) which examined the instructor's performance in classroom. This information was obtained to insure that there were in fact two distinct teaching environments. Two hypotheses were tested: 1) There is no measurable difference in the effects the Teacher-Structured Learning in Science (TSLS) and Student-Structured Learning in Science (SSLS) environments have on the students' concept of science and 2) There is no measurable difference in the effects the TSLS and SSLS environments have on the students' concept of the teaching of science (p. 79).

To test hypothesis one, a 12 question survey entitled, "Self-Perceptions in Science," was used. It was administered on a pre-post basis with the pre-test results compared to the post-test results utilizing a chi-square test. To test hypothesis two, a 30 question survey utilizing a five point Likert scale was used on a pre-post test basis and the results were subjected to an analysis via the chi-square test. To further test hypothesis two, each subject was asked to, "write an optional anonymous evaluation of the course" (p.80).

Findings

The findings for hypothesis number one indicated that there was a significant difference between scores in the SSLS section while there was no significant difference in scores in the TSLS section. This indicates that the SSLS environment has a significant effect on the students' concept of science. The effect can be considered a "favorable" one in that it encouraged the students to take a more active approach to science (p.80).

The findings for hypothesis number two indicated that the two different learning environments did not have measurably different effects on the students'

concept of science. As far as the course evaluations were concerned, most of the students in the TSLS section did not submit one. Those that were received from that section were quite short and contained opinions which would lead one to a neutral view. On the other hand, the situation in the SSLS section was just the reverse. Most of the students wrote evaluations which were fairly long and highly favorable (p.80). Thus, it appeared that students in the SSLS section enjoyed the "open" style of teaching and realized that they had benefited from it.

Interpretations

This study demonstrates that it is possible for an open and investigative college science course to affect a future teacher's approach to science. Furthermore, it has shown that something can be done to improve the present state of science teaching. College science courses can be constructed to leave the future teacher with a more positive attitude about science than that with which he started. Students also seemed to recognize the value of the "investigative" strategy (p.80-81).

ABSTRACTOR'S ANALYSIS

It was very difficult for this abstractor to identify and synthesize the major components of this "study". The article is written in a very disjointed and confusing manner as if it were written by the four authors independently and compiled by accident. Furthermore, there was no attempt to ascertain whether or not the use of TSLS and SSLS strategies are even appropriate for college students. The authors appeared to be willing to accept the notion that if tests and strategies are suitable for fifth grade students, they are also suitable for college students! The assignment of each group of students to one of the treatments is highly questionable along with the assessment instruments used. For example: how do they know whether or not the two groups were in fact equal in background and ability? Secondly, why weren't reliability and validity data collected for the tests used for this population? Third, what is the rationale for the statistical analysis used? Fourth, how can any credence be given to the use of, "optional anonymous evaluations"?

In conclusion, the authors state that, "students recognized the value of the 'investigative' strategy, but we did not find reason to believe the course would affect the way they will eventually choose to teach science" (p.81). Hopefully, science education research rests on a firmer foundation and studies such as this one will not have a measurable affect upon the way we choose to prepare our future elementary teachers.

REFERENCE

Shymansky, James A. et al. "A Study of Self-Perceptions Among Elementary School Students Exposed to Contrasting Teaching Strategies in Science." Science Education 58(3):331-342, July-Sept. 1974.

Dumpert, Klaus. "An Inquiry into the Use of Living Organisms in Biological Education in Western German Schools." European Journal of Science Education 1(3):339-346, 1979.

Descriptors--*Biology; *Educational Research; Elementary Secondary Education; Expenditures; Environmental Education; *Instructional Aids; Outdoor Education; *Science Instruction; Science Education

Expanded abstract and analysis prepared especially for I.S.E. by David H. Ost, California State College, Bakersfield.

Purpose

The intent of the study was to collect data representative of the Federal Republic of Germany concerning the extent and manner in which living organisms are used in the teaching of biology. Teacher attitudes and commitment concerning the use of living organisms was assessed. Support of the schools for such instruction was also considered.

Rationale

Biology educators have long contended that there is considerable educational value associated with the use of living organisms in the classroom. The notion that the science dealing with the study of life should somehow include living things is a common position taken by teacher educators. It is believed that there is an increasing need to familiarize the student with living organisms from his/her enviornment as cities increase in size and students are cornered in an urbanized environment.

The value of studying living things is frequently associated with teaching children how to care for plants and animals. The development of responsibility and positive attitudes about the enviornment are supposedly fostered. The "common knowledge" is that an "original encounter" between student and organism is an important contribution to genuine knowledge/experience.

Assuming that the above if not true, at least has basic merit, it was deemed useful to determine how much correlation existed between theory and practice. For example, to what extent do teachers endeavor to use living organisms in instruction? What are the supportive measures provided by the school?

Research Design and Procedure

A questionnaire study was conducted (year not provided) of biology teachers in FR Germany. A random sample of 650 was initially identified and communication made by sending the director of each school a questionnaire and a cover letter. It was requested that one of the school's biology teachers complete the questionnaire. Only 231 responses were obtained.

The specifics included on the questionnaire are not provided in the paper. Items are referred to in a general manner. Reference is made to such questions as:

1. (the number of schools) maintaining living organisms,
2. types of living organisms which are kept,
3. which individual science disciplines use living organisms in teaching,
4. the sources of the living organisms,
5. use of living organisms in instruction outside the classroom,
6. methods of instruction employed in connection with living organisms,
7. the value of living organisms in biology instruction,
8. reasons why living organisms are not used, and apparently
9. a series of questions related to school support provided for the use of living organisms.

Specific response data are not provided. Percentages of responses are included for some items. Apparently some statistical analysis was attempted but not elaborated any further than, ". . . no significant differences were found. . . ."

Findings

The results of the survey reported in the paper can be summarized as shown below.

1. Percent of schools maintaining organisms:

a.	single-cell	18%
b.	insects	18%
c.	molluscs	10%
d.	crabs	8%
e.	other invertebrates	5%
f.	fish	38%
g.	amphibian	22%
h.	reptiles	12%
i.	birds	15%
j.	mammals	18%
k.	flower plants	60%
l.	other plants	48%

2. Use made of living organisms in instruction:

a.	observation	46%
b.	demonstration	36%
c.	student experiments	25%
d.	dissection	7%
e.	did not use	8%

3. In answer to the questions about the value of living organisms in instruction, 52% of the teachers felt it was imperative, 42% responded that it was not necessary but useful, 1% felt it was of no particular value, and 5% did not respond.

4. Reasons given by teachers who do not keep living organisms in their school were:

a.	care and maintenance	59%
b.	lack of space	35%
c.	professional reasons (e.g., serve no purpose)	12%
d.	other	10%

5. No significant differences were found to exist among schools in communities of different sizes.

6. No significant differences were found to exist among schools about how living organisms can be used outside the classroom.

7. Teachers in larger schools tend to purchase organisms from local sources (pet stores, forests, etc.)

Interpretations

The investigator suggests that, although there is considerable discrepancy between pedagogical theory and practice concerning the use of live organisms in the classroom, many teachers do attend to living things in some manner. It is pointed out that the number is not large and that there are two possible explanations.

1. Teachers do not consider work with living organisms necessary.
2. Although teachers consider the use of living organisms a valuable component of instruction, other factors inhibit their use.

Although no definite conclusions are drawn, it is suggested that much of the inhibition is due to "unfavorable circumstance in actual school practice."

ABSTRACTOR'S ANALYSIS

A descriptive study such as this can be of interest in several ways. It is of general use to researchers to have access to data concerning classroom practice on topics such as the use of living organisms. However, the design of the particular investigations does not provide the reader with confidence in the data. The return of only 231 of 650 questionnaires provides considerable opportunity for error. The investigator reports that it is 8 percent. This means that any result must be interepreted as $\pm$ 8 percent, which is a considerable spread.

The study does provide insight into the difference between highly vocalized theories of instruction and curriculum and the actual practice in the schools. There has been limited research on such dichotomies. The development of explanatory hypotheses and further research is needed. Questionnaire studies only scratch the surface. Follow-up studies would provide additional detail which does not usually result from blindly distributed and selectively returned questionnaires. For example, it might be of interest to compare various populations of teachers (inexperienced vs. experienced) or simply do a careful study of individuals who are identified as "successful" biology teachers. The application of educational theory in the classroom might be better understood. As is frequently the case with descriptive studies, the one reported here raises more questions that it answers.

Shayer, Michael, Philip Adey, and Hugh Wylam. "Group Tests of Cognitive Development Ideals and a Realization." Journal of Research in Science Teaching, 18(2):157-168, 1981.

Descriptors--*Cognitive Development; *Cognitive Measurement; *Cognitive Tests; *Developmental Stages; Elementary School Science; Elementary Secondary Education; Evaluation Methods; *Group Testing; Intellectual Development; Science Education; Secondary School Science

Expanded abstract and analysis prepared especially for I.S.E. by Edmund A. Marek, The University of Oklahoma.

Purpose

The purpose of this research report is to describe seven Science Reasoning Tasks (1979) beginning with the development of the SRTs through the validation and utilization. Item discriminations, reliability and validity are assiduously reported.

Rationale

Cognitive development has traditionally been determined by the methode clinique developed by Jean Piaget. The investigators identify four essential features of the classical method of measuring levels of cognitive development: 1) allowance for the child to be influenced by his perceptions and the apparatus; 2) opportunity to investigate the reasons for the child's responses; 3) ability to observe the child's reaction to interviewer feedback; and 4) opportunity to question the child's response. Interview methods are very time-consuming, making it impossible or very difficult to collect large quantities of data for this type of research. The researchers developed tests of cognitive development which can be used to assess individuals, in groups of twenty or more, simultaneously.

Research Design and Procedure

Development of Science Reasoning Tasks is summarized by the investigators in five statements:

1. Selecting from the tasks devised by Piaget et al. (1956, 1958, 1964, 1974) those which cover the range of stages to be studied, and which seem most likely to be transposable to a group situation.
2. Writing test items from questions reported by Piaget and Inhelder in their interview tasks, together with appropriate instructions for administration.
3. Ascribing developmental stages to each possible reply to each item, followed Piagetian protocols. In practice, almost all items test the attainment of just one level or sublevel and a complete task must include items covering a suitable range of stages.
4. Devising an overall marking scheme by which a level may be ascribed to a pupil on the basis of his replies to a series of items. In general a two-thirds rule is followed: if there are six items at stage n, then four must be correct to indicate achievement of that stage.
5. Trying the task on a sample of pupils and assessing each by the provisional marking scheme.

Item discriminations and reliability, of the tasks developed by this process, were determined using item discrimination diagrams for tests assessing the range from 2B to 3B. Content validity was established by producing an adequate number of items at each of the levels 2B, 2B/3A, 3A and 3B. Internal consistency was measured by the Kuder-Richardson formula 20 at each stage of development of a task. Reliability was estimated with test-retest correlations tests given three months apart.

Findings

Examination of KR20 correlations indicates that the SRTs are virtually the same as the original Inhelder or Piaget Tasks. The predictive validity was determined with two SRTs -- Tasks II and III. Above average 11-13-year-old students were tested with a 50-60 item content examination in physics, chemistry, and biology. The researchers report that these questions measure understanding rather than recall. Predictive validity correlations of .77 and .78 were measured for the two sections of the course.

Construct validity was established by administering SRT Tasks III-VII to approximately 560 students, age fourteen years. Factor analysis showed a single factor accounting for 59 percent of the total variance. Population surveys with the group tasks have been conducted and population norms established for cognitive development of British school students. A wide range in rates of cognitive development was found and the correlation of Piagetian stages and age was 0.35. Population norms of these Piagetian measures do not increase after the adolescent growth spurt and the researchers predict this could account for sex differentials researchers in the United States have found on formal operations with college students.

The SRTs have also been used in studies in other cultures: southeast Asian countries, West Indies, the Philippines, Palestine, Zimbabwe-Rodesia and Swaziland.

Interpretations

The Science Reasoning Tasks have been thoroughly and carefully developed, validated, and reported. SRTs have much potential in applied research. The researchers conclude that the most powerful use of the tasks may be with matching teaching/learning activities with the cognitive developmental level of the learner. Interpretations from this instrument development research are reported throughout the manuscript and summarized in this statement:

> By monitoring the progress of groups of individuals, whose performance on SRTs has been recorded, through the curriculum and noting areas of success and failure, we can gain real insight into the levels of cognitive development needed to successfully complete each small section of the curriculum. In this way difficulties can be differentiated into those which can be remedied by changes in teaching approach and those which demand restructuring or even complete reframing of the curriculum.

ABSTRACTOR'S ANALYSIS

Development and validation of group tests of cognitive development are important achievements for research in science education. The investigators of this study conducted a comprehensive research and development program to produce

the Science Reasoning Tasks. This well written research report thoroughly documents the instrument development and field testing of those instruments. The investigators thoroughly reported each step of the research and included the essential historical influences--i.e., the Piaget interview tasks of cognitive development--providing the chronological evolution of the SRT development.

Group tests of cognitive development should provide data to identify the students' thought as preoperational, concrete operational, formal operational or transitional between levels. The SRT developers state that pupils' development should be categorized on at least six levels of cognitive development:

1	preoperational
2A	early concrete operational
2B	late concrete operational
2B/3A	transitional between late concrete operational and early formal operational
3A	early formal operational
3B	late formal operational

Previous research by Longeot (1965), Warburton (1966) and Raven (1973) used paper-and-pencil tests to measure logical reasoning and categorize thinking as concrete operational or formal operational. Tisher (1971) used a paper-and-pencil test on which the students were required to answer all the questions as thought experiments. Rowell and Hoffman (1975) developed a group test which required a set of apparatus for each student. Recognizing the limitations of these group tasks, the investigators developed a valid and reliable set of demonstration plus paper-and-pencil tasks of cognitive development. The Science Reasoning Tasks were then field tested and refined.

The SRTs utilize a set of apparatus to demonstrate various experiments and a series of questions to which the subjects respond in writing. The abstractor agrees with the developers of the SRT that giving a SRT is more like teaching a lesson than giving a standardized test.

REFERENCES

Inhelder, B., and J. Piaget. The Growth of Logical Thinking. London: Routledge and Kegan Paul, 1958.

Inhelder, B., and J. Piaget. The Early Growth of Logic. London: Routledge and Kegan Paul, 1964.

Longeot, F. "Analyse Statitique de Toirs Tests Genetiques Collectifs." Bulletin de l'Institut National d'Etude du Travail et d'Orientation Professionnelle, 1964.

Piaget, J., and B. Inhelder. The Child's Conception of Space. London: Routledge and Kegan Paul, 1956.

Piaget, J., and B. Inhelder. The Child's Construction of Quantities. London: Routledge and Kegan Paul, 1974.

Raven, R. J. "The Development of a Test of Piaget's Logical Operations." Science Education, 57:377-385, 1973.

Rowell, J. A., and P. J. Hoffmann. "Group Tests for Distinguishing Formal from Concrete Thinkers." Journal of Research in Science Teaching, 12:157-164, 1975.

Science Reasoning Tasks. Seven Piagetian Group-tests and a General Handbook. Windsor: NFER, 1979.

Shayer, M. "The Analysis of Science Curricula for Piagetian Level of Demand." Studies in Science Education, Leeds, 5:115-130, 1978.

Tisher, R. "A Piagetian Questionnaire Applied to Pupils in a Secondary School." Child Development, 42:1633-1636, 1971.

Warburton, F.W. "Construction of the new British Intelligence Scale: Progress Report." Bulletin of the British Psychological Society, 19:68-70, 1966.

Schade, Wayne R. and Rolland B. Bartholomew. "Analysis of Geology Teaching Assistant Reaction to a Training Program Utilizing Video-Taped Teaching Episodes." Journal of Geological Education, 29:92-102, 1980.

Descriptors--College Science; *Geology; Graduate Students; Higher Education; Science Education; *Science Laboratories: Science Programs; *Teacher Education; *Teaching Assistants; Teaching Methods; *Videotape Recordings.

Expanded abstract and analysis prepared especially for I.S.E. by Patricia H. Suter, Del Mar College.

Purpose

The purpose of this study was to ascertain reactions of graduate teaching assistants in three geology departments to the use of video-taped teaching episodes as a method of teaching these students effective teaching methods.

Rationale

Undergraduate students receive a large part of their instruction from graduate teaching assistants in many of the colleges and universities in this country. Criticism concerning the quality of the instruction received by undergraduate students is common on college campuses. These complaints center around four areas:

--Immaturity: Some graduate students are younger than the students they are supposed to teach.

--Lack of knowledge of teaching methods: Most have no prior training in teaching methods.

--Lack of interest in teaching: Many TA's do not like classroom teaching and do not aspire to make teaching a profession.

--Lack of specific matter knowledge: Students tire quickly of the "I don't know" response.

The authors of this article contend that the failure of teaching assistants to furnish quality instruction in their assignments is usually due to a lack of knowledge of how to provide competent instruction rather than to a lack of willingness.

Research Design and Procedures

The preparation of video-taped teaching episodes involved three phases. First, more than 20 geology teaching assistant volunteers were video-taped for one hour in their regularly assigned laboratory sections. Second, each one-hour video-tape was analyzed to ascertain what teaching techniques were being used. Third, the one hour video-tapes were edited into video-taped teaching episodes. Examples chosen for inclusion in each episode represent those that demonstrated the particular teaching techniques as distinctly as possible.

The prototype episodes were used in a pilot Teaching Assistants Training Program to ascertain teaching assistants' reactions to the prototype video-taped teaching episodes as an instructional medium and to their organizational structure. They reacted favorably to the video-taped teaching episodes as an instructional medium and to all other aspects of their organizational structure.

Following this initial tryout and taking into account the suggestions of the teaching assistants in the pilot program, 13 additional episodes were produced. These lasted from five to nine minutes and consisted of a composite of several assistants using the same teaching technique.

Of the total of 16 episodes prepared, four were chosen for use in the research program. The titles were:

1. Using the chalkboard
2. Nonverbal cues
3. Student-initiated talk
4. Closure

The four episodes were chosen for the following reasons. One, these episodes represent the best overall playback quality. Two, the applicability of the

teaching techniques used in these episodes were widespread. Three, the episodes had been used in the pilot program and were judged successful in their ability to initiate discussion. Four, the episodes could be used satisfactorily in the time available to conduct the Research Training Program.

The training program involved assistants from three universities. Three one-hour luncheon meetings were scheduled at each university. During the first and second meetings - held 48 hours apart - two video-taped teaching episodes per meeting were viewed and discussed. The third meeting was scheduled one week after the second one, during which the participants completed a questionnaire. Guide questions for each episode were distributed and were used as the basis for the discussion following the showing of the tape. Then the episodes were viewed a second time. At the third meeting the Research Training Program Questionnaire was administered.

Findings

The authors posed six questions to be answered from which they could determine reactions of the graduate teaching assistants. These questions were:

Question 1. Do teaching assistants participating in the training program utilizing video-taped episodes increase their awareness of importance of the four teaching techniques used in the episodes studied?

Question 2. Do teaching assistants have a more positive reaction toward the training program after participating than before?

Question 3. Do teaching assistants participating in a training program utilizing video-taped teaching episodes view this instructional format as an acceptable way to learn about teaching techniques?

Question 4. Do teaching assistants participating in a training program utilizing video-taped teaching episodes view this instructional format as an enjoyable way to learn about teaching techniques?

Question 5. Do teaching assistants become aware of specific teaching techniques during participation in a training program utilizing video-taped teaching episodes?

Question 6. How do teaching assistants participating in a training program utilizing video-taped teaching episodes rank the training sessions regarding their value to the teaching assistants in their teaching situations?

The Research Training Program Questionnaire was used to answer most of the questions. The Wilcoxian matched-pairs signed-ranks test was used to give the significance level for questions 1 and 2. The Kendall Coefficient of Concordance was applied to measure the amount of agreement among the rankings for question 6.

The conclusions found were that the video-taped teaching episodes offer teaching assistants the opportunity to view others of similar persuasion doing something the assistants do and, thus, they can identify with the teaching event. The ability to identify with the event apparently has made a very positive impact on the teaching assistants.

The video-taped teaching episode used with a group discussion provides the teaching assistants with a non-threatening environment focusing on teaching techniques.

Interpretations

The results reported by the authors were very favorable toward the use of video-taped episodes to instruct graduate assistants in teaching techniques. The participants indicate that they would favor this type of training to acquire the necessary knowledge about teaching to enable them to provide competent instruction in their undergraduate classes.

The authors conclude that "in the final analysis, it will be the undergraduate students and the institution that will benefit from this improved instruction."

ABSTRACTOR'S ANALYSIS

This study was designed to evaluate a possible method for use in improving the teaching techniques of graduate teaching assistants in geology. The amount of work on the part of the authors in preparing the video-taped episodes appears large. Just the job of editing the 20 one-hour video-taped segments into teaching episodes would preclude many others from attempting like methods of instructing their teaching assistants in pedagogy.

This abstractor believes that the process of showing TAs video-taped sessions of other TAs performing like duties is an effective way of imparting knowledge on teaching techniques. A video-tape of the individual TA in action, which is then reviewed by the TA and supervisor, would seem to me to be a valuable adjunct to the "canned" tapes considered by the graduate students in groups.

This study addresses one of four problems the use of graduate teaching assistants poses. These are immaturity, lack of knowledge of pedagogy, lack of interest in teaching, and lack of subject matter knowledge. Of course, if the other three are missing, knowing how to use the chalkboard helps very little.

This study appears to be well designed to acquire the information desired. The use of video-tapes for teaching techniques is not new, but the design of the study gives fresh applications of this tool. Providing for more than one viewing of the taped teaching episode with discussion group activity in-between gives the TA the chance to compare his/her reactions with those of others in the group.

The conclusions drawn by the authors - that the TAs participating in the study concluded that the use of video-taped episodes for imparting teaching techniques was an effective and enjoyable method - seems to be statistically valid. It would have been helpful to this reader if they had included a copy of the Research Training Program Questionnaire. They used the replies to this questionnaire to answer some of the questions they posed from which their conclusions were drawn.

The procedures used by the authors in this study are novel as far as this reviewer has been able to ascertain. It would be interesting if the authors would do a follow-up study of undergraduate ratings relating to their teaching techniques of the TAs who participated in the study. These could be compared to the ratings of TAs who did not participate in the training sessions.

One aspect of the use of graduate teaching assistants which the authors did not mention is the use of foreign students in these positions. Their English may be correctly spoken, but their accent may be heavy enough to prove distracting. Because foreign students are significant part of our graduate enrollment and population of assistants in major colleges and universities, application of this method of video-taped instruction might prove helpful in improving the effectiveness of TA instruction.

TESTING

Perry, Glen Richard, Jr. and Dale G. Merkle. "Validity and Reliability of the SCIS Test for the Organism Unit." Journal of Research in Science Teaching 13(2): 243-247, 1976.

Descriptors--*Educational Research; Elementary Education; *Elementary School Science; Evaluation; Science Education; Science Course Improvement Project; *Test Reliability; Tests; *Test Validity

Expanded abstract and analysis prepared especially for I.S.E. by Russell H. Yeany, University of Georgia.

Purpose

The stated purpose of the study was to assess the reliability and content validity of the first grade test developed for the Science Curriculum Improvement Study unit titled "Organisms."

Rationale

According to the authors, at the time of their study, no data were available on the effectiveness of the SCIS Organisms test.

Research Design and Procedure

The Concept/Process evaluation subsection of the test was administered pre and post to 80 first-grade students. Five boys and five girls were randomly selected from eight different schools in one school district. Some sections of the subtest were administered as individual tests while other parts were given as group tests. Each of five different sections of the posttest were administered after a teacher had completed teaching activities related to a particular section. The attitudes in science and the perception of classroom environment sections of the test were not administered. The authors acknowledged this as a delimitation of the study.

Split-half reliability procedures were used to estimate the reliability of the scores. There were a total of 41 items on the five sections of the Concept/Process subtest. A panel composed of the authors and two

teachers judged the content validity of the items by comparing them to the objectives in the SCIS teacher's guide.

Findings

The split-half reliability of the Concept/Process scores of the SCIS Organisms test was calculated as 0.577. The panel agreed that the test items are consistent with the objectives of the Organisms teacher's guide.

Additional data analysis indicated that there were no mean score differences between boys and girls and between rural and urban students on the pre- and posttests. There was a significant mean gain from pre to post.

Interpretations

The Concept/Process section of the assessment instrument of the SCIS Organisms unit was judged to be both valid and reliable and the authors recommended it as an acceptable method of evaluating the degree to which pupils attain the Organisms unit objectives.

ABSTRACTOR'S ANALYSIS

The desire to establish and assess the reliability and validity of measures used in research and evaluation should be a high priority of science educators. However, one disappointment in the Perry and Merkle study is its lack of scope. The authors examined a test which measures the effects of only one unit of the SCIS program. There are 12 units in the total program: one life and one physical science for each grade level. They then further reduced their analysis to examine only a subtest score of that test, one of three subtests for the unit. To make the study of any real value, parallel data collection and analysis should have been carried out on a more complete or at least representative set of the SCIS evaluation materials. There was no explanation or rationale as to why the study was so restricted.

In relation to methodology employed in the study, there are several points which need to be addressed. First is the inadequacy of the sample to facilitate generalizability. The reliability data were collected on only 77 students in a single school district in the rural Cumberland Valley in Pennsylvania. Even though four of the study schools were labeled as "urban," the population of the town is only 25,000 persons and hardly meets the definition of an urban setting. An explicit description of the sample was lacking but an implicit definition of the students represented does not allow for a broad generalizability to target populations. A second concern relates to the conditions under which the tests were administered. That is, five subsets of items which represent activities were administered at different points in time and the data for each individual were pooled across time to construct a Concept/Process subtest score. If that is the usual administration procedure, some decisions on individual or group performances are probably made at the level of the activity and reliability values (which are probably low due to the reduced number of items) should be reported out at this level. Also, three of the five subsets of items were administered on an individual basis. Variation in test scores could have been influenced by unreliability in the individual test administrator. No assessment of this influence was reported.

The authors reached a questionable conclusion when they stated categorically that an $r = 0.577$ was significant and therefore the test was reliable. The question of test reliability is not one of a statistically significant correlation between two halves of a test. The question is: Is that correlation high enough to reduce the error variance in the score to a degree that a test score is considered to be a fairly stable estimate of the true score? A value of 0.577 should be considered as a low reliability for a group test (especially for a concept/process test) and is totally unacceptable as a measure of individual performance. The 95 percent confidence interval around a score representing achievement on the organisms Concept/Process test is 9.2 score units (points) wide. This represents a lot of measurement error for a 41-point test.

Two judgments related to the test's validity are incongruent. The first was that the test was valid. The second was that portions of the instrument need to be altered or deleted; or, if not deleted, teachers are warned not to place too high a value on these items.

In general, the article provides a bit of interesting information but is too limited in scope to be considered more than a pilot effort. Much more work should have been completed before it was reported.

Warren, Gordon, "Essay Versus Multiple Choice Tests," Journal of Research in Science Teaching 16(6), 563-567, 1979.

Descriptors--*Educational Research; Educational Theories; *Evaluation Methods; Instructional Materials; *Science Education; Teaching Methods; *Test Construction; *Test Items; Testing

Expanded abstract and analysis prepared especially for I.S.E. by Richard B. Baldauf, Jr. and John Edwards, James Cook University of North Queensland, Australia.

Purpose

The purpose of this research report was to compare essay and multiple choice tests as a means of testing the factual content of science learning.

Rationale

The study approaches the topic strictly as an applied testing problem. The problem is important to science teachers because the type of assessment used has an effect on instruction and learning. The study assumes that:

(1) Essay and multiple-choice tests have the same evaluation purposes, i.e., testing factual content.

(2) The percentage correct on each type of test provides an absolute measure of achievement which is valid for comparing the two types of tests.

Research Design and Procedure

The non-random sample used in this study came from three classes containing 70 building industry employees in the 18- to 30-year old age range who were taking a second-year course in building construction at a London college. The design, which the author incorrectly describes as "a pre-post no-control-group design", can be better conceptualized as a delayed, parallel forms, test-retest

reliability design (Thorndike, 1976). The problem should have been conceptualized in measurement rather than research design terms.

An essay and a multiple-choice achievement test, each consisting of five questions measuring the same factual material, were given with a one week interval between testing. Errors made on the essay test were used to construct distractors for the multiple-choice test. To compensate for order, the procedure was then reversed with the essay test being administered after the multiple-choice test. To avoid inflating the retest results with differences due to learning, the alternate form re-test was given without prior warning.

Findings

The results of the study are as follows:

	Order of Presentation	
	Essay v. M.C.	M.C. v. Essay
Essay	39.55%	40.4%
Multiple-Choice	59.08%	54.3%
Difference	19.53%	13.9%

Results also showed that for the essay-multiple choice order of presentation, 33.33 percent of students repeated the original mistake, 14.82 percent made a new mistake, and 51.85 percent who made a mistake originally correctly answered the question where it was later presented in a multiple-choice format. Some opinions volunteered by students about the relative value of essay and multiple choice examinations are also listed. Test-retest correlations, essential to the interpretation of the data, are not provided.

Interpretations

The author claims the results indicate that:

(1) it is easier to obtain high marks with multiple-choice tests than with essay tests,
(2) some students believe that in an essay-type test quantity will compensate for lack of quality, and
(3) essay tests reveal weaknesses that can be hidden in the multiple-choice tests.

ABSTRACTORS' ANALYSIS

This research report illustrates many common errors and inappropriate assumptions found in classroom assessment programs. It also provides an example of inconsistencies in argument and a lack of appropriate detail in presentation too often found in the research literature.

Use of the Literature. In a brief research report, an exhaustive review of the literature is not expected. However, the sources cited ought to provide a basis for research which follows and they ought to be the most relevant available. Of the four studies cited, Frisbie (1973) and Oosterhof and Glasnapp (1974), in their comparisons of multiple-choice and true-false test formats, raise methodological issues which ought to have informed the research report. Issues discussed included the need for a systematic and objective procedure for conversion from one type of test format to another, the need to consider a correction for guessing and the need to consider time difference as a confounding factor in comparing the test formats. None of these procedures were incorporated in the report.

Warren's report cites Gronlund (1976) to establish that multiple-choice and essay tests have a different focus and differing strengths and weaknesses. However, this delineation is completely ignored in the conceptualization of Warren's study - the two formats being compared only as means of testing factual content. The final paper by Voss (1974) seems irrelevant to the issues involved in the research report. There is no evidence of it contributing to the formulation of the study.

With the printed indices and information retrieval techniques available to researchers today, there is little reason for overlooking significant, relevant research. McCloskey and Holland's (1976) paper would have provided a sound basis for Warren's research. This paper, which is easily found through ERIC, is a conceptually well-developed and methodologically sound comparison of student performance in answering essay and multiple-choice type questions.

Unawareness of Appropriate Contexts. Any regular reader of the literature, or user of modern curriculum materials, should be aware that science education has shifted in focus away from factual recall of content. Even the literature Warren cites (e.g., Frisbie, 1973) deals with higher level cognitive processes. It is disappointing that such higher level processes and problem solving and practical skills are not reflected in the testing program.

Although few studies have empirically compared essay and multiple-choice tests, the philosophical and historical issues related to these two types of testing have been widely discussed in science education journals (Ford, 1973; Ongley and Houk, 1969; Thomson, 1970). These studies introduce the researcher to the philosophical and judgemental issues which "must provide the value framework within which a final decision is made" (Thorndike and Hagen, 1977, p.21).

Furthermore, although there are situations where either essay or multiple-choice tests could be used appropriately, it is generally recognised that each test format has its particular strengths and weaknesses. These differences have been described both in widely used texts like Thorndike and Hagen (1977, p. 257) and in the popular literature (e.g., Roth, 1978). It is both unwise and unproductive to use one test format where the other is better suited. In this research report, the reduction of the questions to simple factual responses, for the sake of comparability, avoids the strengths of either test format, producing a comparison which has little value to the serious user of either technique.

While it could be argued that general issues such as those discussed in the preceding paragraphs are not the province of a brief research report, an understanding of these contexts does have implications for the way the study was conceptualised and conducted.

Definition and Use of Test Formats. The essay questions provided as examples in the research report are better described as short answer questions (Thorndike and Hagen, 1977, p. 256) in which a response of one well-written sentence would provide the required answer. Furthermore, the question-answer link for these "essays" is rather tenuous because of the very specific nature of answers required to make the questions scorable on a "factual" basis. To score a general essay question as either totally correct or incorrect based on a single, narrow, factual statement, negates the whole notion of essay tests.

One of the main advantages of the multiple-choice technique is that it provides a breadth of sampling of the materials covered. To suggest an equivalence between one essay question and one multiple-choice question is to deny this advantage. In addition, an examination of the multiple-choice questions themselves causes concern. With such excellent references as Klopfer (1971) and Hedges (1968) available, science educators are in a position to set imaginative and well-constructed multiple-choice items. The questions presented in the report are poor examples of the multiple-choice format.

Definition and Comparison of Difficulty. Warren's report is based on a philosophically absolutist view of test difficulty. This position ignores the fact that test difficulties are influenced by who writes the test, and the test format used. Measurement experts have long agreed that there is no valid absolute meaning of test difficulty except in certain types of criterion-referenced test situations. The myth that marks and standards are in some way absolute has deceived students, teachers, parents, and employers for far too long. Poor student placement and misleading vocational advice are but two results of such misplaced faith.

The difficulty of test questions is always determined to some degree by the person constructing the test. Test questions can be made easier or more difficult by changing question phrasing, or altering item alternatives, or altering the demands made of the student by the question. Question difficulty is therefore more likely to be a function of question design than of test format.

Multiple-choice test items are affected to a much more marked extent by guessing than essay questions. Furthermore, Thorndike and Hagen (1977, p. 205) indicate that changing the number of multiple choice distractors provided influences the difficulty of the test. These factors lead one to expect an inbuilt difference in difficulty between essay and five option multiple-choice test where questions are written to maximise test reliability.

Finally, Warren's report implies that learning has not taken place between tests, yet the results cited refute this. The figure of 51.85 percent for "original mistake, now correct" for the essay test followed by the multiple-choice test strongly suggests learning has occurred. This apparently significant result is not discussed in the analysis.

Any of the factors mentioned in this section, alone or in combination, could account for the "difficulty differences" that are used as a basis for the report's conclusions.

Presentation and Interpretation of Data. A major problem in reading this report is that there is no detailed summary of data which permits the reader wishing to explore the problem further to verify or replicate the results. The reader must, of necessity, accept "on faith" many of the statements presented. Detailed results could have been presented without denying the need for compactness necessary in a research report.

The use of averaged percentages as the method of presenting the data is also of concern because it fails to provide individual item results for equivalent pairs of questions. Unlike the McCloskey and Holland (1976) article, where a question by question analysis lends support through replication to the authors' hypotheses, this study's use of overall averages leaves the reader without the important detail required for interpretation of the results. In addition, the lack of standard deviations for these percentages makes it impossible to determine whether the differences presented are statistically significant or artifacts of the conversion to percentage process.

Finally, it is unclear how the test questions for the second order of presentation were written, nor are any examples given. It is not possible

therefore to determine if the study's two "experiments" were comparable. Without comparability, the reversal of the testing procedure adds nothing to the study.

Relationship Between Results and Findings. Unfortunately, the report lacks any detailed discussion or interpretation. The "discussion" section presents the study's findings. Nowhere is a clearly argued relationship developed between the data and the findings drawn from that data. Perhaps it is not surprising then that the specific findings indicated are not supported by the evidence provided.

For the first finding, no real evidence is presented to show that "it is easier to obtain high marks with multiple-choice tests than with essay tests." Uncontrolled factors such as lack of question equivalence, the recognition versus recall factor, learning, or the specificity of the correct essay answer, suggest alternative hypotheses for explaining the reported differences.

The second finding that "some students believe that in any essay type test quantity will compensate for lack of quality" is supported by three student comments, but negated by several others. The seventeen comments presented do not generally support, nor negate, this position.

Finally, Warren concludes that "essay tests reveal weaknesses that can be hidden in multiple-choice." This appears to be a value judgement which is not supported by any data in the report.

Discussion. This is a very weak piece of research. As such, it is disturbing to find it in a widely respected, refereed research journal. However, our major concern arose from our discussion of the report with our science education students. For many, publication in a refereed research journal puts the stamp of respectability on a piece of research. Some readers may therefore be tempted to accept uncritically, or at least less critically, what they read therein. Readers who are not already aware of the major weaknesses in this research report ought to have their attention drawn to them. We hope this process will help to improve the quality of research in science education.

REFERENCES

Ford, D.F. "The Future of Biological Assessment." Journal of Biological Education, 7(2), 8-11, 1973.

Frisbie, D.A. "Multiple Choice Versus True-False: A Comparison of Reliabilities and Concurrent Validities." Journal of Educational Measurement, 10: 297-304, 1973.

Gronlund, N.E. Measurement and Evaluation in Teaching (3rd ed.). New York: Macmillan, 1976.

Hedges, W.D. Testing and Evaluation for the Sciences. Belmont, California: Wadsworth, 1968.

Klopfer, L.E. "Evaluation of Learning in Science." In B.S. Bloom, J. T. Hastings and G. F. Madaus (Eds.) Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw Hill, 1971.

McCloskey, D.T. & R.A.B. Holland. "A Comparison of Student Performances in Answering Essay-Type and Multiple-choice Questions." Medical Education 10, 392-385, 1976.

Ongley, P.A. & C.C. Houck. "Provocative Opinion, Examinations: Essay or Objectives -- Two Views." Journal of Chemical Education, 46, 830-832, 1969.

Oosterhof, A.C. & D.R. Glasnapp. "Comparative Reliabilities and Difficulties of the Multiple-Choice and True-False Formats." Journal of Experimental Education, 42(3), 62-64, 1974.

Roth, M.S. "Design your own Evaluation Tools." Audiovisual Instruction, 23(8), 21-23, 1978.

Thomson, P. "Multiple Choice and Essay Test Items for Classroom Testing." Australian Science Teachers Journal, 16, 45-50, 1970.

Thorndike, R. L. & E. P. Hagen. Measurement and Evaluation in Psychology and Education. (4th Ed.) New York: John Wiley, 1977.

Thorndike, R. M. "Reliability." In Brian Bolton (Ed.), Handbook of Measurement and Evaluation in Rehabilitation. Baltimore: University Park Press, 1976, 15-37.

Voss, J. F. "Acquisition and Nonspecific Transfer Effects in Prose Learning as a Function of Question Form." Journal of Educational Psychology, 66, 736-740, 1974.

Fraser, Barry J. "Developing Subscales for a Measure of Student Understanding of Science." Journal of Research in Science Teaching, 15(1): 79-84, January 1978.

Descriptors--*Aptitude; Educational Research; Elementary Secondary Education; *Junior High Schools; Middle Schools; *Science Education; *Test Construction; Test Interpretation; *Test Validity; *Tests.

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Rodney L. Doran and Samuel J. Aliamo, State University of New York at Buffalo.

Purpose

The purpose of this study was to develop a "new instrument suitable for measuring understanding of the nature of science among upper elementary and junior high school students and to report validation data from the administration of the instrument..."

Rationale

A widely used research instrument in science education, Test on Understanding Science (TOUS), was developed by Cooley and Klopfer (1961). This instrument measures the understanding of the nature of scientific inquiry, of science as an institution, and of scientists as people. A version of TOUS Form W was designed for senior high school and a simplified version for elementary school (Form Ew) and junior high school (Form Jw). This study evolved in response to two concerns: (1) numerous criticisms of the TOUS and (2) Form Ew and Jw yielded only a single unidimensional score.

Research Design and Procedure

A panel of 12 people consisting of educational measurement experts, scientists, science educators, and school science teachers reviewed TOUS Forms Ew and Jw. Modifications and elimination of items were made after checking for face validity and item faults. The panel modified items to a seventh-grade readability level and also wrote new items and allocated items to the three subscales, described below:

Subscale	Measures
Philosophical	Relationships among types of statements in science and limitations of scientific explanation.
Historical-Social	Historical perspective; relationships among science, technology, and economics; social and moral implications of science.
Normality of Scientists	Student appreciation that scientists are normal people.

These subscales were consistent with Klopfer's (1971) classification of science education objectives and several of the criticisms of the TOUS. According to the author, "the validity of the present subscales rests initially and primarily on the judgment of these experts. Although the statistical techniques described later are useful for refining scales, they certainly cannot be used to transform grossly inadequate scales into satisfactory ones."

The revised instrument was administered to 176 seventh-grade students randomly selected from fourteen public and private schools located in the Melbourne metropolitan area, chosen to be representative of area schools. Data from this sample were analyzed with the aim of "identifying faulty items whose subsequent removal would optimize the overall scale statistics." One technique used to enhance internal consistency of scales was to remove any item whose item-remainder correlation was not significantly greater than zero ($\alpha = .0$). A second technique to improve discriminant validity was the removal of any item whose correlation with "assigned" scale was smaller than its correlation with either of the other two subscales. These two methods reduced the original item pool to 30 items.

These "refined subscales" (30 items) were administered to a second sample, called the crossvalidation sample, to check the stability of the statistics with a different group of students. The crossvalidation sample consisted of 1158 seventh-grade students from 46 schools from the Melbourne area. According to the author, this sample "was representative of the population of schools."

Reliability estimates (KR-20) and subscale intercorrelations were calculated from the data available with both the validation and the crossvalidation samples.

TABLE 1

Number of Items in KR-20 Reliability of, and Intercorrelations between Subscales

Subscale	No. Items	KR-20 Reliability Valid	Crossvalid	Intercorrelations[a] P	H	N
Philosophical (P)	12	0.55	0.51	1.00	0.44	0.41
Historical-Social (H)	12	0.61	0.62	0.42	1.00	0.41
Normality of Scientists (N)	6	0.60	0.53	0.39	0.54	1.00

[a]Subscale intercorrelations are given below the diagonal for the validation sample and above the diagonal for the crossvalidation sample.

From Table 1, one can see that the KR-20 reliabilities of the three subscales ranged from 0.55 to 0.61 in the validation study and from 0.51 to 0.62 in the crossvalidation study. The subscale intercorrelation coefficients were within the range: .39 to .54.

Interpretation

The reliabilities of the three present subscales compare favorably with reliabilities of the TOUS Form W three subscales which ranged from 0.52 to 0.58 for a sample of 2,535 students, as reported by Klopfer (1961). This i noteworthy as the present subscales are considerably shorter than those of the original TOUS (12 items as compared to 20). Similarly, the reliability of the entire present instrument (30 items), 0.77 and 0.78, compares well with values from the original TOUS forms which ranged from 0.58 to 0.76. The author claimed that, although the scale intercorrelations are reasonably large (+0.39 to +0.54) and suggest substantial overlap, "the sizes of the intercorrelations are low enough to indicate that subscales do not measure exactly the same thing." Consequently, he concluded that "satisfactory discriminant validity for the subscales" existed.

The author admits that the reliabilities of the present three subscales preclude their use to measure individual student performance but

states they could be used in several ways by teachers and researchers in measuring group performance. Also, he concludes that the new instrument can be used to identify common specific misunderstandings of the nature of science among students. The author also states that a third use of the new subscales would be to compare understanding of the nature of science after using alternative teaching techniques or differences in such attributes as sex, social class, race, personality, or attitudes. Lastly, the author concludes that the new instrument differs from the TOUS Forms Ew or Jw as it provides three separate scores rather than one single score. Based on the data from two separate samples of Australian seventh graders, the author claimed that "the three subscales possessed satisfactory internal consistency and discriminant validity for use in measuring the performance of groups of students."

ABSTRACTOR'S ANALYSIS

The purpose of this paper was to prepare a "new" instrument in the nature of science area. Many science educators have argued that we do not need a proliferation of more "new" instruments, rather modifications and crossvalidation of existing instruments. Despite the author's repeated reporting of the development of a "new" instrument, I think he really modified and validated existing TOUS forms. However, the reader is not well informed as to how many items from TOUS Form Ew or Jw were among the items administered to the validation sample or the crossvalidation sample. Therefore, one can only guess to what degree the instrument is "new," beyond the utilization of three subscales.

It is clear from the widespread use and substantial criticism of the TOUS, that it could profit from some research oriented toward improvement and strengthening. It assesses outcomes in an area of science education that are attracting increased attention, especially with middle/junior high school students.

The use of a "panel of experts" is widely used in all areas of education and can contribute considerably. The size and breadth of the panel Fraser employed is adequate. However, it is not clear what criterion the panel used to determine face validity or item faults.

Was a listing of the three scales and representative elements available for the panel? Was a list of errors common to multiple-choice items available to help the panel judge item quality? How many new items were written by the panel?

The three scales Fraser chose were consistent with other researchers in the field. Fraser described his third Scale--the Normality of Scientists--as quite distinct and new while it appears to be very similar to the original TOUS subscale--Understanding about Scientists. Any further discussion of what constitutes each subscale can be best illustrated by representative items--noticeably missing from the report. Realizing that journal space is always tight--even sample items from the three subscales would have added considerably to the impact of the report.

The selection of schools and students seemed to be appropriate for a validation study. The size of the validation sample--176--is marginal for that important first stage of validation. However, the reliability estimates were not different from those obtained with the much larger crossvalidation sample. The number of items administered to the validation sample was not specified, although 30 of the items satisfied the two selection procedures used. These statistical procedures were thorough, well-described and relevant to the study. These 30 "psychometrically acceptable" items were subsequently administered to a separate sample to ascertain stability of the statistical parameters. None of the data from the item analysis techniques were summarized, only reliability estimates and subscales intercorrelations were given.

Most literature suggests a minimum of .70 test reliability for use to assess group performance, although the author cited references that suggest that reliabilities "like those of the present subscale are quite adequate to justify their use in measuring the performance of graphs of students." Even for such small subtests (6 and 12 items) there must be concern about the low reliability.

Similarly, the large intercorrelations among the three subscales are discounted as being a stumbling block. Perhaps a procedure such as factor analysis or discriminant analysis would be useful to determine the

unidimensional or multi-dimensionality issue. No one would disagree with three scores providing more information provided they really are distinct measures.

The author suggests several possible uses, some already being implemented by himself, for this nameless research instrument. Research into the assessment of these outcomes is important. It is hoped Fraser continues to pursue "fine-tuning" of these TOUS-like instruments.

REFERENCES

Cooley, W. W. and Klopfer, L. E. Test on Understanding Science, Form W. Princeton, N.J.: Educational Testing Service, 1961.

Klopfer, L. E. "Evaluation of Learning in Science," in B. S. Bloom, J. T. Hastings, and G. F. Madaus (Eds.), Handbook on Summative and Formative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

CURRICULUM

78

Fraser, Barry J. "Use of Content Analysis in Examining Changes in Science Education Aims over Time." Science Education, 62(1): 135-141, 1978.
Descriptors--*Educational Objectives; *Educational Research; *Educational Trends; *Objectives; Science Education; *Science Education History; *Science History

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Ronald D. Anderson, University of Colorado.

Purpose

The purpose of this study was to identify the variations over time in the relative emphasis given to the goals of science education as stated in the science education literature. The stated goals of science education found in the literature for the period of 1932 through 1974, at all levels, were identified. A systematic process then was employed to identify changes in these goals that took place over time.

Rationale

The number of previous studies conducted in this area is very small. Whether one discusses specifically research on stated goals or research on science education goals broadly, it is an area of relatively little activity. The author cites previous work by Newport (1965) and by Ogden (1974, 1975) and goes on to contrast his current work with these earlier studies, both of which used content analysis. One of the previous studies (Newport) had shown little change in the goals of elementary school science over several decades while the other study (Ogden) had shown "pronounced differences" in the emphasis given to the various aims of high school chemistry.

The focus of this study, as well as the earlier ones, was on stated goals of science education. The actual pursuit of goals in the classroom was not studied nor was it assumed that stated goals would necessarily be reflected in school practice.

Research Design and Procedure

The author found 117 publications released between 1932 and 1974 which contained stated aims of science education. These references pertained to all grade levels and included 1,547 stated aims. These aims were classified using a slightly modified form of a classification system developed by Klopfer (1971). For purposes of the analysis of their data, the 117 references were divided into three categories: 1) those which pertained specifically to curriculum projects, 2) all other papers published prior to 1960, and 3) all other publications dated 1960 or later. In collating the results of this analysis, two indices of importance were used for the goals identified. The first index was the number of stated aims in each major category of the classification system employed. The second index was the number of references stating at least one aim that fit within a given category. These numbers also were expressed as proportions of the total number of references in a given group (1, 2, or 3 above).

Further analysis was conducted by converting the proportions of stated aims and references to ranks and using Spearman's rank-order correlation to determine the correlation among the ranks of the various categories for the three groupings of the studies cited above.

Findings

The analysis established a correlation between studies published after 1960 and studies pertaining to curriculum projects of .89 for the references and .95 for the stated aims. When comparing those studies published before 1960 and those published in 1960 or later, the correlations were .84 for references and .81 for stated aims. The correlations between studies published prior to 1970 and those pertaining to curriculum projects were .63 for references and .68 for stated aims. While describing these relationships as relatively high, the author does cite two "noteworthy differences" which are apparent in the tabulated data. A category pertaining to theoretical models was rated as being relatively more important in the references published in 1960 or later than in the

earlier references. Theoretical models were rated even more highly in those references pertaining to the curriculum projects. A second difference noted was the consistently lower ratings of importance given to applications of knowledge in those references pertaining to the later period or to curriculum projects as compared to those references published prior to 1960.

Interpretations

The major conclusion reached in this study is that there is a "relatively high overall similarity" between the emphasis given to the various aims of science education in recent years relative to the emphasis given in the earlier period and the emphasis within curriculum projects. In addition to this major conclusion, various applications and implications of the study are noted. It is cited as a basis for curriculum evaluation where it can serve as a baseline against which the stated aims of a curriculum project could be compared. A further application in the realm of curriculum evaluation would be as the basis for selecting the battery of scales to be used in curriculum evaluation. Finally the author notes that the study illustrates the usefulness of content analysis as a science education research technique, an application of it which is not frequently employed.

ABSTRACTOR'S ANALYSIS

Although the author of this study is persuaded that "perhaps the greatest merit of this article is that it has illustrated the general usefulness of the technique of content analysis in science education research," this writer is convinced that greater significance should be attached to the fact that the study is directed to a largely unresearched area: science education goals. While goals are not infrequently discussed, in the research arena they are usually assumed and not critically examined. Goals are largely unresearched. Many different facets of this arena are deserving of more careful study as will be discussed in more detail. This study was directed toward changes in science education goals over time.

Its major contribution is the identification of their relative stability over time.

Although some reservations about this conclusion will be noted below, the conclusion stands--science education goals have been relatively stable over recent decades. This conclusion raises the question of whether or not science, technology and society have been as stable as these goal statements. This is obviously an area that needs study. It is not simply a matter of historical interest; major attention needs to be directed to the question of whether or not the goals of science education as promulgated today are consistent with the setting to which they pertain, i.e., our science, technology, and society.

While the validity of the study at hand is by and large not the subject of challenge here, a few reservations could be expressed which may temper the conclusions to some degree. First of all, the correlation coefficients found are described as being relatively high, as is the inferred congruence between the emphases placed on different aim categories. While this description certainly is appropriate for correlation coefficients such as .89 and .95 and probably even .84 and .81, this conclusion is not so certain with respect to correlation coefficients such as .63 and .68. If one assumes the size of these coefficients is not restricted due to violation of assumptions such as homoscedasticity and unrestricted variability (assumptions not examined in the article), it is difficult to attach an adjective such as "high" to these correlation coefficients. Squaring the correlation coefficients and coverting them to a percentage, of course, tells us how much of the variance of the one variable can be predicted from the other. These two correlation coefficients, .63 and .68, give percentages of 40 and 46 respectively. Using these more appropriate numerical indicators of the degree of relationship, it would seem more appropriate to describe it with a word like "moderate" rather than "high."

A second matter worthy of mention is the fact that changes of emphasis within goal categories were not analyzed, i.e., we do not know how much shift in emphasis took place within each of the several goal categories used in the analysis. The author clearly identifies this limitation

of the study and it should be so noted. It leaves one with the question, however, of whether or not the conclusion about the high degree of relationship between goals over time would have been moderated if an analysis had been done of shifts in emphasis that took place within each of these major categories. Have the specific goals within each of these categories shifted with time to remain consistent with science, technology and society as we know them today?

Upon reviewing the overall design of the study, one wonders if further useful information would not have been obtained if the data had been broken down further according to their source, e.g., scientists, teachers, and science educators. If significant differences were found in this realm, the knowledge could have a significant bearing on our attempts to understand such matters as the apparent failure of many of the modern science curriculum project materials to be fully implemented in the schools in the manner intended by their developers.

Reviewing this study and its relationship to the sparse matrix of extant studies in this arena makes obvious the lack of attention to this important area of research. In contrast to the plethora of research in some other areas, there is a dearth of work in this realm. Yet it is an area of critical importance; the old question of "what knowledge is of most worth?" has not lost any of its relevance. Further, there are many facets of this arena that are amenable to scholarly attention with a variety of research methodologies.

There are some aspects of this arena that are fairly well understood. Just what the stated goals of science education are is reasonably well established through studies such as the one under consideration here. In addition, but not quite as obvious, the goals which are actually sought by teachers in the classroom are not the same as the stated ones. This conclusion is one of the major inferences to be drawn from the recent work of Stake et al. (1978). This extensive research, utilizing trained observers in selected school systems across the country, provides abundant evidence of this inference.

This apparent discrepancy between stated and actual goals leads directly to consideration of some of the following research questions which are posed as among those needing attention.

To what extent do stated goals vary among teachers, scientists, and various segments of the public?

To what extent do teachers agree (as reflected in classroom practice) with the commonly stated goals?

Why is there such a discrepancy between stated goals and those actually sought?

What science education goals would be inferred from a systematic analysis of today's science, technology and society?

How can the process of goal setting (stated and actual) be influenced?

Obviously the above listing of questions is not exhaustive but may give some indication of the fertility of this research area. With such a range of questions, this area lends itself to a variety of research techniques such as philosophical analysis, delphi techniques, various survey techniques and many others.

The importance of this area of research may be illustrated by the findings of recent "time-on-task" research such as that of Berliner (1975). By and large we can teach whatever we take time to teach. Yet we continue to devote our research to how to go about teaching science and largely ignore what should be taught in the time available. The question of "What knowledge is of most worth?" was never more relevant than it is today. It deserves our attention with whatever help can be obtained from the many modern research techniques at our disposal.

REFERENCES

Berliner, D. C. The Beginning Teacher Evaluation Study: Overview and Selected Findings, 1974-75. Berkeley, Calif.: Far West Laboratory for Educational Research and Development, 1975.

Klopfer, L. E. "Evaluation of Learning in Science." In Handbook on Formative and Summative Evaluation of Student Learning, B. S. Bloom, J. T. Hastings, and G. F. Madaus (Eds.). New York: McGraw-Hill, 1971.

Newport, J. F. "Are Science Objectives Changing?" School Science and Mathematics, 65: 359-363, 1965.

Ogden, W. R. "Secondary School Chemistry Teaching, 1918-1972: Objectives as Stated in Periodical Literature." Journal of Research in Science Teaching, 12: 235-246, 1975.

Ogden, W. R. "An Analysis of Published Research Pertaining to Objectives for the Teaching of Secondary School Chemistry, as Reflected in Selected Professional Periodicals 1918-1972." School Science and Mathematics, 74: 120-128, 1974.

Ogden, W. R. "An Analysis of Authorships of Articles Dealing with Objectives of Secondary School Chemistry Teaching 1918-1967." Science Education, 58: 181-184, 1974.

Stake, R. et al. Case Studies in Science Education. Urbana, IL: The University of Illinois, 1978.

TEACHER EDUCATION

Rubba, Peter, "Do Physics Teachers Have Special Inservice Needs?" School Science and Mathematics, 82(4):291-294, April, 1982.

Descriptors--*Educational Research; *Inservice Education; *Inservice Teacher Education; *Physics; Science Education; *Science Instruction; Secondary Education; *Secondary School Science; Teacher Education

Expanded abstract and analysis prepared especially for I.S.E. by William R. Brown, Old Dominion University

Purpose

Two questions were part of this inservice needs assessment. (1) What are the inservice needs identified by at least 65 percent of the physics teachers? (2) Do physics teachers' top inservice needs differ from those of the science teachers in general?

Rationale

The investigator presents the premise that inservice education is more effective when consideration is given to the participating teachers' needs. Teacher opinion of inservice sessions is the definition of effectiveness. One text is cited.

Research Design and Procedure

The sample size was 79 physics teachers out of a total of 403 science teachers who returned needs assessment instruments. The return rate was 41 percent. The sample was stratified across 78 Office of Education Regions in the state. One science teacher was selected at random for every five in the region (20 percent).

The instrument used was the Moore Assessment and Profile. It consists of 117 statements organized under six categories. The reliability is .97 using Hoyt's analysis of variance method. Construct validity was established using factor analysis. The 13 identif able factors accounted for 73 percent of the

87

total variance. Data were reported by percentage of "much needed" and "moderately needed" categories grouped together. A t-test was used to compare the physics teachers' group mean to that of the entire science teacher sample.

Findings

The physics teachers identified 13 inservice needs. The needs can be summarized to indicate that physics teachers desired to gain knowledge and skills which could help them make physics instruction receptive to individual learners.

Six of the 13 needs were shared with all secondary science teachers. Four of the six needs were in the category of better understanding of students.

Interpretations

It would appear that physics teachers share a number of their top inservice needs with secondary science teachers. Physics teachers also identify certain knowledge and skills associated specifically with physics instruction.

ABSTRACTOR'S ANALYSIS

The purpose of this assessment was to determine practicing teachers' perceptions of their own inservice needs. Although self-identification of needs probably helps to make inservice sessions more palatable for most participants, external identification may be necessary to improve teacher effectiveness. For example, if physics teachers are relying heavily on mathematical abstractions, they may be turning-off many students. The teachers may need to be "told" of this problem. Inservice sessions might be necessary to help mathematically oriented physics teachers convert to other approaches. Perhaps inservice sessions should be designed based in part on internal assessment and partly on external input.

Another factor to be considered is how to determine the effectiveness of inservice. Teacher opinion may be one way, but change in teacher-student behaviors may be another gauge of effectiveness.

A couple of questions arise as to how useful is the information collected in this study? Why was 65 percent established as a level to identify major needs? A rationale should be provided to clarify this figure.

The return rate was low, although not unusual for a survey study. What follow-through procedures could have been used to increase the 41 percent return rate? Only 8 percent of the total science teacher population from grades 6-12 responded to the needs assessment instrument (20 percent sampled x 41 percent return rate). Physics teachers contributed 20 percent of the returned instruments (79 out of 403). Although the return rate for physics teachers was high, the overall return rate was low. How much confidence can we assign to extrapolation based on only 8 percent of a population? Needs assessment questionnaires are of little value for making decisions unless measures are taken to assure a high return rate.

IN RESPONSE TO THE ANALYSIS OF

Rubba, Peter. "Do Physics Teachers Have Special Inservice Needs?" by William R. Brown. Investigations in Science Education, 9( ): ,1983.

Peter A. Rubba
Southern Illinois University at Carbondale

In a recent issue of Investigations In Science Education Brown (1983) presented an analysis of the research report, "Do Physics Teachers Have Special Inservice Needs?" (Rubba, 1982). This response has been prepared to clarify five points of confusion which may exist among those who have read the report and Brown's analysis.

First, the author's professional curiosity and the utility of the inservice teacher needs data which would be collected provided provided the rationale for the study. This, the author believes, is implied in the opening sentence of the report's second paragraph.

> In preparation for designing a number of inservice activities for science teachers, the author completed an inservice needs assessment on a random sample of the 4956 Illinois science techers in grades six through twelve. (Rubba, 1982, p. 291)

Second, nowhere in the report is an attempt made to define teacher effectiveness or deal with means for determining the effectiveness of inservice teacher education, nor does the author see the need to do so. The study was a survey of practicing science teachers' perceptions of their inservice needs.

Third, further support for the "...premise that inservice education is more effective when consideration is given to the participating teachers' needs" (Brown, 1983) can be found in Edelfelt and Johnson (1975) and Tyler (1979). Still, it needs to be understood that the author neither stated in the report nor holds the view that teachers' perceptions of their needs provide a sufficient base of information upon which to develop inservice activities. External input also is a necessary source of teacher needs information.

However, it is the author's experience that attention to teachers' perceived needs is particularly valuable in initiating a continuing program of inservice activities and in establishing an atmosphere of trust with a group of teachers. External input on the teachers' needs, whether collected by the inservice educator or another source, is more likely to be accepted as valid by the teachers once the inservice educator has helped the teachers meet their perceived needs.

Fourth, the 65 percent need identification level was established by the author to represent an appropriate level of consensus. Another researcher may have selected a different standard on a similar basis.

Fifth, based upon respondent comments written on a number of the instruments, the author speculates that the low percentage of returns was due, in part, to the length of the Moore Assessment Profile, and in part, to a code number having been placed on the instrument to identify those who would receive another instrument during a second mailing (which occurred two weeks after the first) (Rubba, 1981, p. 273)

REFERENCES

Edelfelt, R. & M. Johnson (Ed.). Rethinking Inservice Education. Washington, D.C.: National Education Association, 1975.

Rubba, P. A. "A Survey of Illinois Secondary School Science Teacher Needs." Science Education, 65(3):271-276, 1981.

Rubba, P. A. "Do Physics Teachers Have Special Needs?" School Science and Mathematics, 82(4):291-294, 1982.

Tyler, R. "Accountability and Teacher Performance; Self-Directed and External-Directed Professional Improvement." In L. Robin (Ed.). The Inservice Education of Teachers: Trends, Processes and Prescriptions. Rockleigh, NJ: Allyn & Bacon, Inc., 1978.