DOCUMENT RESUME

ED 197 939 SE 030 997


AUTHOR Walberg, Herbert J.; And Others
TITLE A Meta-Analysis of Productive Factors in Science Learning Grades 6 Through 12. Final Technical Report.
INSTITUTION Illinois Univ., Chicago. Coll. of Education.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Jun 80
GRANT NSF-SED78-17374
NOTE 339p.

EDRS PRICE MF01/PC14 Plus Postage.
DESCRIPTORS Academic Ability; Academic Achievement; Cognitive Development; Educational Environment; Educational Policy; Educational Research; Elementary Secondary Education; *Literature Reviews; *Science Education; Science Instruction; Sex Differences; Synthesis
IDENTIFIERS *Meta Analysis; Science Education Research



ABSTRACT

Presented is a review and synthesis of research in science education conducted with students in grades 6 through 12 during the years 1964-78. In the meta-analysis process, eight constructs linked to learning outcomes were used as the theoretical constructs for the literature selection. The constructs include quality and quantity of instruction; student ability; motivation; age or developmental level; and home, peer, and classroom environments. Included in this report are nine research-synthesis papers and related appendices. Five papers are focused on the dependence of science learning on one or more of the constructs used in the study, one summarizes the implications of this project for future research syntheses and for conducting future primary empirical studies, while other papers are concerned with sex differences in science learning and the effects of the science curriculum efforts after 1958. Appendices include a discussion of the significance of research synthesis, a computer codebook for characterizing the studies, and an interim report on the project. (Author/PB)

ED197939

78173 74

A Meta-Analysis of Productive Factors in

Science Learning Grades 6 Through 12



Herbert J. Walberg
Research Professor of Education

F. David Boulanger
Assistant Professor of Education

Barbara K. Kremer
Research Associate

Geneva D. Haertel
Research Associate

Thomas Weinstein
Research Assistant

College of Education

University of Illinois at

Chicago Circle

Box 4348

Chicago, Illinois 60680

June, 1980

SE 030 997

| SEDR DATA FORMAT | 1. Award Number<br>SED78-17374 | 2. Award Date<br>September 15, 1978 | 3. Termination Date<br>August 15, 1980 |
|---|---|---|---|

**6. Title**
A Meta-Analysis of Productive Factors in Science Learning in Grades 6 through 12

**4. Amount of Award**
$129,200 (Cum. Amt.)

**5. Type**
Final Tech. Report

**7. Performing Organization**
University of Illinois
Box 4348
Chicago, IL 60680

**8. Pagination**
328 pages

**9. Accession Number**
00373

**11. Principal Investigator; Field or Specialty**
Herbert J. Walberg [Education

**10. Performing Organization Report Number**

**12. NSF Program Manager**
Raymond J. Hannapel

**13. SEDR Program**
RISE

**14. SEDR Subprogram**

**15. Abstract**

The purpose of this project was to provide a review and synthesis of research in science education conducted with students in grades 6 through 12 during the years 1964-78. Eight constructs linked to learning outcomes were used as the theoretical constructs for the literature selection. These constructs include the following: quality and quantity of the instruction; student ability; motivation; age or developmental level; and home, peer, and classroom environments.

Nine research-synthesis papers and related appendices are included in this report. An introductory paper provides an overview of the scope, purpose, method, results, and research recommendations of the project. Five papers deal with the dependence of science learning on one or more of the constructs listed above. One paper summarizes the implications of the project for future research syntheses and for conducting future primary empirical studies. Other papers are concerned with sex differences in science learning and the effects of the science curriculum efforts carried out after 1958. The appendices include a discussion of the significance of research synthesis, a computer codebook for characterizing the studies, and an interim report on the project.

**16. Descriptors**

Science Education
Synthesis
Literature Reviews
Learning
Research Reviews
Evaluation
Educational Policy

**17. Identifiers**
Meta-Analysis

**18. Field**
Multidisciplinary Sciences

**19. Target Audience**
Grades 6-12; Researchers

**20. Availability**

**21. Supplemental Notes**

Table of Contents

## Appendices

# Introduction and Overview

Herbert J. Walberg

F. David Boulanger

Barbara K. Kremer

Geneva D. Haertel

Thomas Weinstein

College of Education

University of Illinois at

Chicago Circle

Box 4348

Chicago, Illinois 60680

June, 1980

This brief introduction and overview is intended to provide an overall perspective for the reader on the nine research-synthesis papers and related appendices that constitute the remainder of this report of research syntheses carried out under the support of the National Science Foundation. The nine papers are each self-contained to a large extent. Most are either in press or submitted to journals and thus contain independent statements of purpose, method, findings, and educational and research implications. Only in this complete report, however, can the interested reader find all the papers in one document together with detailed supportive material about the project that is unlikely to be published in a journal but which, we hope, will save future researchers a great deal of time and effort in carrying out similar research syntheses in education.

As indicated in the Table of Contents, the sections of this report are identified by letter; and the pages are numbered within sections. Thus, for example, this page is the first of the first section and is identified as A 2.

Section B, "Science Education Research," is a brief overview of the scope, purpose, method, results, and research recommendations of the project. It will appear along with 45 other research syntheses in a special 1980 issue of Evaluation in Education: International Progress, edited by Herbert J. Walberg (principal investigator for this project) and Edward H. Haertel. In addition to more specific details on these points, the remaining papers also draw implications of the research syntheses for the improvement to science teaching and learning.

Sections C and D concern the dependence of science learning on student age, ability, and developmental level. The paper on ability has been accepted for publication in the Journal of Research in Science Teaching; and the paper on age and development has been submitted to the same journal.

Sections E, F, and G concern syntheses, instructional techniques, and classroom social environments in relation to science learning. The paper on social and psychological influences has been accepted for publication in Science Education; and the paper on instruction will be published in the Journal of Research in Science Teaching. The paper on social environments has been submitted to Educational and Psychological Measurement.

Section H gathers together the implications of the project for not only future research syntheses in science education but also for conducting future primary empirical studies. The paper has been submitted to the journal Science Education.

The remaining sections go beyond the original proposed scope of the project. The first, a preliminary synthesis of sex differences in science learning, served as the basis of a proposal for a full-scale research synthesis by Barbara Kremer and a collaborator at the University of Illinois at Urbana with the principal investigator of the present project as consultant. The last section is a full-scale research synthesis of the effects of the large science curriculum efforts carried out after 1958. It has been submitted to the Review of Educational Research.

The three appendices provide material that may be valuable to investigators who plan research synthesis of research in science education. The first appendix discusses the potential and significance of research synthesis in education as well as prior efforts. The second appendix is the computer codebook for characterizing the studies, which required a great deal of effort and group discussion. The final appendix contains the interim report on the project.

In conclusion, we wish to acknowledge the support of the National Science Foundation, our project officer Raymond J. Hannapel, and his colleagues Mary Budd Rowe and F. James Rutherford. We are also grateful to James Kulik and Wayne Welch who served as consultants to the project. Perhaps it goes without saying, however, that errors and opinions in this report are strictly our own.

# Science Education Research

Barbara K. Kremer, F. David Boulanger
Geneva D. Haertel, and Herbert J. Walberg
University of Illinois at Chicago Circle
College of Education, Box 4348
Chicago, Illinois 60680

This paper summarizes systematic syntheses, and integrative reviews of 15 years of science education research spanning the years 1964-78 conducted with students in grades 6 through 12. This project was initiated partly in response to recommendations of the NARST-NIE commission on Research in Science Education Report (Yager, 1978) that stated the need for more broadly based theoretical models, research reviews, and field studies to explain science learning. The selection of literature for this synthesis was guided by a psychological model of learning productivity (Walberg, 1980). This model identifies eight constructs that are linked to learning outcomes. The constructs are quality and quantity of instruction; student ability; motivation; age or developmental level; and home, peer, and classroom environments.

The principal goals of the synthesis were to investigate the dependence of science learning on each of the eight constructs represented in the productivity model, to identify promising directions for science education research, and to provide policy makers with a comprehensive, quantitatively based guide to what is known about the major factors influencing science learning.

The limitation of grade levels 6-12 was chosen so as to include the usual range of science course offerings in the United States beginning with required science courses at grade 6 and concluding with the elective program of the high school. The age period represented by these grade levels ecompasses the onset of formal operational thinking (Inhelder and Piaget, 1958). The fifteen-year literaure period was chosen since it represents time of major curriculum reform in science education and a corresponding increase in the quality and quantity of published research,

Five reviews based on the eight constructs in the productivity model were conducted by the authors (the social-psychological constructs motivation, home, and peer-group environment were combined in one review). Features of the literature sampled and major conclusions are summarized in Table 1. In selecting literature for these reviews, the authors examined studies published in referred journals, unpublished and unreferred research reports, and dissertations. Because of the large volume of studies, dissertations were not searched for the quality of instruction construct; and selection was limited to studies with instructional variables represented in five or more published studies within this construct area.

All studies selected for this synthesis related some measure of a construct variable in the model to a science-learning outcome. Where statistical reporting was adequate, quantitative methods of research synthesis, involving effect sizes and correlations, were applied (Glass, 1978). Where statistical reporting was incomplete and therefore precluded the calculation of effect sizes, research findings were synthesized using modified, quantitative techniques including box scores and visual displays of data points.

## A Brief Summary of Results and Recommendations for Research

Table 1 shows that 922 summary numerical-data points such as correlations and effect sizes could be extracted from 151 published studies. A great number of research findings are available in some areas such as the social environment and quality of instruction, but science education research on other important constructs such as motivation, home, and peer environments is meager.

The results summarized in the table speak for themselves, but several overall points seem worth noting here. The results support a key notion of the productivity model--that learning is not a function of only one or a few major constructs, as assumed in much published research, but is consistently correlated and undoubtedly causally implicated with at least eight constructs. With the exception of quantity of instruction, results for which are thin, the findings in science education generally coincide in sign, consistency, and magnitude with previous synthesis conducted in other school subjects, particularly reading and mathematics; and indicate that all eight constructs require consideration in efforts to improve the productivity of academic learning. Current work is devoted to extending the findings, making more explicit comparisons of the productivity of constructs in science and other subjects, making final preparations of detailed technical reports for journal publication, and writing non-technical articles on the synthesis implications for policy and practitioner audiences. The rest of this brief article summarizes methodological recommendations.

Future research should include more consistent reporting procedures, more studies of construct areas slighted in science education research, replication of consistent findings within construct areas, and implementing more rigorous design and sampling procedures. Study reports should routinely include means and standard deviations of all experimental and comparison-group outcomes to make future quantitative syntheses possible, and more comprehensive. The generalizability of individual studies as well as future syntheses of research stand to benefit from greater attention to the description of the populations represented by the sample. This description should at least include the occupational composition of the community, or SES; and whether the community is urban, suburban, or rural in character. It should also include a description of the type of curriculum (whether academic, general, or vocational) in which sampled students are enrolled, and student ethnicity.

The reliabilities of instruments measuring construct variables and science learning outcomes should be reported, including the reliability of treatment implementation in experimental studies. Correlations have been observed to vary as a function of measurement reliability.

Descriptions of instruments, including types of items and content, should be included in reports in order that judgments of validity and learning domain will be facilitated. This point is especially important in the case of unpublished, and locally-developed instruments.

Surveys of literature in motivation, home environment, and peer environment constructs reveal that science education researchers have paid little attention to these variables. Nevertheless, the consistent, positive direction of findings observed in studies of these constructs makes a strong case for their consideration in future research. The consistency and parallelism of results observed in studies of student motivation and home environment with previous work in general education suggests the need for further direct investigation of these constructs as control or stratification factors in studies of curriculum and instruction.

Studies replicating major findings in the quality of instruction, ability, and classroom environment construct areas are recommended. However, the replication of these results need not rigorously follow the details of previous studies. Instead, it is recommended that future studies employ more robust designs incorporating multiple outcomes, and independent variables representing different construct areas. Experimental designs would be improved if such factors as ability, motivation, and classroom environment could be overtly partialed out and their control not be assumed by random assignment. This approach would lead to a better accounting of the sources of variances in outcomes and lead to better prediction and control.

## References

Boulanger, F. D. Ability and science learning: A quantitative synthesis. Journal of Research in Science Teaching, 1980, in press.

Boulanger, F. D. Instruction and science learning: A quantitative synthesis. Manuscript submitted for publication, 1980.

Boulanger, F. D., and Kremer, B. K. Age and developmental level as antecedents of science learning. Manuscript submitted for publication, 1980.

Glass, G.V. Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education, Vol. 5 Itasca: Peacock, 1978.

Haertel, G. D., Walberg, H. J., and Haertel, E. H. Social psychological environments and learning: A quantitative synthesis. Manuscript submitted for publication, 1980.

Inhelder, B. and Piaget, J. Growth of logical thinking from childhood to adolescence. New York: Basic Books, 1958.

Kremer, B.K., and Walberg, H. J. A synthesis of social and psychological influences on science learning. Science Education, 1980, in press.

Walberg, H. J. A psychological theory of educational productivity. In F. H. Farley and N. Gordon (Eds.) *Contemporary Perspectives on Educational Psychology*. Chicago: National Society for the Study of Education, 1980, in press.

Yager, R. E. Priorities for research in science education: A study committee report. *Journal of Research in Science Teaching*, 1978, *15*, 99-107.

Table 1

Science Education Research on

Eight Construct Areas

| Construct(s) Reviewed | Authors | Number of Studies | Source of Articles | Number of Data Points Summarized | Major Conclusions |
|---|---|---|---|---|---|
| Age/Developmental Level | Boulanger and Kremer (1980) | 27 | ERIC science education bibliographies and annual reviews, and articles published in the Journal of Research in Science Teaching and Science Education. Two recent dissertations were included. | 17 median correlations<br>21 median correlations | Age was found to be a poor predictor of conceptual outcomes or logical operations in science achievement. The mean within-grade correlation of developmental level with cognitive achievement was .40. Annual increments in cognitive achievement averaged 10 percentile points, and developmental level 14 percentile points. Interventions to increase increments are reported under the quality of instruction construct. |
| Ability | Boulanger (1980) | 34 | ERIC science education bibliographies and annual reviews, and articles published in the Journal of Research Teaching and Science Education. | 67 median correlations | Relationship between ability and achievement is very stable. Ability accounts for an average of 25% of the variance in science learning. Ability measures are better predictors of cognitive achievement than developmental measures. |
| Motivation, Home Environment, and Peer Influence | Kremer and Walberg (1980) | Motivation/5<br>Home Environment/13<br>Peer Environment/5 | Journal of Research in Science Teaching and Science Education were reviewed for 1964-1979. School Science and Mathematics, Journal of Educational Psychology, Developmental Psychology and Sociology of Education for 1971-1977 were also searched. ERIC and Social Science Citation Index as well as science education bibliographies and annual reviews were also consulted. | Motivation-5 correlations<br>Home-environment-12 study-median correlations<br>Peer environment-5 correlations | The mean correlation for student motivation and science learning was .37. Higher correlations were obtained with standardized scales than with specially constructed measures.<br>Ten out of thirteen studies showed positive relationships between parental socio-economic status and science learning. The mean correlation was .25. Parent education and aspiration, and involvement in the child's science education yielded a correlation of .36 with achievement. |

Table 1
(Cont.)

| Construct(s) Reviewed | Authors | Number of Studies | Source of Articles | Number of Data Points Summarized | Major Conclusions |
|---|---|---|---|---|---|
| | | | | | No consistent trends were observed in the peer construct. Isolated positive effects were found in the few studies located, b.t most showed no effects. |
| Quality of Instruction | Boulanger (1980) | 52 | Published articles found through ERIC science education bibliographies and annual reviews and articles, published in the Journal of Research in Science Teaching and Science Education. Dissertations excluded. | 57 effect sizes | Percentile improvements in cognitive achievement due to interventions were: Preinstructional strategies, 34; training in scientific thinking, 30; high structure verbal content over lower structure, 27; realism or concreteness in adjunct materials, 22. Indirect and inductive strategies showed no differences compared to direct and deductive strategies. |
| Quantity of Instruction | Boulanger (1980) | 3 | Same as Quality of Instruction. | 4 effect sizes | Amount of time spent on a given unit of material holds no significant overall relationship to amount learned in the limited number of studies found. |

Table 1
(Cont.)

| Construct(s) Reviewed | Authors | Number of Studies | Source of Articles | Number of Data Points Summarized | Major Conclusions |
|---|---|---|---|---|---|
| Social Environment of the Classroom | Haertel, Walberg, and Haertel (1979) | 12 | A search was made of 15 years of the Dissertation Abstracts, Education Index, Psychological Abstracts, Social Science Citation Index and the annual research summaries sponsored by the National Association for Research in Science Teaching (1963-1978). | 734 correlations | Correlations reported for science did not differ from those in other subject areas. Learning outcomes and gains correlate positively with Classroom Cohesiveness, Satisfaction, Task difficulty, Formality, Goal Direction, Democracy and Material Environment. Negative correlations are found with Friction, Cliqueness, Apathy, and Disorganization. |

# Ability and Science Learning:
# A Quantitative Synthesis

F. David Boulanger

University of Illinois at Chicago Circle

Running head: Ability and Science Learning

Abstract

The quantitative relationship of measured ability to measured science learning was synthesized from the reported correlations in 34 studies on grade 6 through 12 students over a 16 year period. The findings indicate a stable central tendency and deviation of correlations across ability and cognitive learning outcome categories and across several study variables such as sample size. Reliability of measures had the greatest and only statistically significant influence on ability-cognitive outcome correlations. Ability was found to account for an average of 23 percent of the variance in science learning.

## Ability and Science Learning

### A Quantitative Synthesis

One of the uncontested findings of educational research is the relationship between measures of ability and school learning. In 1930, St. John could conclude, "The intercorrelations of all the criteria of intelligence and educational achievement are without exception positive..." (p. 141). A more recent large scale national survey, Project Talent (Flanagan, Davis, Dailey, Shaycoft. Orr, Goldberg, and Neyman, 1964), reaffirmed this general finding; all reported correlations between measures of ability and achievement were positive. In a review of research on cognitive characteristics that influence learning, Bloom (1976) reported universally positive relationships between past achievement or ability and learning in several subjects areas.

Although consistent in direction, past studies, whether large scale or small, reported different estimates of the size of the ability-learning correlation. A scan of published research in science education reveals a wide variation in correlations of ability with science learning. It appears that the quality of different measures of ability or of science learning at different grade levels and under different study conditions might account for some of the variation in the reported correlations. Correlations may differ with such study conditions as sample size, subject matter, ability level of students and research design.

The purpose of this study is to review, analyze and synthesize published studies relating ability measures to science learning measures in order to establish the best estimates of such correlations under various study conditions. The estimates provide science education researchers and practitioners summary statistics for comparing the ability factor with other factors influencing science learning.

Of particular interest for future reviews and syntheses are the eight constructs enumerated in Walberg's (1978) Productivity Model which draws on the general education empirical literature and provisionally identifies the primary factors influencing general school learning. The constructs are: ability, motivation, and age or developmental level; quality and quantity of instruction; and home, peer, and classroom social environments. The unique features of science instruction such as laboratories, the use of quantitative skills, and the cumulative nature of the subject matter suggest that estimates of correlations of the constructs with general learning outcomes may not be accurate for science learning. The present study provides an estimate of the ability influence on science learning with future studies providing estimates for the other constructs. Such broadly based reviews drawing on general education research findings to inform and augment science education research are nationally identified needs (Yager, 1978).

## Literature Search, Selection, and Coding

To assemble a body of literature reflecting the best current science education research, yet sufficiently extended in time to include the recent period of growth in curriculum development and research, the published research in science education over the 1963-1978 period was included in the literature search. The search was further limited to studies conducted with subjects in grades 6 through 12 to include the pre-college science program from the grade that is typically the beginning of required, specially taught science courses through the elective senior high courses taken by a minority of students.

Ability was initially defined as any cognitive measure that predicts science learning. Using this definition, thirty-four published studies were identified that correlated one or more measures of ability or past achievement with a science learning outcome. Studies including ability measures as blocking variables or as covariates in ACOVA were excluded, unless a zero-order correlation was reported between ability and outcome measures. (Calculated estimates of r from blocking factors were judged inaccurate). The Appendix contains a bibliography of included studies.

All assembled studies were numerically coded according to the following study-variables: the type, source and reliablility of the ability and the outcome measures; the type of intervention; and the elapsed time between measures; grade level, ability level, and science subject area of the sample; the ethnic, urban-rural, and SES character of the community; the design of the study, unit of analysis and methodological flaws; and reported correlations.

In total, over forty study-variables were recorded for each study on prepared code sheets. An independent check by a second researcher of the reliability of coding routinely revealed about 90 percent agreement. Codebook, codesheets, and raw data are available in the project final report (Walberg, Boulanger, Kremer, & Haertel, in preparation).

The coding process yielded three ability categories and four learning outcome categories forming a 3 x 4 or 12 cell ability by outcome matrix. The three ability categories were general ability, prior achievement, and quantitative-spacial reasoning. The four outcome categories were: factual, product, process, and attitudinal learning. Table 1 presents the definition and example measures for each of the ability and outcome categories.

Insert Table 1 about here.

## Results and Analysis

To insure the independence of each correlation in a given ability-outcome cell, each study's median correlation for a given cell was computed, reducing the original 207 raw correlations to 67 median correlations which were used throughout the analysis. When combining correlations, ordinary means and standard deviations were computed following the arguments of Glass (1978) and empirical results of Uguroglu and Walberg (in press) that z-transformations make little difference in means when combining correlations in the range of values of correlations in this study.

Only five of the 67 correlations related ability to attitudes. Based on two (.16 and .28) and three (.30, .24, and .38) correlations respectively, the mean of study-median correlations of general ability and prior achievement with attitudes are .22 and .31. The overall ability-attitude mean correlation is .27 with a standard deviation of .07. Given the small number of correlations, no further analysis of the ability-attitude relationship was attempted.

The three cognitive outcomes (factual, product, and process) were analyzed together. Table 2 shows the means and number of study-median correltaions in each of the nine ability-cognitive outcome cells along with marginals combining correlations across categories. No correlations of quantitative-spacial ability with process outcomes were found which leaves that cell empty. One cell mean, prior achievement with factual outcome, was based on only one correlation; while the general ability with product outcome cell contained the most, 16, correlations. The range of the mean ability-outcome correlations across the eight cells (empty cell excluded) was .41 (prior achievement with factual outcome) to .53 (quantitative-spacial ability with product outcome).

Insert Table 2 about here

A two-way analysis of variance was conducted to determine if the differences among categories were attributable to chance. Main effects (Ability: F = .46, p = .64; Outcome: F = .38, p = .69) and interactions (F = .11, p = .95) were non-significant, leading to the simplification that all three ability categories were, within statistical error, equally good predictors of any of the three cognitive outcomes. Combining the 62 correlations across all cells for the best overall estimate of the ability-cognitive outcome correlation yielded a mean of .48 with a standard deviation of .15.

Since no statistically valid distinction could be made among the correlations relating the various subcategories of ability and cognitive outcome, the analysis of the influence of other study variables such as sample size, subject matter, and study design was conducted on the entire 62 correlation data set. To determine if any study-variable systematically biased the reported ability-outcome correlations, the values of study variables were dicotomized into approximately equal subgroups and the t-test applied to compare each resulting subgroup pair. Study-variables whose values were constant or nearly constant across studies (e.g. mixed sex of sample, individual as unit of analysis) or were rarely reported (e.g. ethnic composition, community type, SES) were dropped from this analysis since it was clear they would not be significantly associated with systematic differences among the correlations. Table 3 reports dicotomized values and t-test results of the variables included.

Insert Table 3 about here

The results in Table 3 indicate only one difference significant at the $P < .05$ level: cognitive outcome measures with reliabilities higher than .80 yielded higher correlations with ability than cognitive outcome measures with reliabilities less than .80. Two variables had differences at the $p = .10$ level: published (usually standardized) ability measures yielded higher correlations with cognitive outcome measures than locally produced ability measures, and higher reliability (greater than .90) ability measures gave higher correlations with cognitive outcome measures than lower reliability (less than .90) ability measures.

In addition to the t-tests reported in Table 3, correlations of continuous study-variables with associated ability-cognitive outcome correlations were computed and are reported in Table 4. Grade level and the reliabilites of the ability and outcome measures show positive relationships with ability-cognitive outcome correlations, but only the reliability of the outcome measure reached significance ($p < .05$)

Insert Table 4 about here

## Discussion and Conclusions

The five ability-attitude outcome correlations gave a mean of .27 with a standard deviation of .07, while the sixty-two ability-cognitive outcome correlations had a mean of .48 with standard deviation of .15. Clearly, ability predicts cognitive outcomes better than attitudinal outcomes, a finding which is not surprising given the cognitive character of ability measures.

Regarding the ability-cognitive outcome correlations, the consistency in correlational means regardless of the ability or cognitive outcome category gave a solid estimate of the degree to which ability is associated with cognitive learning in grades 6 through 12. The .48 mean correlation translates into 23 percent of the variance in cognitive learning accounted for by ability. The standard deviation of .15 means that in 2 out of 3 studies, the variances in cognitive learning accounted for by ability was somewhere between 11 and 40 percent. The stability of the standard deviation and thus of this estimate of variance accounted for, is evident with examination of the SD columns in Table 3, where 19 of 20 SD's are in the range of .13 to .17

According to t-test (Table 3) and correlational (Table 4) results, only one study variable had a significant ($p < .05$) effect on the size of the ability-cognitive outcome correlation while other variables had a minor or no impact. The reliability of the outcome measure had the greatest impact, accounting for 11 percent ($r = .33$) of the variance in the ability-cognitive outcome correlation. This finding is probably related to the higher correlations associated with published outcome measures. The use of published ability measures of high reliability also raised the correlations, although statistical significance was not attained. Both of the above findings are in agreement with the well known tendency for correlations to rise as the reliability of measurement improves, e.g. Iverson and Walberg (1979) found the correlations of the home environment with school learning increased with the reliablility of the outcome measure. The correction for attenuation formula (Thorndike and Hagen, 1977) was developed to correct correlations for this effect. Study variables having no systematic impact on the ability-cognitive outcome correlations were sample size, subject matter, group ability level, and time elapsing between measures.

The primary methodological flaw throughout the 34 studies was the use of convenience sampling which is related to the primary reporting flaw of not sufficiently identifying the population under study. No study provided population parameters of ethnic composition, urban-rural community type, and SES level along with evidence of random sampling of the population. Without this information, generalization of the findings from any individual study is greatly limited. If it is assumed, however, that there is randomness of selection of groups studied across the 34 studies synthesized, then the reported .48 correlation is representative of the grade 6 to 12 population, almost exclusively however, in the United States. This assumption might be questioned given the location of institutions conducting educational research and the tendency of researchers to choose convenient and accessable schools, often in university communities or under some kind of university influence.

To cross-validate the .48 general estimate of the ability-science cognitive outcome correlation found in this study, Educational Achievement in Relation to Intelligence (St. John, 1930) and the Project Talent study (Flanagan et al., 1964) were consulted to find if comparable correlations had been reported. St. John identified eight studies containing 16 correlations between intelligence test scores and teachers' marks in natural science in secondary and higher grades. The mean correlation reported was .46.

Project Talent did not report an ability test score but did identify an *a priori* IQ composite consisting of Reading Comprehension, Abstract Reasoning, and Mathematics 1 test scores. The mean correlation of the IQ composite with Physical Science and Biological Science test scores for grades 9 through 12 was .51.

The two estimates of .46 and .51 are in excellent agreement with the finding of .48 in this study. The congruence of these estimates is even stronger if reliabilities of measures are considered. It can be assumed that teachers' marks will have lower reliabilities than the average of outcome measures used in the 34 studies which yielded the .48 correlation; whereas, Project Talent measures reported reliabilities higher than this average.

## Implications

The tenet that ability and past learning are among the best predictors of future learning is well established among educational researchers and practictioners. What is less well established is the degree to which this tenet is true for different subject areas under different study conditions. The estimates developed in this study should provide the researcher in science education with a guide for estimating the influence of ability on science learning in untested populations, as well as a norm for comparing new findings on the extent to which various factors influence learning. Educational practictioners will find the results of value in moderating their a priori judgements on placement of students in ability groups or raising or lowering expectations for individual students based solely on test scores. The results of this study highlight the fact that measured ability, on average, does not account for a great amount of variance in science learning. Several other factors are known to influence learning and thus compensate for ability differences. Major among these other factors are student motivation; the quality and quantity of instruction; and home, peer and classroom social environments. As improved estimates of the effects of these other factors on science learning become available, science education research and teaching practice can be directed at optimizing those influences most potent in improving science learning, keeping the less manipulable ability factor in proper perspective.

## References

Bloom, B. S. Human characteristics and school learning. New York: Mc Graw Hill, 1976.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., and Neyman, C. A. (Jr.). Project Talent: The American high school student. Pittsburgh: The Project Talent Office, 1964.

Glass, G. V. Integrating findings. The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education (Vol. 6.) Itasca, Illinois: F. E. Peacock, 1978.

Iverson, B. K. and Walberg, H. J. Home environment and learning: A quantitative synthesis. A paper presented at the American Educational Research Association Annual Meeting, San Francisco, April, 1979.

St. John, C. W. Educational achievement in relation to intelligence. In Harvard Studies in Education (Vol. 15). Cambridge: Harvard University Press, 1930.

Thorndike, R. L. and Hagen, E. P. Measurement and evaluation in psychology and education (4th ed.). New York: John Wiley, 1977.

Uguroglu, M. E. and Walberg, H. J. Motivation and achievement: A quantitative synthesis. American Educational Research Journal, in press.

Walberg, H. J. A psychological theory of educational productivity. In F. H. Farley and N. U. Gordon (Eds.), Contemporary perspective in educational psychology. Chicago: National Society for the Study of Education, in press.

Walberg, H. J., Boulanger, F. D., Kremer, B. K. and Haertel, G. D. A meta-analysis of productive factors in science learning in grades 6 through 12, in preparation.

Yager, R. E. Priorities for research in science education: A study committee report. Journal of Research in Science Teaching, 1978, 15, 99-107.

Table 1

Ability and Outcome Categories and Measures

| Category | Definition | Example Measures[1] |
|---|---|---|
| General Ability | General, verbal, or subject matter specific aptitude or ability | Lorge-Thorndike (Johnson, 1969)<br>SAT Verbal (Wasik 1971)<br>IQ from school records (Hardy, 1970) |
| Prior Achievement | Past general or subject matter specific science achievemnt or knowledge | Gr. 9 math achievement (Rothman 1966)<br>Nelson Biology Test (Schock, 1973)<br>SRA Battery (Sheehan, 1977) |
| Quantitative-Spacial Ability | Quantitative, mechanical or spacial ability or reasoning except where specifically based on Piagetian tasks or logical operations | ITED Quantitative (Benson & Howell, 1968)<br>DAT Mechanical Reas. (Tanner 1969)<br>NFER Spacial Test (Marjoribanks, 1978) |
| Factual Learning | Recognition or recall of specific information, e.g., facts, names, definitions | Retention Test (Holliday & Brunner, 1977)<br>Environmental Info. Test (Hart, 1978)<br>Biology Info. Test (Tamir & Jungwirth, 1975) |
| Product Learning | Requires generalization or application of concept (s) to new situations. May also include factual items as in standardized achievement tests. Not identified in study report as process or factual outcome | BSCS Comp. Final (Engen & Smith, 1968)<br>ACS Chem. Exam (Jones, 1963)<br>SCCT Science Comp. (Raven & Polanski, 1974) |
| Process Learning | Requires use of thought processes or logical operations associated with scientific thinking, e.g. hypothesizing, controlling variables. Must be identified in study report as such a measure | Controlling Variables (Bredderman, 1973)<br>Watson-Glaser TCT (George, 1968)<br>Science Process Inv. (Welch & Pella, 1968) |
| Attitudinal Learning | Attitudes toward or interests in scientists, science careers, science instruction | Science Attitude Scale (Engen & Smith, 1968)<br>Environmental Attitude (Hart, 1978)<br>Inventory of Science Attitudes (Swan, 1966) |

[1] Parentheses contain a study using this measure. Studies are listed in the appendix.

Table 2

Ability - Cognitive Outcome

Mean Correlation Matrix

| Ability | Cognitive Outcome: Factual | Product | Process | Combined |
|---|---|---|---|---|
| General Ability | .46 (5) | .49 (16) | .49 (13) | .49 (34) |
| Prior Achievement | .41 (1) | .48 (11) | .42 (7) | .46 (19) |
| Quantitative-Spatial | .49 (3) | .53 (6) | (0) | .51 (9) |
| Combined | .46 (9) | .50 (33) | .46 (20) | .48 (62) |

Notes: Parentheses contain number of study-median correlations used to compute the mean.

Two-way ANOVA (ability by outcome) yielded no signigicant main effects or interactions.

Table 3

Means, Standard Deviations, and t-test

Comparisons of Subgroups of Studies

| Study Variable | Subgroups Compared (1 vs 2) | Subgroup 1 n | Subgroup 1 Mean r | Subgroup 1 SD | Subgroup 2 n | Subgroup 2 Mean r | Subgroup 2 SD | t | p |
|---|---|---|---|---|---|---|---|---|---|
| Sample Size | n< 200 vs n > 200 | 31 | .46 | .16 | 31 | .50 | .14 | 1.10 | .28 |
| Grade Level | 5-9 vs 10-12 | 30 | .47 | .14 | 32 | .49 | .17 | .47 | .54 |
| Subject Matter | Physical Science vs Life and Earth Sci[a] | 22 | .47 | .12 | 24 | .50 | .17 | .71 | .48 |
| Group Ability | High and Above Average vs Average | 25 | .48 | .16 | 31 | .47 | .16 | .24 | .81 |
| Experimental Intervention Between Meas. | yes vs no | 30 | .45 | .13 | 32 | .51 | .17 | 1.48 | .14 |
| Reliability of Ability Measure | R < .90 vs R ≥ .90 | 25 | .45 | .14 | 20 | .53 | .15 | 1.68 | .10 |
| Reliability of Outcome Measure | R < .80 vs R ≥ .80 | 27 | .42 | .13 | 13 | .55 | .14 | 2.83 | .01 |
| Source of Ability Measure | Local vs Published | 11 | .44 | .15 | 48 | .50 | .15 | 1.70 | .10 |
| Source of Outcome Measure | Local vs Published | 29 | .45 | .08 | 33 | .51 | .16 | 1.48 | .14 |
| Time Between Measures | Time < 4 wk vs Time > 4 wk | 20 | .48 | .14 | 40 | .48 | .16 | .02 | .99 |

Note: Dependent variable is the median ability - cognitive outcome correlation per study for each cell in Table 2.

[a] Physical science is physics, chemistry or physical science; life science is biology or life science.

Table 4

Correlations of Continuous Study Variables with Ability-Cognitive Outcome, Study-Median Correlations

| Study Variable | n | r | p |
|---|---|---|---|
| Sample Size | 62 | -.01 | .48 |
| Grade Level | 62 | .07 | .29 |
| Reliability of Ability Measure | 45 | .12 | .21 |
| Reliability of Outcome Measure | 39 | .33 | .02 |
| Time Between Measures | 62 | .01 | .48 |

# Bibliography of Published Studies Correlating Ability with Science Learning: 1963-1978

Benson, B.W. and Howell, J.E. A programmed unit in statistics for secondary science students. School Science and Mathematics, 1968, 68, 691-698.

Bredderman, T.A. The effects of training on the development of the ability to control variables. Journal of Research in Science Teaching, 1973, 10, 189-200

Bredderman, T. Elementary school science experience and the ability to combine and control variables. Science Education, 1974, 58, 457-469.

Crocker, R.K., Bartlett, K.R. and Elliot, H.G. A comparison of structured and unstructured modes of teaching science activities. Journal of Research in Science Teaching, 1976, 13, 267-274.

Doran, R.L. and Sellers, B. Relationships between students' self concept in science and their science achievement, mental ability and gender. Journal of Research in Science Teaching, 1978, 15, 527-533.

Durkee, D. An analysis of the appropriateness and utilization of TOUS with special reference to high-ability students studying physics. Science Education, 1974, 343-356.

Engen, H.B., Smith, D.D. and Yager, R.E. Student outcomes with rotation of teachers in secondary biology. Journal of Research in Science Teaching, 1968, 5, 230-234.

George, K.D. The effect of critical thinking ability upon course grades in biology. Science Education, 1968, 52, 421-432

Hardy, C.A. Chem. Study and traditional chemistry: an experimental analysis. Science Education, 1970, 54, 273-276.

Hart, P.H. Examination of BSCS biology and nonbiology students' ecology comprehension, environmental information level, and environmental attitude. Journal of Research in Science Teaching, 1978, 15, 73-78.

Holliday, W.G., Brunner, L.L., and Donais, E.L. Differential cognitive and affective responses to flow diagrams in science. Journal of Research in Science Teaching, 1977, 14, 129-138.

Johnson, J.O. The relationship between science achievement and selected student characteristics. Science Education, 1969, 53, 307-318.

Jones, K.J. Predicting achievement in chemistry: a model. Journal of Research in Science Teaching, 1963, 1, 226-231.

Kastrinos, W. Teaching for critical thinking in high school biology. Journal of Research in Science Teaching, 1963, 1, 341-350.

Kastrinos, W. The relationship of two methods of teaching to the development of critical thinking by high school students in advanced biology. Science Education, 1964, 48, 187-195.

Linn, M.C., Chen, B. and Thier, H.D. Teaching children to control variables: investigation of a free-choice environment. Journal of Research in Science Teaching, 1977, 14, 249-255.

Majoribanks, K. The relation between students' convergent and divergent abilities, their academic performance, and school-related affective characteristics. Journal of Research in Science Teahing, 1978, 15, 197-207.

Nordland, F.H. and Kahle, J.B. An analysis of the effectiveness of audio-tutorial instruction measured by student achievement and predicted by standardized measures. School Science and Mathematics, 1975, 277-284.

Pella, M.O. and Ziegler, R.E. Use of mechanical models in teaching theoretical concepts. Journal of Research in Science Teaching, 1968, 5, 138-150.

Phillips, D.G. The development of the concept of displacement volume: a hierarchical model and its partial testing under two methods of presentation. Journal of Research in Science Teaching, 1971, 8, 9-19.

Raven, R.J. and Calvey, S.H. Achievement on a test of Piaget's operative comprehension as a function of process-oriented elementary school science programs. Science Education, 1977, 61, 159-166.

Raven, R. and Polanski, H. Relationships among Piaget's logical operations, science content comprehension, critical thinking, and creativity. Science Education, 1974, 58, 531-544

Rothman, A.I. The determination of criteria for the selection of students for the PSSC physics course. Journal of Research in Science Teaching, 1966, 4, 98-101.

Schock, N.H. An analysis of the relationship between cognitive and affective educational objectives. Journal of Research in Science Teaching, 1973, 10, 299-315.

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

Sheehan, D.S. and Hambleton, R.K. A predictive study of success in an individualized science program. School Science and Mathematics, 1977, 77, 13-20.

Smeltz, J.R. A study of the retention of learning in high school chemistry for a period of one year. Science Education, 1966, 50, 359-370.

Swan, M.D. Science achievement as it relates to science curricula and programs at the sixth grade level in Montana public schools. Journal of Research in Science Teaching, 1966, 4, 112-123.

Tamir, P., and Jungwirth, E. Students' growth as a result of studying BSCS biology for several years. Journal of Research in Science Teaching. 1975, 12, 263-279.

Tanner, R.T. Expository-deductive versus discovery-inductive programming in physical science principles. Journal of Research in Science Teaching, 1969, 6, 136-142.

Troxel, V.A. and Snider, C.F.B. Correlations among student outcomes on the Test on Understanding Science, Watson-Glaser Critical Thinking Appraisal, and American Chemical Society Cooperative Examination-General Chemistry. School Science and Mathematics, 1970, 70, 73-76.

Ukens, L.L. and Merrifield, P.R. The structure-of-the-intellect model applied to a COPES learning sequence. Journal of Research in Science Teaching, 1976, 13, 221-225

Wasik, J.L. A comparison of cognitive performance of PSSC and non-PSSC physics students. Journal of Research in Science Teaching, 1971, 8, 85-90.

Welch, W.W. and Pella, M.O. The development of an instrument for inventorying knowledge of the processes of science. Journal of Research in Science Teaching, 1968, 5, 64-68.

# Age and Developmental Level as Antecedents of Science Learning

F. David Boulanger and Barbara K. Kremer
University of Illinois at Chicago Circle

Running head: Age, Development and Science Learning

Abstract

Over the past decade, developmental theory has occupied a central role in science education instructional theory and empirical research. The purpose of the present study is to quantitatively synthesize studies relating age (or grade) and developmental level to science learning among grade 6-12 students over the 1967-1978 period. Twenty-seven studies were reviewed. Annual increments observed in measures of developmental level were consistent with current theory, and annual increments in cognitive achievement were relatively constant over the grade 4-9 interval. Measures of student ability were found to be better predictors of cognitive achievement than developmental measures; and age and grade level were weakly related to developmental level and cognitive achievement, only showing significant correlations across grade levels.

## Age and Developmental Level as Antecedents of Science Learning.

Over the past decade, developmental theory has occupied a central role in science education instructional theory and empirical research. Each annual Summary of Research in Science Education, e.g., Petersen and Carlson (1979), over the 1973-1977 period devoted a separate section to this area of research and focused almost exclusively on Piagetian based studies. Chiappetta (1976) and Levine and Linn (1977) conducted multi-year, qualitative reviews of Piagetian-related science education literature. These included descriptive studies on the general developmental level of various components of the population and on the relationship of training studies to development and achievement. Other than the occasional count of studies reporting a certain kind of result, and the listing of percentages of persons at various developmental stages, no attempt has been made to provide a quantitative synthesis of the findings of related studies.

A quantitative synthesis of studies has the advantages of a more objective process for summarizing each study and a more concise means of displaying and interpreting trends than qualitative approach. The objectivity arises from the use of a numerical coding scheme that provides for ease of replication and tests of agreement among raters. The quantitative summary of the studies allows tables and graphs for concise presentations as well as the use of descriptive statistics.

Another value of quantitative synthesis is the comparability of findings relating different independent or predictor variables to a common dependent variable. For example, the question of whether measures of ability or of developmental level are, in general, better predictors of science achievement could be addressed through quantitative synthesis. Additional comparisons might be made with other major influences on science achievement.

Walberg (in press) has identified eight constructs in the general education literature as substantially related to learning. The eight constructs are: student ability, motivation, age or developmental level; quality and quantity of instruction; and home, peer, and classroom environments. The relationship of the constructs to student learning in science is the theme of other quantitative syntheses concurrent to the present study. This comprehensive view of influences on science learning based on general education literature is in harmony with the recommendations of the NARST - NIE Commission on Research Priorities in Science Education (Yager, 1978). A general report of findings in all construct areas is in preparation (Walberg, Boulanger, Kremer & Haertel, in preparation).

## Purpose and Scope

The purpose of the present study is to quantitatively synthesize studies relating age (or grade) and developmental level to science learning among grade 6-12 students over the 1963-1978 period. The grade levels 6-12 were chosen to focus on that interval in the school curriculum that typically begins with the first required junior-high school science courses and concludes with elective, senior-high school courses. This interval is also characterized in Piagetian theory as the period of transition from concrete to formal thinking. The science education literature of the 1963-1978 time period reveals the emergence of the developmental perspective in educational psychology and the most recent growth period in the quantity and quality of science education research.

## Methodology

The assumptions and procedures advocated by Glass (1978) were adopted for this synthesis. Glass argues that all studies have flaws and limitations and that combined results give better estimates of outcomes and trends than any single flawed study. A weakness in one study is often balanced by a strength in another; an effect or relationship persisting across diverse studies on a variety of populations is more robust than any single result.

In the present synthesis, the individual study results of interest were either correlations or effect sizes. Zero-order correlations between similar predictor variables and similar outcome variables were recorded as data points for analysis. Where two age levels or, more often, two grade levels were compared, an effect size (ES) was calculated using one of two formulas:

$$ES = \frac{\overline{X}_H - \overline{X}_L}{S_H} \qquad ES = t\sqrt{\frac{1}{n_H} + \frac{1}{n_L}}$$

$\overline{X}_H$ and $\overline{X}_L$ are the dependent variable means of the higher and the lower grades respectively. $S_H$ is the standard deviation of the higher grade scores. t is the computed t-test statistic and the n's are group sizes. An F-ratio comparing two groups was considered equal to $t^2$ and, $\sqrt{MS_w}$ was considered equal to $S_H$. F-ratios based on comparisons of more than two groups were not used in computing effect sizes.

## Literature Search and Selection

The goal of the literature search and selection was to identify two kinds of grade 6 through 12 studies in the 1963 through 1978 science education literature: 1) studies that reported a correlation of age, grade, or developmental level with some measure of science learning, and 2) studies that reported measures of developmental level or science learning at two different grade levels in a manner that allowed computation of an effect size. The search had three components: scanning of all available ERIC annual reviews of science education research for the period; a study-by-study search of all 1963-1978 volumes of the Journal of Research in Science Teaching and Science Education; and a computer search of Dissertation Abstracts and Social Sciences Citation Index for the period in question.

Since the literature search revealed only Piagetian-based studies, the definition of developmental level was limited to any measure of Piagetian stage or related logical operations whether obtained via interview techniques (e.g., Lawson & Blake, 1976) or other measure validated against Piagetian theory (e.g., Raven & Polaski, 1974). Among the studies meeting the selection criteria, developmental level appeared as a predictor variable for cognitive achievement, as a criterion variable for age or grade predictors, and as a dependent variable in grade level comparisons. Cognitive achievement was defined as any measure of factual and/or conceptual learning of science content, while science process learning was restricted to scores on the Science Process Inventory (Welch & Pella, 1968). The above definitions of developmental level, cognitive achievement, and science process learning evolved with the selection and coding of studies. A total of 27 studies met the selection criteria.

## Coding

All assembled studies were numerically coded according to the following study variables: the type, source and reliability of independent and dependent measures; grade level, ability level, and science subject area of the sample; the ethnic, urban-rural, and SES character of the community; the design of the study, unit of analysis,and methodological flaws; and reported correlations or computed effect sizes. In total, over forty study variables were recorded for each study on prepared code sheets. An independent check by a second researcher of the reliablility of coding routinely revealed about 90 percent agreement. Codebook, codesheets, and raw data are available in the project final report (Walberg, Boulanger, Kremer & Haertel, in preparation).

## Analysis and Results

Study code sheets were sorted in terms of similarity of independent and dependent variables and type of summary statistic, i.e., correlation or effect size. The resulting five classifications and associated summary tables are: correlations of developmental level with cognitive achievement--seven studies in Table 1; correlations of age or grade with developmental level or cognitive achievement--six studies in Table 2; grade level comparisons (effect sizes) in terms of developmental level, cognitive achievement, and science processes--15 studies all in Table 3. If a study reported more than one effect size for a given dependent variable category and grade level comparison, the median effect size was identified for later analysis. Likewise, if a study reported more than one correlation in a given predictor and criterion category, the median correlation was selected for analysis. Median values were used to insure independence, since multiple effect sizes or correlations from the same study population would be related. An annotated bibliography of studies by category is provided in the Appendix.

Insert Tables 1, 2 & 3 about here

Mean correlations of developmental level with cognitive achievement (Table 1) rise from .28 in grade seven to .63 in grade 9, and decline to .32 in grade 12.

The grand mean is .40 with a standard deviation of .11. Only one study (Sayre & Ball, 1975) reported correlations at each grade, 7-12, based on the same measures which were Piagetian interviews (developmental level) and student grades in science (cognitive achievement). Figure 1 is a plot of the Table 1 mean values against grade level and the Sayre and Ball data against grade level. The plot indicates that the trend in the Sayre and Ball data is maintained by the other studies in grades 10 through 12.

Insert Figure 1 about here

The grade 7 through 9 correlations are based on data from required junior high courses; while grade 10 and 11 mean correlations are from three biology course-related and three chemistry course-related situations, respectively. The grade 12 data is from one physics based study and from a group of British fifth and sixth form students.

It might be hypothesized, from a developmental perspective, that the increasing correlation over grade 7 through 9 required courses is due to differing developmental rates causing an increase in variation within classes as they move from seventh to ninth grade. The decline in correlations from grade 10 through 12 most likely is due to the self-selection of students in the elective advanced science courses, diminishing the variation within classes by removal of the cognitively less-developed, and lower-achieving students. However, both explanations are vulnerable to competing interpretations, such as changes in interest and motivational factors which influence performance on both developmental and cognitive achievement measures.

The correlations in Table 2 of age or grade with developmental level range from .00 when based on the ages of grade 11 science students, to .57

when data spanning six grade levels (grade 4 through 9) is included. These correlations emphasize the inappropriateness of strongly associating age or grade with levels of intellectual development or ability to use logical operations. The .57 correlations would mean that only about 30 percent of the variation in developmental level across the developmentally diverse series of grade levels is accounted for by grade or chronological age. As will be seen in the next section, the low correlations of grade with developmental level may be explained by the fact that within-class variation is greater than between-class variation. Table 2 also indicates that age or grade level is a poor predictor of cognitive achievement.

When studying the calculated effect sizes in Table 3, it should be noted that a mean effect size comparing one grade level to the next is simply the difference in means between the lower and higher grade converted into standard deviation units. The distribution of the higher grade's score is assumed to be normal and the lower grade's mean is to the left of the higher grade's central mean on the normal curve by the amount of the effect size.

The grade comparison mean effect sizes presented in Table 3 are best visualized by plotting the cumulative mean effect size against grade level. The incremental effect size to be added each year is based on the average of the mean effect sizes which apply to the grade interval in question. For example, examination of the first entries in the far right column in Table 3 in conjunction with the far left column will indicate that .261 and .399 yearly increments both apply to the grade 5 to 6 interval and thus should be averaged when plotting the grade 5 to 6 increment. Following this method of calculation, Figure 2 parts a, b, and c displays three plots of the cumulative effect sizes of developmental level, cognitive achievement, and science process learning respectively over grade levels in the data being plotted.

Insert Figure 2 about here

While inspecting Figure 2, certain quantities and trends should be noted. Based on six and ten studies respectively, both Figure 2a and Figure 2b show fairly smooth curves with gradually increasing increments in the case of developmental level and relatively constant increments in the case of cognitive achievement. The increasing developmental increments are in correspondance with developmental theory which poses a concrete operational to formal operational transition beginning about grade six for many children. Even if individual transitions were fairly sharp for most children, group data would show only a gradual upward swing of the mean accompanied by the increased variation noted earlier in the correlational results. The linearity of the cognitive achievement cumulative effect size over the same grade intervals suggests that the developmental upward swing is not simply an artifact of increasing achievement.

A second trend worth noting is the relationship between within-class and between-class variation. Developmental effect size increments sum to .932 between grades 4 and 7. This means that the average seventh grade student is approximately one standard deviation above the average fourth grade student on developmental level measures. Thus, the upper 16 percent of the fourth grade is developmentally above the median level of the seventh grade. The between class variation is small compared to the within-class variation. Similar statements can be made about cognitive achievement; e.g., a change of nearly four grade level mean values is analogous to a change of one standard deviation (one effect size unit) of within class variation.

The cumulative effect size of science process learning against grade (Figure 2c) is more irregular than the other two plots. The large gain (.754) in the grade 9 to 10 interval is based on one study (Tamir, 1972) where knowledge of science processes was measured at the end of each grade level year. The gain is largely a measure of the effects of tenth grade science, the character of which is unclear from the study report. The irregularity of the plot in general may be an artifact of combining results of only two studies (Tamir, 1972, & Welch and Pella, 1968) conducted in quite different educational systems (Israel and Wisconsin, respectively).

The mean annual effect size increments for the three Figure 2 plots are: developmental level, .36; cognitive achievement, .28; and science process learning, .43. Expressed as percentiles, the increments indicate the approximate advance of the mean class score each year from the previous year's 50 percentile point. Average yearly percentile increases would be: developmental level, 14; cognitive achievement, 11; and science process learning, 17.

The analysis to this point has focused only on correlations and effect sizes and their relationship to grade levels. Additional information about each study was coded to provide normative values and to determine if study variables such as instrument reliability, sample size, etc. had any across study systemtaic influence on correlations or effect sizes.

The reliablilities of cognitive achievement, developmental level and science process measures were comparable in average values (.73, .72 and .76 respectively) and were unrelated to either correlation or effect size values. The ranges of reported reliabilities were .49 - .98, .50 - .92, and .74 - .79 respectively. Only 12 of the total 27 studies reported instrument reliabilities.

All developmental level measures were research measures, ie., not standardized over any representative local, regional or national sample. Written measures of logical operations, eg., Raven and Polanski (1974) were, in general, more reliable than task based measures, eg., Lawson and Blake (1976). Four studies with written measures yielded a mean reliability of .79, while five studies with task measures yielded .67. Assessing the validity of either kind of measure is difficult since both deviate from the Piagetian clinical approach and are analyzed in terms of parametric statistics; yet, the content of both kinds of measures is founded in Piagetian theory.

Among correlational studies relating developmental level to cognitive achievement, average or heterogenious groups registered higher correlations (eight correlations with mean of .45) than high ability groups (four correlations with mean of .31). This trend is related to the self-selection in higher grade levels referred to earlier. The high ability groups are all in elective eleventh and twelth grade courses.

Several study variables, eg., population demographics, were too infrequently reported for analysis. Sample size was reported for all studies but bore no relationship to correlation or effect size values.

Threats to the validity of study designs were primarily of two kinds: convenience sampling which threatened generalizability, and use of cross-sectional data in grade level comparisons. No longitudinal study was found which traced the development of a group of students over a period of time (other than pre and post measures bracketting an instructional treatment).

Developmentally related instructional studies are reported in Boulanger, 1979b.

The usual caution in the interpretation of all tables, plots, and quantitative values presented above is appropriate here. All figures and interpretations are based on a relatively small number of diverse studies. The case for this kind of quantitative synthesis rests on the argument that the combined results carry more general validity than any single study, as well as showing trends not apparent in studies considered singly. All the above interpretations should be considered hypotheses for further investigation; all average correlations and effect sizes should be considered as only tentative norms based upon data available in the 1963-1978 period.

## Discussion

The grand mean correlation of .40 between developmental level and cognitive achievement might be compared to the correlation between ability measures and cognitive achievement reported in another research synthesis (Boulanger, 1979 a). Ability was defined as any measure of prior achievement, general ability or quantitative-spacial ability. The mean correlation between ability and cognitive achievement was .48 with a standard deviation of .15, significantly ($p<.01$) higher than the developmental-level-as-predictor correlation reported in this study. Since general ability or prior achievement measures are usally available in school records, the value of administering time-consuming developmental measures for achievement prediction, in general, makes little sense unless it can be shown that developmental measures account for significant amounts of unique variance not accounted for by ability measures. However, the more common defense for the use of developmental measures is the diagnostic value of knowing student capabilities in the various kind of theory related logical operations. Ability measures may tap many of the same skills, but developmental measures make logical operations and student weaknesses in applying them more explicit on an individual basis.

Another research synthesis (Boulanger, 1979 b) which examined the effects of training in scientific thinking skills has implications for the findings of this study. In the present study, the annual mean percentile gain in developmental level was found to be approximately 14 percentile points in the grade 4-9 interval with the annual increment increasing in the higher grades (Figure 2a). Based on 11 training in scientific thinking studies, 9 of these training in Piagetian logical operations, a mean effect size of .89 or 30 percentile points was found when trained groups were compared with untrained control groups. These training effects occurred in grades 5 through 9 primarily as a result of short term (two to ten hours) tutorial type training of individual students by special teachers. Long term effects of the training were not investigated in the studies; but the studies strongly suggest that the annual increments in such developmentally related traits as logical reasoning patterns can be increased with appropriate instruction.

## Summary

Twenty-seven studies were identified in the 1963-1978 science education research on students in grades 6 through 12. The studies related age or grade, developmental level and science learning in terms of either correlations or computed effect sizes. Major findings were:

a) The mean within grade level correlation of developmental level with cognitive achievement is .40, with individual grade level correlations reach-

ing a maximum in grade nine.

b) Ages and grade level are weakly related to developmental level and cognitive achievement, only showing significant correlations when computed across several grade levels.

c) Annual increments in developmental level effect size average .36 (14 percentile points) and increase over the grade 4-9 interval in agreement with developmental theory. Training studies reported elsewhere indicate that it may be possible to increase these increments through carefully designed instruction.

d) Annual increments in cognitive achievement are relatively constant at an average of .25 (10 percentile points) over the grade 4-9 interval.

e) Ability measures are better predictors of cognitive achievement than are developmental measures.

## Recommendations

Piagetian based developmental measures are founded in hypothesized intellectual structures and operations which emerge in stages over the years of childhood and adolescense. Traditional ability measures are norm referenced and are founded in observed reasoning skills often in the context of culturally defined situations. Both kinds of measures correlate with culturally defined cognitive achievement. To sort out the unique contribution of each kind of measure to the prediction of science learning, both should be administered and later related to both cognitive achievement (as defined in this study) and developmental growth. A longitudinal series of such measures over a period of years would allow the tracking of both individual and group absolute progress in intellectual development and relative standing on ability and achievement measures. This would allow verification of correlational and effect size trends described earlier. Planned, developmentally oriented instructional interventions with selected subsamples would provide time series data to be collected on the short and long term value of such interventions on both development and achievement.

The weak point in the above plan is the present set of developmental measures. A first research priority is the creation of a series of valid and reliable developmental measures which provide quantitative indicators of developmental level comparable over the full range of developmental stages. The measures should account for significant unique variance in science learning when compared to ability measures in order to justify the time and expense of administration. The measures should also possess diagnostic properties to provide direction to the developmental aspects of subsequent instruction.

Chronological age and school grade remain rough indicators of developmental level and science learning and will continue to be routinely recorded for a variety of organizational and cultural reasons. Age is probably better related to physical maturity, general life experience, and broad psychosocial life stages than to intellectual development, and, even less so, to science learning.

Acknowledgement

This research was supported by the National Science Foundation Grant No. 78-17374, Principal Investigator Herbert J. Walberg. Findings and opinions expressed in this paper are solely the responsibility of the author, who wishes to acknowledge the critical contributions of James A.

Kulik, University of Michigan and Wayne W. Welch, University of Minnesota and the comments of colleagues Geneva D. Haertel, Barbara K. Kremer, and Herbert J. Walberg, who are synthesizing research on other factors influencing science learning.

## References

Boulanger, F. D. Ability and science learning: a quantitative synthesis. Manuscript submitted for publication, 1979 a.

Boulanger, F. D. Instruction and science learning: a quantitative synthesis. Manuscript submitted for publication, 1979 b.

Chiappetta, E. L. A review of Piagetian studies relevant to science instruction at the secondary and college level, Science Education, 1976, 60, 253-261.

Glass, G. V. Integrating findings. The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education (Vol. 6). Itasca, Illinois: F. E. Peacock, 1978.

Lawson, A. E. and Blake, A. J. D. Concrete and formal thinking abilities in high school students as measured by three separate instruments. Journal of Research in Science Teaching, 1976, 13, 227-235.

Leon, L. O. The principle of conservation or invariance and its relationships to achievement in science in the junior high school. ED 091 145, 1975.

Levine, D. I. and Linn, M. C. Scientific reasoning ability in adolescence: theoretical viewpoints and educational implications. Journal of Research in Science Teaching, 1977, 14, 371-384.

Petersen, R. W. and Carlson, G. R. A summary of research in science education-1977. New York: John Wiley, 1979.

Raven, R. and Polaski, H. Relationship among Piaget's logical operations, science content comprehension, critical thinking, and creativity. Science Education, 1974, 58, 531-544.

Sayre, S. and Ball, D. W. Piagetian cognitive development and achievement in science. Journal of Research in Science Teaching, 1975, 12, 165-174.

Tamir, P. Understanding the process of science by students exposed to different science curricula in Israel. Journal of Research in Science Teaching, 1972, 9, 239-244.

Walberg, H. J. A psychological theory of educational productivity. In F. H. Farley and N. U. Gordon (Eds.), Contemporary perspective in educational psychology. Chicago: National Society for the Study of Education, in press.

Walberg, H. J., Boulanger, F. D., Kremer, B. K. and Haertel, G. D. A meta-analysis of productive factors in science learning in grades 6 through 12. In preparation.

Welch, W. W. and Pella, M. O. The development of an instrument for inventorying knowledge of the processes of science. Journal of Research in Science Teaching, 1968, 5, 64-68.

Yager, R. E. Priorities for research in science education: A study committee report. Journal of Research in Science Teaching, 1978, 15, 99-107.

Table 1

Mean Correlation by Grade of Developmental Level with Cognitive Achievement

| Grade [a] | Number of Median/Correlations [b] | Mean Correlation |
|---|---|---|
| 7 | 1 | .28 |
| 8 | 1 | .31 |
| 9 | 1 | .63 |
| 10 | 3 | .47 |
| 11 | 3 | .36 |
| 12 | 2 | .32 |
| | Grand Mean | .40 |

[a] One study (Leon, 1975) reported a correlation of .48 based on combined grade 7-9 data.

[b] The total number of studies represented in this table is seven. One study (Sayre & Ball, 1975) reported six correlations, one at each grade level.

Table 2

Correlation of Age or Grade with
Developmental Level or Cognitive Achievement

| Grades Included in Data | Number of Correlations[a] | Age or Grade Correlation with | |
|---|---|---|---|
| | | Devel. Level | Cogn. Achieve. |
| 4,5,6 | 1 | | .01 |
| 4,5,6,7,8,9 | 1 | .57 | |
| 4,6,8,10 | 1 | .39 | |
| 7 | 1 | -.03 | |
| 11 | 1 | | -.11 |
| 11 | 1 | .00 | |

[a] The total number of studies represented in this table is six.

Table 3

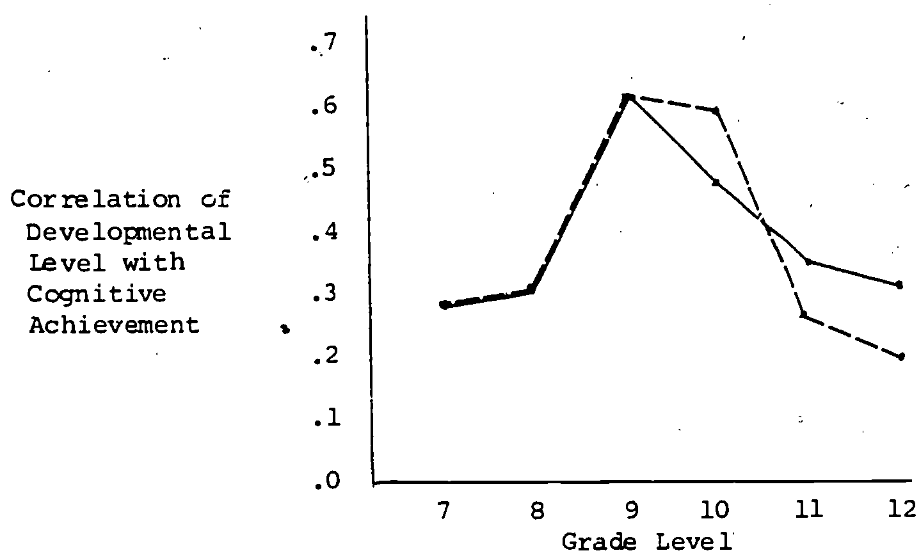
Grade Comparison
Mean Effect Size by Outcome

| Grades Compared | Number of Median Comparisons [a] | Mean Effect Size | Mean Effect Size Per Year |
|---|---|---|---|
| | Developmental Level Outcome | | |
| 4,6 | 1 | .521 | .261 |
| 5,7 | 2 | .797 | .399 |
| 6,8 | 2 | .565 | .283 |
| 7,9 | 2 | .966 | .483 |
| | Cognitive Achievement Outcome | | |
| 4,6 | 4 | .547 | .274 |
| 5,7 | 1 | .525 | .263 |
| 7,9 | 2 | .575 | .288 |
| 9,10 | 2 | .142 | .142 |
| | Science Process Outcome | | |
| 9,10 | 1 | .754 | .754 |
| 10,11 | 2 | .086 | .086 |
| 11,12 | 2 | .442 | .442 |

[a] The total number of studies represented per section are: Developmental Level, 6; Cognitive Achievement, 8; Science Process Outcome; 2. One study appears in two sections; total for table is 15.

Figure 1

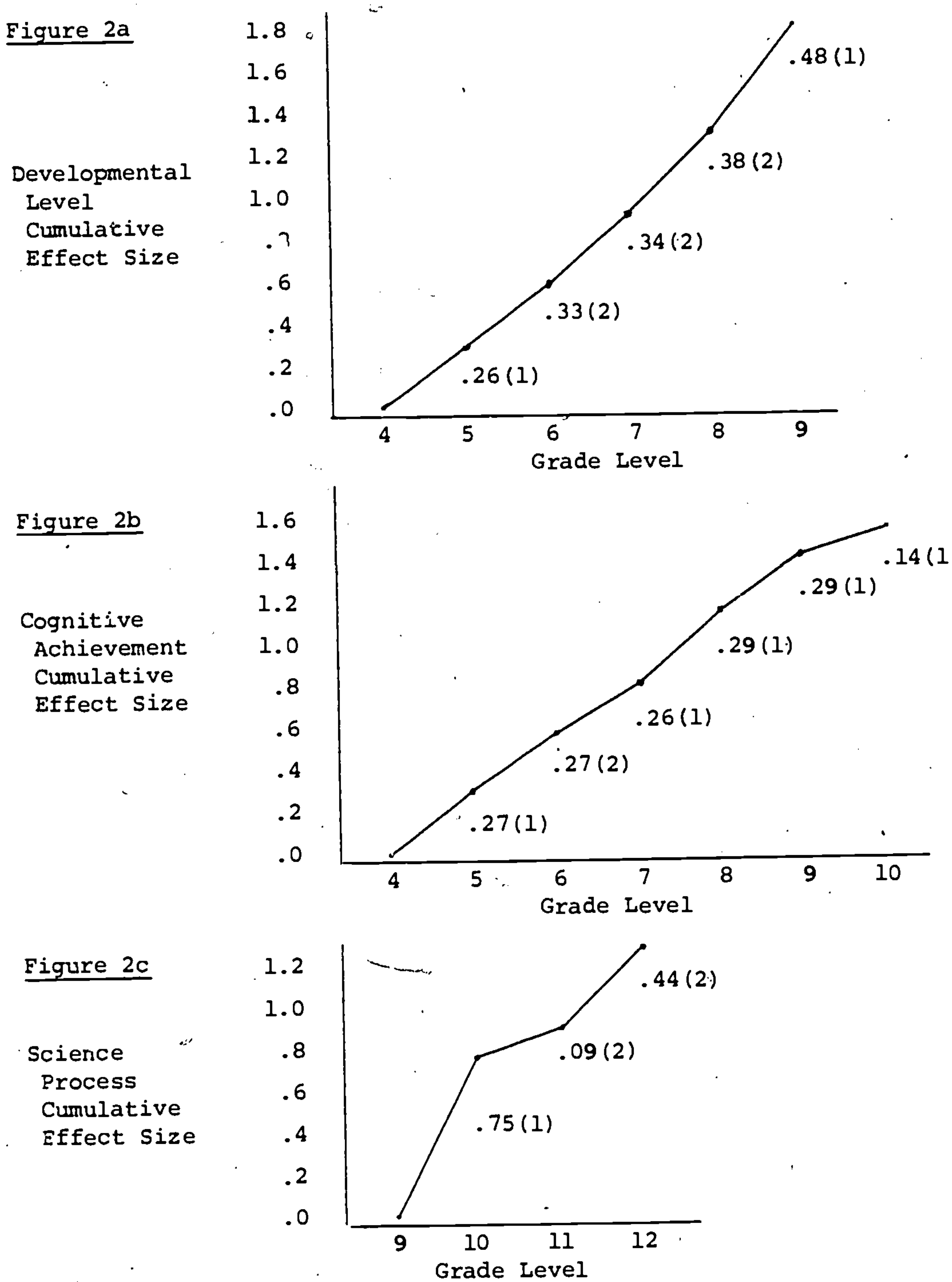
Developmental Level - Cognitive Achievement
Correlation versus Grade Level




Note. Solid line connects Table 1 mean correlations. Dashed line connects data points from Sayre and Ball (1975) who reported a correletion for each grade.

Figure 2

Cumulative Effect Size Based on
Annual Grade Interval Effect Sizes




Note. Number to the right of each line segment is the mean effect size increment with contributing number of values in parentheses.

# Appendix

## Annotated Bibliography of Studies on Development and Science Learning

### Effect Size Studies (Age/grade and developmental level)

Hammond, J. & Raven, R. The effects of a structured learning sequence on the achievement of compensatory tasks. Journal of Research in Science Teaching, 1973, 10, 257-262.

55, grade 6 - 8 students grouped into three ability levels (within grade), randomly assigned to control and programmed instructional groups in compensatory operations. Experimental instructional groups scored higher on a post-test than did control groups.

Lawson, A.E. & Blake, A.J.D. Concrete and formal thinking abilities in high school students as measured by three separate instruments. Journal of Research in Science Teaching, 1976, 13, 227-235.

32 biology students were administered tasks measuring Piagetian Stage, and a test of understanding of concrete and formal biology concepts. Performance on concepts tests varied significantly as a function of stage.

Lewis, W.R. The influence of age, sex, and school size upon the development of formal operational thought. Unpublished doctoral dissertations, University of Oklahoma, 1972.

574 junior and senior high school students were individually administered six Piagetian tasks. Significant differences were observed between grades separated by two or more years, but no significant differences in groups separated by one grade.

Nous, A. & Raven, R. The effects of a structured learning sequence on children's correlative thinking about biological phenomena. Journal of Research in Science Teaching, 1973, 10, 251-255.

246, grade 5, 7, & 9 students receiving identical instruction on correlational thinking. Performance on a post-test varied significantly with grade.

Raven, R. & Polanski, H. Relationships among Piaget's logical operations, science content comprehension, critical thinking, and creativity. Science Education, 1974, 58, 531-544.

Performance of 220 grade 4 & 6 students were compared on tests of general science achievement and critical thinking. A significant difference between grade levels (favoring grade 6) was observed.

Raven, R. J. & Calvey, S. H. Achievement on a test of Piaget's operative comprehension as a function of process - oriented elementary school science programs. Science Education, 1977, 61, 159-166.

Performance of 249 grade 6 & 8 students on a test of logical operations was compared. A significant difference between grade levels (favoring grade 6) was observed.

Effect-Size Studies (Age/grade and science process achievement)

Tamir, P. Understanding the process of science by students exposed to different science curricula in Israel. Journal of Research in Science Teaching, 1972, 9, 239-244.

3500 Israeli grade 9 - 12 students were administered the Welch Science Process Inventory. Norms for Israeli students were established.

Welch, W. W. & Pella, M. O. The development of an instrument for inventorying knowledge of the processes of science. Journal of Research in Science Teaching, 1968, 5, 64-68.

839, grade 10 - 12 students were administered a test of science processes (SPI). No significant differences between grade levels were observed.

Effect Size Studies (Age/grade and cognitive achievement)

Doran, R. L. Misconceptions of selected science concepts held by elementary school students. Journal of Research in Science Teaching, 1972, 9, 127-137.

253, grade 2 - 6 students were administered a test of science misconceptions. Mean test scores increased with grade.

Kauchak, D., Eagen, D., & Kirk, S. The effect of cue specificity on learning from graphical materials in science. Journal of Research in Science Teaching, 1978, 15, 499-503.

82, grade 4 - 6 students randomly assigned to three treatments: Cued questions, non-cued questions, and generalizing questions in passages about plant growth. Performance increased significantly with grade.

Lawson, A. E. & Blake, A. J. D. Concrete and formal thinking abilitites in high school students as measured by three separate instruments. Journal of Research in Science Teaching, 1976, 13, 227-235.

32 biology students were administered tasks measuring Piagetian Stage, and a test of understanding of concrete and formal biology concepts. Performance on concepts tests varied significantly as a function of stage.

Pederson, A.A. & Jacobs, J. E. The effect of grade level on achievement in biology. Journal of Research in Science Teaching, 1976, 13, 237-240.

Performance of 684 grade 9 & 10 biology students was compared on a local achievement test. No significant differences observed.

Pella, M. O. & Triezenberg, H. J. Three levels of abstraction of the concept of equilibrium and its use as an advance organizer. Journal of Research in Science Teaching, 1969, 6, 11 - 21

270, grade 7 & 9 students randomly assigned to three advance organizer treatment groups. A significant difference in performance between grade levels on a test of factual knowledge was observed. No differences were observed among treatment groups.

Raven, R. & Polanski, H. Relationships among Piaget's logical operations, science content comprehension, critical thinking, and creativity. Science Education, 1974, 58, 531-544. Performance of 220 grade 4 & 6 students were compared on tests of general science achievement and critical thinking. A significant difference between grade levels (favoring grade 6) was observed.

Voelker, A. M. Elementary school children's attainment of the concepts of physical and chemical change--a replication. Journal of Research in Science Teaching, 1975, 12, 5 - 14. Performance of 40 grade 4 - 6 students on a post-test of concepts of physical and chemical change was compared (experimental and control groups within each grade had previously received instruction). A significant difference between grade levels (favoring grade 6) was observed.

Walters, L. L. Ninth vs. tenth grade biology--a comparison of achievement. Journal of Research in Science Teaching, 1963, 1, 170-176.

Performance of 144 grade 9 & 10 students on the Nelson Biology Test was compared. No significant differences were observed.

Correlational Studies (Developmental level and cognitive achievement)

Cantu, L. L. & Herron, J. D. Concrete and formal Piagetian stages and science concept attainment. Journal of Research in Science, 1978, 15, 135-143.

16 chemistry students identified as formal operational, and 12 as concrete operational were administered tests of concrete and formal concepts, following instruction. Formal operational students performed significantly better.

Fields, T. W. & Cropley, A. J. Cognitive style and science achievement. Journal of Research in Science Teaching, 1969, 6, 2 - 10.

178, fifth and sixth form students were administered tests of Piagetian operations and science achievement. Level of cognitive operations were found to be significantly correlated with achievement.

Lawson, A. E. & Nordland, G. H. Conservation reasoning ability and performance on BSCS blue version examinations. Journal of Research in Science Teaching, 1977, 14, 69 - 75.

23 biology students were dministered tests of Piagetian conservation and the BSCS achievement test. Significant correlation between test performance and conservation.

Lawson, A. E. & Blake, A. J. D. Concrete and formal thinking abilities in high school students as measured by three separate instruments. Journal of Research in Science Teaching, 1976, 13, 227-235.
32 biology students were administered tasks measuring Piagetian Stage, and a test of understanding of concrete and formal biology concepts. Performance on concepts tests varied significantly as a function of stage.

Leon, L. O. The principle of conservation or invariance and its relationship to achievement in science in the junior high school. ED 091 145, 1975.
182 grade 7 - 9 students were administered the STEP test in science and a test for conservation of quantity. Significant correlation observed between ability to conserve and achievement in science.

Rubley, V. D. An investigation of formal thought and dogmatism during the transition between adolescence and adulthood. Unpublished doctoral dissertation, University of Iowa, 1972.
60 high school chemistry students were administered Piagetian tasks and the ITED background in the natural sciences test. No correlation between age and test performance.

Sayre, S., & Ball, D. W. Piagetian cognitive development and achievement in science. Journal of Research in Science Teaching, 1975, 12, 165-174.
352 junior and senior high school science students were administered Piagetian tasks. Significant correlation between grade in science and tasks performance.

Correlational Studies (Age/grade with cognitive achievement and developmental level.)

Bredderman, T. Elementary school science experience and the ability to combine and control variables. Science Education, 1974, 58, 457-469.
80, grade 4,6,8 & 10 students were administered a test on controlling and combining variables. Significant correlation between age and test performance was found.

Gunnels, F. G. A study of the development in logical judgements in science of sucessful and unsuccessful problem solvers in grades four through nine. ED 026 249.

Inferences drawn by students in grades 4-9 from science texts were related to Piagetian levels of thought. Older students and those at higher grade levels were found to operate more frequently at formal levels of operational thought.

Hardy, C. A. Chem study and traditional chemistry: an experimental analysis. Science Education, 1970, 54, 273-276. Performance of 208 chemistry students and traditional chemistry students were compared on tests of standardized achievement and critical thinking. Ability and past achievement significantly correlated with post-test chemistry achievement.

Nordland, F. H., Lawson, A. E., & Kahle, J. B. A study of levels of concrete and formal reasoning ability in disadvantaged junior and senior high school science students. Science Education, 1974, 58, 569-575.

96 minority junior high, and 506 minority senior high science students were administered tests of Piagetian operations. No correlation between age and task performance.

Pella, M., & Ziegler, R. Use of mechanical models in teaching theoretical concepts. Journal of Research in Science Teaching, 1967-68, 5, 138-150.

72, grade 4,5 & 6 students were administered tests of science achievement after being instructed in concepts relating to particle nature of matter. No correlation between age and test performance.

A Synthesis of Social and Psychological Influences on Science Learning

Barbara K. Kremer and Herbert J. Walberg
University of Illinois at Chicago Circle

Running head: Social Psychology and Science Learning

Abstract

Research on the relationship of social and psychological factors--including student motivation and home and peer environments--to science learning in grades 6 through 12 was synthesized. Twenty-six studies conducted over a 16 year period from 1964-1979 were considered. A quantitative synthesis of findings indicate that motivation, home and peer environments are important correlates of science learning, and results in science are parallel to those observed in previous syntheses of these constructs in general, educational research.

A Synthesis of Social and Psychological

Influences on Science Learning

Beginning with Jones and Fiske (1953), a number of reviewers have urged the quantitative synthesis of educational and psychological research findings (Gage, 1978; Light & Smith, 1971; and Rosenthal, 1976). These reviewers describe a variety of statistical techniques for summarizing and evaluating a series of empirical findings across investigations. As for example, in the natural sciences where estimates of astronomical constants are made (Ash, Shapiro, & Smith, 1967), these techniques are intended to provide objective estimates across investigations of the consistency of observations or coefficients such as means, correlations, and regression weights; their magnitude and margins of error; and their boundaries of application.

The purpose of the present review is to synthesize, through the application of quantitative methods, social and psychological research on science learning in grades six through twelve, conducted under three rubrics -- student motivation, home or family environment,and peer-group environment. The present synthesis is part of a larger effort to synthesize science education research on factors that are productive of cognitive, affective, and behavioral learning. Those considered are student ability (including developmental level and prior achievement) and motivation; amount and quality of instruction; and home, classroom, and peer-group environments (Walberg, 1978). These factors have been frequently investigated in general educational research, and show reasonably consistent and, in most cases moderate to strong associations with learning outcomes.

It seems particularly appropriate to investigate learning productivity in science-education at this time for several reasons. The movement "back to basics," and tightened school budgets threaten to diminish the place of science in the school curriculum as represented by instructional time, quality of lesson preparation, and laboratory facilities. Furthermore, the growing field of science education research has yielded a large number of published reports that appear ready for parsimonious integration and summary. Syntheses of educational research in subjects such as reading and mathematics, focusing on a large number of constructs and subconstructs, have already been conducted (Walberg, Schiller, & Haertel, 1979; Uguroglu & Walberg, 1979; Iverson & Walberg, 1979). It is of interest to know if the results of synthesis carried out in science yield the same general conclusions, or whether a separate set of learning laws or "production functions" seem necessary in the special field of science. The identification of causal factors or constructs, and the importance of objectively reviewing evidence on them, is in substantial agreement with the broad review of science-education research needs carried out by the National Association for Research on Science Teaching and the National Institute of Education (Yager, 1978).

The constructs of motivation and home and peer-group environment are placed together in the present synthesis, and somewhat apart from the others, because these topics have by comparison, been neglected in science as well as in general educational research. It is therefore possible to bring together, and discuss all the selected work on these three constructs in a single paper. Second, these three constructs fit under the general rubric of social psychology rather than the mainstream fields of curriculum, instruction, or cognitive and behavioral psychology that currently seem more influential on educational policy and practive. Work on the social environment of the classroom is also social-psychological, but the sizeable number of large-scale studies necessitates a separate treatment (Haertel, Walberg, & Haertel, 1979). Lastly, motivation and home and peer-group environments, are only semi-manipulable and under the partial control of educators. They seem less fixed than mental ability but, on the other hand, more difficult to influence than teacher behavior or allocation of time in the curriculum. The science teacher can raise motivation, and perhaps also encourage science learning in the home and in adolescent peer-groups; but such changes require the cooperation of other agents such as the students themselves and their families. For these reasons, the three major constructs are synthesized and compared in the present review.

## Literature Search and Selection

Fifteen years of science education literature (1964-1979) were searched to identify studies relating science achievement and learning to each of the three constructs areas under consideration: Student motivation, home or family environment, and peer environment. This time period was selected in order to reflect recent growth in curriculum development and evaluation and to include the most current science education research. In searching the literature, priority was given to selecting studies from refereed journals. Search procedures were extended to unpublished reports and dissertations when the number of studies located in the published journals did not appear to be sufficient.

For the period 1964 to 1979, studies in the two major research journals in science education, the Journal of Research in Science Teaching and Science Education, were scanned. Volumes of School Science and Mathematics, Journal of Educational Psychology, Developmental Psychology, and Sociology of Education for the years 1971-1977 were also searched. Computer searches of studies indexed by the Educational Resources Information Center (ERIC), and the Social Sciences Citation Index (SSCI) were conducted. The collection of science education bibliographies and annual reviews published by the Science, Mathematics, and Environmental Education Information Analysis Center (ERIC/SMEAC) were scanned for citations of dissertations and unpublished reports.

Studies were screened and selected for synthesis on the basis of the following criteria: 1) Concerned with science learning in grades 6-12; 2) That some measure of student learning in science (e.g., achievement, attitude, developmental level attained) be reported; 3) That at least one of the three constructs under consideration serve as a predictor of science learning. Table 1 presents definitons of the motivation, home and peer constructs which guided this search, and examples of how these constructs were conceptualized and operationalized in the literature.

Insert Table 1 about here

The results of this search and selection yielded a total of 20 studies: 5 studies considering student motivation, 13 of home environment, and 5 of peer environment, (two of the studies selected considered 2 or more of these construct variables). While numerous studies of student motivation, home environment, and peer environment constructs were found, many were excluded from the analysis for several reasons: measures of science achievement were either absent or invalid (22 studies); findings relating the effect of the construct variable to achievement were inadequately reported (15 studies); reports were based on opinions rather than evidence ( 5 studies); or studies considered the effects of science learning on some measure of the construct variable such as students' self-concept in science, or compatibility with peers (15 studies). Up to 20 studies of home environment alone, were excluded for these reasons. A complete bibliography of studies selected for inclusion under each construct is contained in the appendix.

## Method of Analysis

All of the studies selected for synthesis were numerically coded using schemes developed by investigators for each construct area. From 40-50 study variables were coded in each construct area.

These included the type, source, and validity of science learning and construct measures; the characteristics of the sample; the type of design employed; and methodological flaws threatening the validity of the study. Statistical information, including correlations and inferential statistics, levels of significance, and the sign or direciton of results were also recorded for each study. Copies of the coding schemes used are available from the authors.

The limited number of adequate studies available under these constructs, precluded the use of multivariate techniques of research synthesis (Glass, 1978).

Instead, findings were synthesized by plotting the correlations, calculating simple statistcs, and tabulating "box scores" denoting the direction (whether positive or negative) of the relationship between construct variables and learning outcomes.

## Results and Discussion

The majority of studies selected for this synthesis, a total of 16 of 20, were correlational. Where correlations were reported in studies, these were recorded for analysis. In studies not reporting correlations between construct variables and learning outcomes, when possible, techniques outlined by Glass (1978) for converting statistics to correlations were applied. In studies with insufficient information to derive correlations from statistics reported, signs or box scores were coded denoting the direction of the relationship between construct variables and science learning outcomes. Studies indicating that as the construct variable increased, science learning or achievement increased, were coded as positive ("+"). Studies showing an inverse relationship or no relationship, between construct and achievement variables were coded as negative ("-").

Subject characteristics, study features and findings, median correlations, and box scores of studies under each of the three construct areas, are summarized and discussed below (see Tables 2, 3 and 4). Unless specified in the table, subjects were from white, middle-class, mixed sex populations in the United States. While the sample of studies represented is limited, the results indicate consistent, positive findings, in studies considering student motivation, home environment, and peer environment as predictors of science learning. Of the total 20 studies considered, 14 indicated positive signs of the findings. The binomial probability of this ratio is $< .01$.

Table 5 presents stem and leaf diagrams (Tukey, 1976) of all correlations in all studies as well as the median correlations for each study. The first decimal place of the correlation is represented on the stem on the left of the vertical line; and the second decimal place is represented as a leaf to the right of the line; for example, the highest and lowest outlying correlations for the student motivation construct are .15 and .58. These diagrams show all the correlations in the studies as well as the study-median correlations that weight each study equally. Mean correlations were computed for each construct area using the raw correlations reported in individual studies. The mean correlations for the three construct areas are .37 for student motivation, .30 for home environment, and .24

for peer environment. Results specific to each construct area are discussed below.

---

Insert Table 5 about here

---

Student Motivation

All of the studies of student motivation and science achievement located, showed positive relationships between motivational variables and learning. These are summarized in Table 2. Three studies considered measures of academic self-concept (Alvord & Glass, 1974 ; Raven & Adrian, 1978; Mancini, 1972), one study (Bart, 1978) looked at reported persistance, and another (Soh, 1973) considered general, need-achievement motivation. Of these studies relating student self-concept to science learning, only one study (Raven & Adrian, 1973), specifically looked at students' concept of their ability in science, as opposed to general academic self-concept.

---

Insert Table 2 about here

---

The mean correlation for student motivation and science learning, .37, is somewhat higher than those obtained in the home and peer environment constructs as shown in Table 5. In part, this may be explained by the fact that standardized scales having the advantage of higher measurement reliability, were used to measure motivation sub-constructs (e.g., self-concept). This of course, was not the case in studies of home and peer environment, as will be discussed below. As noted in a synthesis of student ability and science learning (Boulanger, 1979): construct measures with higher reliability yield higher correlations with learning outcomes (particularly cognitive ones) than measures with lower reliabilities.

Individual correlations reported for the student motivation construct area in Table 5, indicate a median correlation of .33. Previous studies of student motivation and general educational achievement conducted by Bloom (1976) and Uguroglu and Walberg (1979) report median correlations of .35, and .30 respectively. These studies were based on large national samples, and included correlations with achievement data from reading and mathematics. The similarity of the correlations found in this study with those reported by Bloom, and Uguroglu and Walberg suggests that the "productive function" of student motivation in learning and achievement is independent of subject area or content. This possibility warrants further study. Motivational factors in science learning, in general, merit greater attention than they have received from science educators as evidenced by these findings.

Home Environment

All of the studies selected in this construct area, as summarized in Table 3, contained measures of parents' socio-economic status or SES, and science learning. Among the SES indices considered, were parent occupation, parent education, and community SES. Of the 13 studies considered, 9 show positive relationships between parental SES and science learning: Students of higher socio-economic status homes scored higher on achievement measures of logical operations (Bart, 1978), science attitudes and interests (Neujahr & Hansen 1970; Hasan, 1975; James & Pafford, 1973; Keeves, 1975), general cognitive learning in science (Hardy, 1970; Keeves, 1975; Klein, 1971; Troost, 1969), critical thinking (Hardy, 1970), and factual learning (Lynch et al., 1979). Studies showing no significant relationship between SES and science achievement are those considering process learning (Quinn & George, 1975), factual learning (Ashbaugh, 1968), and science attitudes and interests (Wynn & Bledsoe, 1967).

Insert Table 3 about here

The mean value of correlations reported between SES and science learning was computed as .25. White (1976) obtained a mean correlation of .26 between parental social class indices and measures of verbal and mathematics achievement. Again, as in the case of student motivation, the correlation obtained in science is similar to that obtained in earlier work on general educational achievement, based on a larger sample of studies.

In addition to SES, several studies considered other indices of home environment. Among these sub-constructs were parent education (Hasan, 1975), parental aspirations for student achievement (Bart, 1978; Hasan, 1975; Keeves, 1975), parent involvement in the student's education (Bart, 1978), and the presence of science equipment in the home (Neujahr & Hansen, 1970). The mean of correlations reported for these indices was computed as .36. Higher correlations with learning were therefore obtained for these indices than for more general SES measures. Again, this correlation is similar to that reported elsewhere for verbal and mathematics achievement. Iverson and Walberg (1978) obtained a mean correlation of .35 for studies considering parent stimulation of the child with measures of verbal ability and general educational achievement.

A particularly noteworthy study in this construct area is that of Keeves (1975) who considered multiple predictors of achievement, and both learning and interest in science. The effects of father's occupation, parental aspirations for the child, parent involvement in the school, and general SES level on science attitudes and interests, and general cognitive achievement in science were investigated. His study was based on a randomly selected sample of 215 Australian sixth and seventh grade students. Science learning and interest were measured by specially prepared attitude questionnaires and achievement tests in science.

In other studies, the most frequently used methods for collecting home environment information were student questionnaries (Neujahr & Hansen, 1970; Hasan, 1975; Stronck, 1974) and the use of school archives (Hardy, 1970; James & Pafford, 1973; Wynn & Bledsoe, 1967). Three studies failed to report methods used for securing home data (Bart, 1978; Ashhaugh, 1968; and Troost, 1969). The reliabilities of measures used in these studies is seldom reported.

Peer Environment

Of the five studies considering the effects of peer environment and science learning in Table 4, three were concerned with the effects of within class grouping on cognitive science learning: i.e., with the effects of individual vs. group work (Gabel & Herron, 1977; Linn et al., 1977), and homogeneous vs. heterogeneous ability groupings (Bicak, 1964).

Of these three studies, only Gable and Herron show a positive relationship between learning and peer environment. They report that, in their urban sample (the study also included a rural sample), group work had a positive effect on factual learning in general science. This result was not replicated in their rural sample. Bicak found no significant effect for ability grouping on the learning of science material in meteorology; while Linn et al found no effects for individual vs. group work on the acquisition of logical operations.

Insert Table 4 about here

In his study of logical operations in urban adolescents, Bart (1978) reported a correlation of .25 for teacher ratings of students' "rapport with peers". Keeves (1975) considered the effects of friends', or peers' participation in science and mathematics activities on students' cognitive achievement and critical thinking in general science. He reported correlations of .23 and .24, respectively for these measures.

That the number of studies considering the effects of the peer environment on science learning over the past fifteen years is so limited, is noteworthy. This is particularly so, in light of attention previously given to peer influences on achievement in general educational literature (Coleman, 1961). Of note too, is the observation that none of the studies reviewed here considered sociological or extra-curricular aspects of the peer environment on science achievement. They were rather, restricted to the consideration of peer influeces within the classroom. That peers exert considerable influence outside the school on curricular choices and academic achievement, has been demostrated in previous research on adolescence (Bradley, 1977; Spencer, 1976; Kandel & Lesser, 1969).

Conclusions

As the results of the literature search and selection undertaken in these construct areas demonstrate, science educators have paid little attention to student motivation, home environment, and peer environment variables in the study of science achievement. Nevertheless the consistent, positive direction of findings observed in studies of these constructs makes a strong case for their inclusion in future research. Student motivation, and home and peer environment factors appear to be important correlates of science learning. They deserve closer attention from the science educator since academic achievement associated with these constructs is subject to environmental intervention, either through instruction, or counseling.

The consistency and parallelism of results observed in studies of student motivation and home environment with previous work in general education suggests the need for further direct investigation of these constructs. The incorporation of such constructs as control or stratification factors in curriculum and instructional research is recommended; and the value of attempts to manipulate these constructs experimentally in the hope of making science-education more productive is indicated.

References

Ash, M. E., Shapiro, I. I., and Smith, W. B. Astronomical constants and planetary aphemerides deduced from radar and optical observations. Astronomical Journal, 1967, 72, 338-350.

Bloom, B. S. Human characteristics and school learning. New York: Mc Graw-Hill, 1976.

Boulanger, F. D. Ability and science learning: A quantitative synthsis. Manuscript submitted for publication, 1979

Bradley, J. Clique membership and academic achievement. Research in Education, 1977, 18, 1-8.

Coleman, J. S. The adolescent society. New York: Cromwell-Collier Publishing Company, 1961.

Gage, N. L. The scientific basis of the art of teaching. New York: Teachers College Press, 1978.

Glass, G. V. Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education, Vol. 5. Itasca: Peacock, 1978.

Haertel, G. D., Walberg, H. J., & Haertel, E. H. Social psychological environments and learning: A quantitative synthesis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Iverson, B. K., & Walberg, H. G. Home environment and learning a quantitative synthesis. Paper presented at the annual meeting of the American Educational Research Asssociation, San Francisco, April 1979.

Jones, L. V., & Fiske D. Models for testing the significance of combined results. Psychological Bulletin, 1953, 50, 375-382.

Kandel, D. B., and Lesser, G. S. Parental and peer influences on educational plans of adolescents. American Sociological Review, 1969, 34, 212-223.

Light, R. J., and Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 421-471.

Rosenthal, R. Experimenter effects in behavioral research. New York: Irvington Publishers, 1976.

Spencer, W. A. Interpersonal influences on educational aspiration: A crosscultural analysis. Sociology of Education, 1976, 49, 4- 46

Tukey, J. W. Exploratory data analysis. Reading, Mass.: Addison-Weseley, 1976.

Uguroglu, M. E., and Walberg, H. J. Motivation and achievement: A quantitative synthesis. American Educational Research Journal, 1979, in press.

Walberg, H. J. A psychological theory of educational productivity: In F. H. Farley and N. Gordon (Eds.) Contemporary Perspectives on educational psychology. Chicago: National Society for the Study of Education, 1980, in press.

Walberg, H. G., Schiller D., and Haertel, G. D. The quiet revolution. Phi Delta Kappa, 1979, in press.

White, K. R. The relationship between socio-economic status and academic achievement. Unpublished doctoral dissertation, University of Colorado at Boulder, 1976.

Yager, R. E. Priorities for research in science education: A study committee report. Journal of Research in Science Teaching, 1978, 15, 99-107.

Footnote

[1]This research was supported by a National Science Foundation Grant, No. 78-17374, to Herbert J. Walberg, Principal Investigator. Findings and opinions expressed in this paper are solely the responsibility of the authors, who wish to acknowledge the critical contributions of James A. Kulik, University of Michigan and Wayne W. Welch, University of Minnesota and the comments of colleague F. David Boulanger who is synthesizing research on other factors influencing science learning.

Table 1

Definition and Examples of Student Motivation, Home Environment, and Peer Environment Constructs

| Construct | Definition | Example Measures |
|---|---|---|
| Student Motivation | Any measured intrinsic drive or extrinsic reward that influences student performance during an instructional treatment or test situation. | Self-concept, persistence, need-achievement, test anxiety. |
| Home Environment | Any characteristic of environments over which a parent or guardian exerts direct control as opposed to classroom or peer group environment. | Parent occupation (SES), presence of science-related equipment and documents in the home, parent involvement in school work. |
| Peer Environment | Characteristics of the students' beliefs, practices, and social activities associated with peer group beliefs and practices. | Ability tracking (between classes), school activities (extra-curricular), instructional grouping (within classes). |

Table 2

Student Motivation and Science Learning Studies: Subjects, Features, and Findings

| Author (Date) | Subjects | Feature | Finding | Sign/ $r_{xy}$ |
|---|---|---|---|---|
| Alvord & Glass (1974) | 3162 grade 4,7,12 students | Academic achievement in science as measured by NAEP tests, and self-concept | Positive correlation between achievement and self-concept | + .16[a] |
| Bart (1978) | 285 urban high school students, aged 13-19; heterogeneous racial, ethnic and SES backgrounds | Adolescent formal reasoning and teacher's evaluation of task persistence (peer and home environment also considered) | Positive correlation between formal reasoning and persistence | + .26[a] |
| Mancini (1972) | 267 suburban grade grade 7 students | Self-concept of academic ability, and achievement in biology | Students with higher self-concept, evidenced higher achievement | + |
| Raven & Adrian (1978) | 249 grade 9-11 rural, average and above average students | General science achievement, and general self-concept of ability and concept of ability in science | Positive correlation between achievement and general and science self-concepts | + .47[a] |
| Soh (1973) | 170 high ability second year male students from English Grammar Schools | Comparison of the motivational orientations of students with, and without career interests in science | Students with greater preference for science careers, evidenced higher achievement motivation | + |

+ - Positive relationship between construct variable and science learning

- = Negative relationship between construct variable and science learning

lian of reported correlations

Table 3
Home Environment and Science Learning Studies: Subjects, Features, and Findings

| Author (Date) | Subjects | Feature | Finding | Sign/ $r_{xy}$ | |
|---|---|---|---|---|---|
| Ashbaugh (1968) | 430 grade 4-6 students from upper middle class suburban communities | Attainment of geological concepts and SES | No differences in learning as a function of SES level | - | |
| Bart (1978) | 285 urban high school students, aged 13-19; heterogeneous racial, ethnic, and SES backgrounds | Adolescent formal reasoning and parent involvement in the school, parent aspirations for the child, and SES (student motivation and peer environment also considered | Positive correlation between formal reasoning achievement and home environment | + | .30[a] |
| Hardy (1970) | 208 chemistry students 104 enrolled in CHEM study, 104 in traditional chemistry courses | Critical thinking and performance on standardized achievement test correlated with SES | Positive correlation between SES and critical thinking and achievement | + | .26[a] |
| Hasan (1975) | 340 grade 11 Jordanian students | Student interest in science and parents' education (SES), and parent aspirations | No differences in interest as a function of parents' education, but positive relationship found between science careers desired by parents and student science interest | -/+ | |
| James & Pafford (1973) | 84 grade 12 students | Student interest in science and father's occupation (SES) | Students with professional fathers elected more science courses than those of non-professional fathers | + | |





+ = Positive relationship between construct variable and science learning

- = Negative relationship between construct variable and science learning

a = Median of reported correlations

Table 3 (continued)

| Author (Date) | Subjects | Feature | Findings | Sign/ | $r_{xy}$ |
|---|---|---|---|---|---|
| Keeves (1975) | 215 Australian grade 6-7 students | Fathers occupation, parent aspirations, parent involvement in the school, and general SES level; and general science achievement and science attitude (also considered student motivation and peer environment) | Positive correlation between home environment indices and achievement and attitudes | + | .35[a] |
| Klein (1971) | 310 grade 6 students | General science learning and SES | Positive correlation between achievement and SES | + | |
| Lynch et al. (1979) | 1635 grade 7-10 Australian students | Performance on a test of factual science learning, and SES | Postive correlation between test performance and SES | + | .14 |
| Neujahr & Hansen (1970) | 194 students from a high school science honors program | Students interest in science (as evidenced by subsequent academic work in science); and fathers' occupation (SES), and presence of science equipment in the home | Positive correlations between interest and home environment indices | + | .17[a] |
| Quinn & George (1975) | 176 grade 6 students from urban, and suburban schools | Performance on a hypothesis formation task, and SES | No differences in performance observed as a function of SES | - | |
| Stronck (1974) | 700 grade 10 - 12 students from Texas | Performance on a statewide scholarship test of general science learning, and SES | Positive correlation between test performance and SES | + | |

Table 3 (Cont'd)

| Author (Date) | Subjects | Feature | Findings | Sign | / $r_{xy}$ |
|---|---|---|---|---|---|
| Troost (1969) | 54 grade 7-9 students of diverse ethnic origin | Achievement following a summer program in space science, and SES | Positive correlation between achievement and SES | - | .21 |
| Wynn & Bledsoe (1967) | 325 urban, grade 11-12 students | Students' interest in science and SES | No difference in interest found as a function of SES | - | |

Table 4

Peer Environment and Science Learning Studies: Subjects, Features, and Findings

| Author (Date) | Subjects | Feature | Findings | Sign/ | $r_{xy}$ |
|---|---|---|---|---|---|
| Bart (1978) | 285 urban high school students, aged 13-19; hetorogeneous racial, ethnic and SES backgrounds | Adolescent formal reasoning and teacher's evaluation of rapport with peers (student motivation and home environment also considered) | Positive correlation between achievement and rapport with peers | + | .25 |
| Bicak (1964) | 77 grade 8 students | Homogeneous vs. Heterogeneous ability grouping on achievement in a local science course | No differences between homogeneous and heterogeneous ability groups in achievement | - | |
| Gabel & Herron (1977) | 1022 grade 7 ISCS students from county and city schools | Group work vs. individual work on retention | Higher retention shown for city students working with partner. No differences found in the county sample | +/- | |
| Keeves (1975) | 215 Australian grade 6-7 students | Peer participation in science and math, and general science achievement and science attitude (also considered student motivation and home environment) | Positive correlation between peer environment, and achievement and attitudes | + | .24[a] |
| Linn et al. (1977) | 132 grade 5-6 students in a lower middle class urban school | Individual work vs. elective group work on promoting students' ability to control variables | No differences in achievement for individual and elective group work | - | |

+ = Positive relationship between construct variable and science learning

- = Negative relationship between construct variable and science learning

a = Median of reported correlations

Table 5

Stem and Leaf Diagrams of Individual and Study-Median Correlations For

Student Motivation, Home Environment, And Peer Environment Construct Variables

| Student Motivation | | | Home Environment | | | Peer Environment | | |
|---|---|---|---|---|---|---|---|---|
| | INDIVIDUAL | MEDIAN | | INDIVIDUAL | MEDIAN | | INDIVIDUAL | MEDIAN |
| .6 | | | .6 | | | .6 | | |
| .6 | | | .6 | | | .6 | | |
| .5 | 68 | | .5 | 5 | | .5 | | |
| .5 | | | .5 | 03 | | .5 | | |
| .4 | 7 | | .4 | 9 | | .4 | | |
| .4 | | 7 | .4 | 0 | | .4 | | |
| .3 | 6 | | .3 | 5 | 5 | .3 | | |
| .3 | 13 | | .3 | 02 | 0 | .3 | | |
| .2 | | 6 | .2 | 5668 | 6 | .2 | 5 | |
| .2 | | | .2 | 01 | | .2 | 34 | 4 |
| .1 | 57 | 6 | .1 | 677 | 7 | .1 | | |
| .1 | | | .1 | 4 | | .1 | | |
| .0 | | | .0 | | | .0 | | |
| .0 | | | .0 | | | .0 | | |

| | Student Motivation | Home Environment | Peer Environment |
|---|---|---|---|
| Mean = | .36 | .30 | .24 |
| Median = | .33 | .26 | .24 |
| Sd = | .15 | .13 | .01 |

# Appendix

## Bibliography of Studies on Student Motivation Home and Peer Environment and Science Learning

Studies of Student Motivation

Alvord, D. J., & Glass, L. W. Relationships between academic achievement and self-concept. Science Education, 1974, 58, 175-179.

Bart, W. M. Formal operations among high school adolescents from lower and middle socio-economic classes: An investigation of the scholastic import, psychometric structures, and styles of formal reasoning. Unpublished final report to the Spencer Foundation, Chicago, November, 1978.

Mancini, D. An investigation of the relationships between self-concept of ability, classroom verbal interaction, and achievement of seventh grade pupils in biological science in two suburban schools. Dissertation Abstracts International, 1972, 33, 5577A. (University Microfilms No. 7308183).

Raven, R. J., & Adrian, M. Relationships among science achievement, self-concept, and Piaget's operative comprehension. Science Education, 1978, 62, 471-479.

Soh, K. C. Dynamic structures and science bias. Science Education, 1973, 57, 335-341.

Studies of Home Environment

Ashbaugh, A. E. Selection of geological concepts for intermediate grades. Science Education, 1968, 52, 189-196.

Bart, W. M. Formal operations among high school adolescents from lower and middle socio-economic classes: An investigation of the scholastic import, psychometric structures, and styles of formal reasoning. Unpublished final report to the Spencer Foundation, Chicago, November, 1978.

Hardy, C. A. CHEM study and traditional chemistry: an experimental analysis. Science Education, 1970, 54, 273-276.

Hasan, O. E. An investigation into factors affecting science interest of secondary school students. Journal of Research in Science Teaching, 1975, 12, 255-261.

James, A. N., & Pafford, W. N. The relationship between academic achievement in science and father's occupation. Science Education, 1973, 57, 37-41.

Keeves, J. P. The home, the school, and achievement in mathematics and science. Science Education, 1975, 59, 439-460.

Klein, C. A. Differences in science concepts held by children from three socio-economic levels. School Science and Mathematics, 1971, 71, 550-558.

Lynch, P., Benjamin, P., Chapman, T., Holmes, R., Mc Cammon, R., Smith, A., & Symmons, R. Scientific language and the high school pupil. Journal of Research in Science Teaching, 1979, 16, 351-357.

Neujahr, J., & Hansen, R. The prediction of careers in science from high school data. Journal of Research in Science Teaching, 1970, 7, 391-394.

Quinn, M. E., & George, K. D. Teaching hypothesis formation. Science Education, 1975, 59, 289-296.

Stronck, D. R. The sociological backgrounds of scientifically talented secondary school students throughout the state of Texas. Journal of Research in Science Teaching, 1974, 11, 31-37.

Troost, C. J. A study of socio-economically divergent superior students in an aerospace enrichment program. Science Education, 1969, 53, 325-328.

Wynn, D. C., & Bledsoe, J. C. Factors related to the gain and loss of scientific interest during high school. Science Education, 1967, 51, 67-74.

Studies of Peer Environment

Bart, W. M. Formal operations among high school adolescents from lower and middle socio-economic classes: An investigation of the scholastic import, psychometric structures, and styles of formal reasoning. Unpublished final report to the Spencer Foundation, Chicago, November, 1978.

Bicak, L. J. Achievement in eighth grade science by heterogeneous and homogeneous classes. Science Education, 1964, 48, 13-22.

Gabel, D., & Herron, J. D. The effects of grouping and pacing on learning rate, attitude, and retention in ISCS classrooms. Journal of Research in Science Teaching, 1977, 14, 385-399.

Keeves, J. P. The home, the school, and achievement in mathematics and science. Science Education, 1975, 59, 439-460.

Lynn, M. C., Chen, B., & Thier, H. D. Teaching children to control variables: Investigation of a free-choice environment. Journal of Research in Science Teaching, 1977, 14, 249-255.

# Instruction and Science Learning:

# A Quantitative Synthesis

F. David Boulanger

University of Illinois at Chicago Circle

Running head: Instruction and Science Learning

Abstract

The purpose of this study was to quantitatively synthesize quality and quantity of instruction studies with the same or similar independent variables in the published science education grade 6-12 research of the 1963-1978 period. Fifty-two studies formed six clusters and revealed significant positive cognitive outcomes due to the use of preinstructional strategies, training in scientific thinking, increased structure in the verbal content of materials, and increased realism or concreteness in adjunct materials. In general, systematic innovation in instruction was found to produce positive improvements over the norm or traditional practice. Methodologically, improved research design quality was related to larger effect sizes. Recommendations are made regarding replication, use of multiple measures, attitudinal research, use of general education findings, and the reporting of research.

# Instruction and Science Learning: A Quantitative Synthesis

Research on the quality of instruction is extensive, diverse, complicated and often inconclusive. Reviews of hundreds of studies have resulted in disappointment expressed by many reviewers in what they interpret as a lack of substantive research in the quality of instruction and its influence on student learning (Travers, 1973). Yet other reviewers, using quantitative synthesis techniquies, have found positive empirical support for the influence of several factors on learning. Bloom (1976) identified instructional cues, participation, and reinforcement as accounting for up to 25% of the variance in student learning. Rosenshine (1979) summarized the work of several major researchers and found evidence for instructional time, content coverage, and direct instruction strategies as major influences on learning. Walberg, Schiller and Haertel (1979) tabulated the results of recent reviews on the relation of instructional and other educational conditions to learning outcomes and found a number of consistent, positive results.

One reason for the differing views on the summative findings in a given area of research is the qualitative character of attempts at research synthesis. Long narratives citing study after study provide little basis for objective comparisons and accumulation of results. If study characteristies and outcomes could be quantified, research synthesis might gain new precision and objectivity, providing a finer measure of what is known as well as a better knowledge of the gaps and flaws in the accumulated research.

Based on theoretical considerations and the accumulating empirical evidence, Walberg (in press ) developed a productivity model incorporating eight constructs as major factors in student learning. The constructs are: student ability, motivation, and age or developmental level; quality and quantity of instruction; and classroom, home and peer environments. Using quantitative research synthesis techniques estimates of the size of the contributions of each construct to general learning outcomes were prepared (Haertel, Walberg and Haertel, Note 1; Iverson and Walberg, Note 2; Uguroglu and Walberg, in press). The productivity model provides a framework of constructs known to be important factors in general learning and, therefore likely to be important in science learning. Yager (1978) identified the need for reviews of science education research and guidance from the findings of general education research as national priorities for science education. The present study was conducted to meet these needs by quantitatively synthesizing the science education research on learning for two of the constructs, the quality and quantity of instruction.

## Purpose and Method

The purpose of the present study was to quantitatively synthesize the published science education quality and quantity of instruction research performed with subjects in grades 6 through 12 over the 1963-1978 period. This period and grade range were chosen to include the recent growth in research and curriculum development with the precollege students enrolled in the range of general to specialized science courses. A quantitative approach to the synthesis was chosen to provide comparable indicies of the characteristics and outcomes within and among homogeneous groups or clusters of studies. The quantitative techniques of research synthesis advocated by Glass (1978) are employed. Quantitative syntheses is intended to complement traditional qualitative syntheses such as the annual Summary of Research in Science Education, e.g., Petersen and Carlson (1979). Quantitative techniques require multiple studies relating the same or similar variables in terms of compariable statistics such as signs, effect sizes, and correlations.

## Literature Search and Selection

One of the most difficult tasks in research synthesis is deciding what constitutes similar studies suitable for integration. Quality of instruction is a multi-dimensional construct encompassing many definitions and points of view. Rather than defining the construct a prior, it was decided to let the body of science education research define it through a simple count of independent variables receiving the most attention in experimental research on science instruction. The primary source of literature references was the collection of ERIC science education bibliographies and annual reviews. This combined with a scanning of all studies in the two major research journals in science education, Journal of Research in Science Teaching and Science Education, resulted in the identification of 137 published studies in the quality construct and 3 on the quantity of instruction, 2 published and 1 dissertation. (The quantity of instruction studies will be discussed later.) Ninety-five of the quality of instruction studies involved an instructional situation manipulated in an experimental fashion and learning outcomes measured. The additional 42 studies were curriculum comparisons and, due to the poorly defined nature of the treatments, were eliminated from further consideration. The 95 studies were categorized by independent variables and the categories and frequencies tabulated (see Table 1).

---

Insert Table 1 about here

---

A minimum of five studies was set as the criterion for inclusion of an independent variable or clustering of closely related independent variables in the synthesis, since the binomial probability of five independent studies having the same outcome direction is less than .05. This criterion would allow a strong test of the effectiveness of one treatment over another. For example, if the treatment group receiving indirect instruction achieved a higher mean score than the direct instruction group in five out five independent experiments, this would be accepted as strong evidence for the general superiority of indirect instruction.

Applying the above criterion, six clusters totaling 52 studies were identified: preinstructional strategies, indirectness of instruction, inductive vs deductive strategies, training in scientific thinking, structure in the verbal content of materials,and realism or concreteness in adjunct materials. Table 2 gives cluster component variables, operational definitions, and number of studies.

Insert Table 2 about here

## Coding

A numerical coding scheme for study variables (characteristics) was developed prior to study selection and refined as study coding progressed. Each comparison of treatment means in each study was coded according to approximately 40 study variables: dependent measure type, origin, and reliability; subject grade, sex, ethnic group, and academic achievement level; community SES and urban-rural character; subject matter of treatments and sources of curriculum; constructs measured other than quality of instruction; treatment characteristics including group size, elective or required course participation, regular or special teacher, lab or non-lab focus, reliability of implementation, length, and equality of control group access to content; study design and nine categories of threats to validity; sample size; and outcome statistics, ie. direction of effect, level of significance and effect size.

Effect size is a normalized measure of the difference between two treatment groups in performance on a dependent measure. Nearly all effect sizes were computed using one of the following two formulas (Glass, 1978):

$$ES = \frac{\bar{X}e - \bar{X}c}{Sc} \qquad ES = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\bar{X}e$ and $\bar{X}c$ represent experimental and control group means respectively. Sc is the standard deviation of the control group. t is the computed t-test statistic. If an F-test were used in a one-way analysis of variance to compare two groups, the F value was considered equal to $t^2$. If only the total sample size was given, it was assumed that $n_1=n_2$, since equal n's provide a more conservative estimate of effect size than the unequal n's. Finally, in cases where one-way analysis of variance was used, homogeneity of variances was assumed setting $Sc = \sqrt{MSw}$. Two-way analysis of variance tables, however, without other statistics, were insufficient for computing effect sizes.

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

1. Factual learning (recall, recognition of treatment content; retention test)
2. Conceptual learning (conceptattainment, science processes or logical operations, critical thinking; standardized achievement test.)
3. Attitudinal learning (any affective measure of opinion, attitude or interest.)
4. Laboratory performance test.

Methodological flaws (Cook and Campbell, 1976) were examined and coded as either 1) "potential threat" or 2) "adequately minimized." Flaws examined were: reliability of treatment; statistical power; error rate; maturation; history; selection bias; contamination, compensation or differential incentives; mortality, and generalizability. A simple sum of these ratings yielded an over all index of design quality.

Given the wide range in the number (1 to 11) of comparisons in different studies, and given the limited number of studies in any one cluster, it was decided to use the median effect size from each study in each outcome category. The median effect size has the advantages of greater stability than the mean and meets the criticism of lack of independence when multiple effect sizes are drawn from the same study. The 52 quality of instruction studies yielded 160 raw comparisons which reduced to 69 median comparisons. (A few studies were useful without computable effect sizes; therefore, there are a few more comparisons than effect sizes). Based on a small sampling of studies read independently by two raters, 90 percent agreement between raters was readily attained in coding the 40 study variables. The appendix contains a bibliography of all studies by cluster. Abstracts of each study, a code book, code sheet and a table of coded values are available in the project final report (Walberg, Boulanger, Kremer and, Haertel, Note 3).

## Analysis and Discussion

With the completion of coding it was apparent that many study variables were not available in the study reports (ie. subject ethnic group and community SES and urban-rural character) or were constant across studies (ie. mixed sex of sample and local origin of the treatment) and would, therefore, provide little help in identifying sources of variation across studies. Only study variables adequately reported and with non-constant variable values were considered in the analysis.

Across all studies, the distribution of median effect sizes in dependent variable categories was: 38 conceptual, 14 factual, 4 attitudinal, and 5 laboratory performance outcomes. Since the trends in size and direction of the factual outcomes conformed closely with the conceptual outcomes in any given cluster, and given the great overlap in content of factual and conceptual measures, the two outcome categories were combined into one category named cognitive outcome. The number of positive comparisons and the mean of median effect sizes for cognitive outcomes in each cluster and associated 95 percent confidence interval for each mean are summarized in Table 3 and discussed below in terms of trends in other coded study variables. Later, the entire set of quality of instruction studies will be analyzed and discussed. Reference to "significance" in the following sections refers to statistical significance at the .05 level.

---

Insert Table 3 about here

---

1) Preinstructional strategies. Three subgroupings of studies form the preinstructional strategies cluster: four studies on advance organizers, five on behavioral objectives and two on set induction. Each included study compared the effect of the strategy with a comparable instructional treatment where no preinstructional strategy or a placebo strategy was used.

Eight studies on a total of 1204 subjects resulted in a mean cognitive effect size of 1.03, significantly positive and favorable to the use of a preinstructional strategy. Seven of nine effect sizes (one study contributed two) were associated with significant differences, all favorable to the strategies. The strongest contributors to the large effect were studies on the use of behavioral objectives and set induction, with 5 of these studies having significant findings. Inspection of the two weakest effect sizes in this cluster of studies revealed that both originated from the same source (Santiesteban, 1977) a study with the shortest treatment length, less than 1 hour, of any study in the cluster. By contrast, the greatest effect size (Olsen, 1973) resulted from a course length treatment on the largest sample in the cluster and with the highest design quality rating. Examination of other study variables indicated that the most effective strategies were conducted by trained regular teachers using prepared materials with their own students rather than materials used alone without teacher intervention.

2) Indirectness of instruction. Two subgroupings formed this cluster. One subgroup of seven studies, called "non-direct versus direct," compared teacher or workbook controlled instruction with instruction allowing, by comparison, greater student choice in content and/or method. The other subgroup (two studies), called "indirect versus direct," used Flanders Interaction Analysis to monitor the degree of teacher indirectness in lecture-discussion settings. The learning of students of high indirect versus low indirect teachers were compared _ex post facto_. For coding purposes these later two studies were classified as quasi-experiments.

Eight studies totaling 1135 subjects resulted in a mean cognitive effect size of .11, favorable though not significantly, to the non-direct approach. Five of the 10 effect sizes yielded significant differences, three favorable to the non-direct or indirect, and two to the direct approach. These results indicate no general tendency for one approach to be superior to the other. A trend was noted in the four positive effect sizes: all were from studies conducted in grade 10 or above.

The two reported attitudinal effect sizes almost exactly cancelled each other for a mean of .002. The study (Campbell, 1971) showing a significant effect size favorable to the indirect approach had the weakest design quality for this cluster, while the study (Kline, 1971) with the opposite outcome had the strongest design quality rating. Both studies were with required junior high courses.

3) Inductive vs. deductive strategies. This cluster of studies bears some resemblance to the indirectness of instruction cluster in that the deductive or expository strategies always involved a greater degree of teacher and / or printed materials verbal directness in the instructional process. This cluster differed from the indirectness cluster in that the sequencing of instructional components in the two competing treatments always had the flavor of one being the reverse of the other, e.g., from rule-to-example compared to from example-to-rule. There were no subgroupings of studies in this cluster.

Seven studies with cognitive outcomes gave a mean effect size of -.22 favorable to the deductive strategy. In terms of direction of effect, seven of nine comparisons favored the deductive strategy, but only one was significant. The two largest effect sizes (Babikian, 1971 and Thomas, 1969) both involved regular teachers using prepared meterials with their own student in 8th grade required science courses. The stronger of the two studies (Babikian: a true experiment with higher design quality rating) yielded the highest and only significant effect size. However, a study (Tanner, 1969), comparable to the highest effect size study in many respects (true experiment with regular 9th grade teachers in required course over similar treatment length), but using materials only (no teacher intervention) and with the highest design quality rating in this cluster resulted in no significant differences on conceptual outcomes. As in the case of indirectness cluster studies, the mean effect size was not significantly different from zero and no conlusion can be drawn about the superiority of one approach.

Comparing the inductive vs. deductive cluster with the indirectness cluster, there was evidence of a continuation of the pattern suggested earlier, namely, that one teaching strategy may be more effective with upper grade students, while the other stategy is more effective in lower grades.

Figure 1 is a scatter plot of all cognitive outcome effect sizes for both the indirectness of teaching role cluster and the inductive vs. deductive cluster against grade level. The correlation (r = .48) of effect size with grade level is significant. The trend is worthy of further research.

Only one attitude outcome was reported in this cluster. It favored the deductive approach but was non-significant.

4) Training in scientific thinking. Two subgroups formed this cluster, seven studies attempting to train subjects in some aspect of Piagetian related logical operations and two studies of the effects of training in the processes of science. The mean cognitive outcome effect size for the cluster, based on 716 subjects in eight studies, was .89 significantly positive and favorable to training students to use logical operations or processes of science. Eight of the 11 median effect sizes were based on significant differences, all favorable to the effectiveness of whatever training strategy was used in the study. However, only one of the eight significant differences was from a study where the control group had equal access to the content being taught.

Examining other study variables, the strong mean effect size is a clear statement that progress in scientific thinking can be made in a wide range of grade levels (grades 5-9), in relatively short treatment periods (2 to 10 hours), as part of required courses where a special teacher or special materials present carefully designed instruction to individual students. Only one study (Howe, 1977) of the significant studies had the regular teacher working with a class size group. That study was a quasi-experiment with a very low (10) design quality rating. No attitudinal or laboratory outcomes were reported.

5) Structure in the verbal content of materials. This is the most tightly defined cluster of the six discussed in this paper. All five studies in the cluster use Anderson's (1971) analysis and operational definition of structure. The operational definition takes the form of formulas used for computing certain structural coefficients based on a careful analysis of printed materials. In each study, the learning of subjects using high structure materials was compared to subjects using lower structure materials. The cognitive outcome mean effect size was .74, significantly positive and based on six effect sizes all favorable (three significantly) to the higher structure treatment.

The homogeneity of this cluster of studies is evident in a brief examination of study variables. All are short (one hour or less) treatments in biology or life science, administered to individuals in true experiments where the control group has, with one exception, equal access to the content. All treatments are without teacher intervention, based only on printed (or audio taped) materials in, with one exception, non-laboratory settings. All studies are of high design quality.

One laboratory outcome effect size, 1.364 was found. It was significant and in favor of the higher structure treatment.

6) Realism or concreteness of adjunct materials. Studies in this cluster have a common feature of comparing instructional treatments differing in their positions on the instructional materials concrete-symbolic continuum. Comparisons might involve manipulative vs. pictorial materials, laboratory based vs

lecture based instruction or more commonly, pictorial vs verbal presentations in printed matter. In coding each study, the experimental group was always the group receiving the more concrete or realistically illustrated instruction.

All nine studies with cognitive outcomes were favorable to the more concrete or realistic instructional mode, yielding a significant mean effect size of .58 based on a total of 512 subjects. Five of the nine outcomes were significant with four of these from true experiment studies of high design quality. Six of the nine studies used instructional materials only with no teacher intervention, seven were of short duration (less than ten hours) and seven involved individuals working alone with the materials. The major exception to these trends was a year long study (Yager, 1969) comparing a laboratory based approach to a comparable content, expository approach. This study resulted in one of the lowest, yet positive, effect sizes ( .131). Over all, the evidence was strongly supportive of the value of realism and concreteness in adjunct instructional materials to teach conceptual content.

Only one lab performance outcome (effect size 1.540) was reported. It was favorable to realism or concreteness, while the one attitudinal outcome (effect size - .848) was favorable to an expository over a laboratory approach. Both were from the Yager study.

## Conclusions: Quality of Instruction Clusters

Based on the published science education research on subjects in grades 6 through 12 for the period 1963 through 1978, conclusions about the impact on student learning of certain aspects of the quality of science instruction are stated here.

1) Preinstructional strategies, especially the use of behavioral objectives and set induction but also advance organizers, can improve student concep-

tual learning when used with other instructional activities by classroom teachers. The mean effect size 1.03 is significantly positive and is equivalent to an improvement of about one standard deviation (34 percentile points) when the treated group is compared to a control group having access to the content of instruction but without the focusing of a preinstructional strategy.

2) Non-direct or indirect instruction compared to direct instruction resulted in no difference in the general effectiveness of one approach over the other. This cluster of studies was characterized by design weaknesses and significant findings both for and against a given instructional strategy.

3) The mean cognitive outcome effect size of -.22 though slightly favorable to the deductive over the inductive teaching strategy, must be accepted with caution since it is not significantly negative and only one of the ten studies reported significant differences between the outcomes of the two strategies. As in the previous cluster, no firm general conclusion can be drawn regarding the effectiveness of one strategy over the other.

4) When the indirectness cluster findings are combined with the inductive versus deductive cluster findings, a pattern of effect sizes against grade level led to the conclusion that deductive or direct instruction tends to be more effective in terms of cognitive outcomes with junior high level (grades 6-8) students in required courses, while indirect, non-direct or inductive instruction was more effective with senior high (grades 10-12) students in elective courses.

5) Training in scientific thinking, especially the use of logical operations, is effective in terms of cognitive outcomes when conducted on an individual basis by a special teacher. Only two studies with significant effect sizes involved class size groups, one with the regular and one with a special teacher. The mean effect size for training .89 is significantly positive and equivalent to a 30 percentile point improvement when compared to untrained control subjects.

6) More highly structured verbal content in printed or audio materials is more effective in promoting cognitive learning than less structured content. The mean effect size .74 is significantly positive and equivalent to about 27 percentile points between the low structure group and high structure group means.

7) An insufficient number of studies were found reporting attitudinal or laboratory outcomes to draw any general conclusions about what aspects of quality of instruction have favorable or unfavorable effects.

## Cognitive Outcome General Trends

Examination of the 57 comparisons of cognitive outcomes including 52 median effect sizes provides some insight into the general effectiveness of systematic innovation in instruction. All studies were coded such that the experimental treatment represented a departure from the norm or "traditional" instructional practice. Twenty-three of the 57 comparisons (Table 3) were significantly positive while only three were significantly negative. The mean cognitive effect size was .55, significantly positive and favorable to experimental treatments. Removing those comparisons (14) where control group access to content was less than the experimental group slightly lowers the mean effect size to .51, still significant and equivalent to an improvement of approximately 20 percentile points over a control group.

The influence of study variables on effect size was investigated by computing and comparing the effect sizes corresponding to various subgroups of studies. Only studies with comparable content access by both treatments were included. Nine major study variables, subgroups of values, corresponding effect sizes, and F-test results are shown in Table 4. None of the differences among subgroups is significant. One trend deserves noting: published outcome measures tended to yield larger effect sizes.

Insert Table 4 about here

Correlations between effect size and the study variables of sample size, grade level, and reliability of outcome measure were computed and found to be -.02, -.10 and .00 respectively. All three were non-significant.

## Methodological Quality

An index to the general quality of the studies synthesized is the breakdown of design characteristics and threats to validity. Seventy-two percent of the studies were true experiments using random assignment of subjects to treatments; 28 percent were quasi-experiments, 5 percent of these using matching. Based on references in the study reports to precautions taken to insure the treatments were reliably implemented, 44 percent of the studies were judged to have low reliability of treatment implementation, 37 percent adequate reliability, and 19 percent high reliability. This weakness in design (or reporting) might be remedied using a verification of treatments approach such as that described by Leonard and Lowry, (1979). The percent of studies judged to be probably flawed by other threats to validity is given here with the associated threat: 65 percent, inadequate statistical power; 2 percent, error rate; 24 percent, maturation; 49 percent, history; 27 percent, selection bias; 59 percent, contamination, compensation or differential incentives; 19 percent, mortality; and 34 percent generalizability. Since no study involved random selection of subjects from a larger, well identified population, the last percent is a rather conservative estimate.

The rather high rate of threats to study validity might call into question the .51 over all cognitive outcome effect size reported earlier. To check the ralationship between design quality and effect size, the design quality index defined earlier was correlated with effect size for the 38 studies where both treatment groups had equal access to the content of instruction. The correlation was .21 ($p = .09$) indicating a trend toward higher effect sizes with improved research design. As the number of design flaws diminishes, the difference between experimental and control means increases,

Results, Analysis and Conclusions: Quantity of Instruction

Three quantity of instruction construct studies we found by searching both the published and dissertation literature. Two of the studies were very similar in design, Welch's (1968) and Economos' (1972). Both studies had teachers of physical science (Harvard Project Physics and Introductory Physical Science respectively) keep track of their total teaching days over set units of material. The unit test served as a criterion measure of cognitive achievement.

Welch found a non-significant -.08 correlation between teaching days and Unit 1 achievement based on the class means of 41 teachers. Economos found two non-significant correlations, .29 for Unit 1 and .17 for Unit II, based on the 20 class means of five teachers.

The third study (Tomera, 1974) compared the effect of two weeks of training in observation and comparison skills in seventh grade life science with four weeks of similar training (total n=80). After five months, no significant difference in ability to use the skills was found.

Taken as a whole, the three studies indicate that simply expanding the amount of time spent on a given unit of material holds no special relationship to amount learned. Since how the time was spent in each classroom was not reported, nothing about how to teach to a comparable level of achievement in a shorter period of time can be concluded.

## Summary and Recommendations

The task of this study was to identify quality of instruction clusters of five or more studies of the same or similar independent variables in the published science education grade 6-12 research of the 1963-1978 period, to quantitatively synthesize the studies within and across clusters, and to comment on the general quality of and gaps in the research. Fifty two of 95 studies met the 5 study criterion and revealed significant positive cognitive outcomes due to each of four types of instructional interventions: the use of preinstructional strategies, training in scientific thinking, increased structure in the verbal content of materials, and increased realism or concreteness in adjunct materials. Indirectness of instruction and inductive strategies showed no effect in general over direct or deductive strategies, but a trend toward more effectness of the indirect or inductive approaches in grades 10-12 and direct or deductive approaches in grades 6-8 was found. Combining the results of all clusters, systematic innovation in instruction resulted in significantly positive improvements over the norm or traditional practice.

Methodologically the research was judged particularly weak in reliability of treatment implementation and particularly vulnerable to threats of history and contamination, compensation or differential incentives. Improved design quality was related to larger effect sizes.

Certain recommendation evolve from the findings and the general experience of conducting this kind of research synthesis.

1) The replication of studies is important but the replication need not rigorously follow in detail an earlier study. All studies are flawed and limited in some way. Variation in flaws and strengths as well as in subject population can add to the generalizability of the cummulative results. To be useful in a practical sense, instructional interventions must be sufficiently robust to give positive results under a variety of less than optmal situations.

2) Several constructs, besides the quality of instruction, compete in explaining science learning. More of these constructs should be measured and brought into the analysis, especially in quasi-experiments. Even experimental designs would be improved if such factors as ability, motivation and classroom environment factors could be statistically removed and not

assumed to be neutralized by random assignment. This multivariate approach would also allow a better accounting of the sources of variance in outcomes and thereby lead to better prediction and control.

3) Research on the attitudinal impact of various instructional interventions is needed. Routinely, studies should consider multiple outcomes on both an immediate and long term basis. Few studies had delayed follow-up measures of any kind.

4) Findings from general education research should inform science education research. For example, the research on direct instruction techniques (Rosenshine, 1979) in lower grades should be examined and applied in science lessons to determine its limits of effectiveness.

5) Study reports should typically include the means and standard deviations of all treatment group outcome measures to make future quantitative syntheses possible and easier. Also, the generalizability of individual studies as well as future syntheses would benefit from greater attention to the description of the populations represented by the sample of students actually receiving experimental treatments. This should include at least community occupational composition, SES, and urban-suburban-rural character.

6) More attention needs to be given to insuring the reliability of treatment implementation and minimizing associated threats to study validity.

Without a quantitative synthesis of the research, the findings of this study would have remained qualitative and directional at best. The quantification of effect sizes and study variables has allowed a more objective and precise representation of the literature reviewed. The relatively small number of studies in each cluster has meant larger confidence intervals making significance of the findings more difficult to attain, but, where attained, more convincing. As the body of research literature grows, additional studies will form new data points in new clusters, building toward confidence in the general pattern of research findings.

## Reference Notes

1. Heartel, G.D., Walberg, H.J. & Haertel, E.H., Social psychological environments and learning: A quantitative synthesis. A paper presented at the Annual Meeting of the American Educational Research Association, April, 1979.

2. Iverson, B.K. and Walberg, H.J. Home environment and learning: A quantitative synthesis. A paper presented at the American Educational Research Association Annual Meeting, San Franciso, April 1979.

3. Walberg, H.J., Boulanger, F.D., Kremer, B.K. and Haertel, G.D. A meta-analysis of productive factors in science learning in grades 6 through 12. In preparation.

## References

Anderson, O. R. Quantitative analysis of structure in teaching. New York: Teachers College Press, 1971.

Babikian, Y. An empirical investigation to determine the relative effectiveness of discovery, laboratory, and expository methods of teaching science concepts. Journal of Research in Science Teaching, 1971, 8, 201-209

Bloom, B.S. Human characteristics and school learning. New York: Mc Graw Hill, 19

Campbell, J.R. Cognitive and affective process development and its relation to a teacher's interaction ratio. Journal of Research in Science Teaching, 1971, 8, 317-323.

Economos, J.M. Achievement gains as a function of teaching duration for Introductory Physical Science. Unpublished doctoral dissertation, University of Mississippi, 1972.

Glass, G.V. Integrating findings. The meta-analysis of research. In L.S. Shulman (Ed.), Review of Research in Eudcation (Vol. 6). Itasca, Illinois: F. E. Peacock, 1978.

Kline, A.A. A study of the relationship between self-directed and teacher-directed eighth grade students involved in an open-ended supplementary ESCP laboratory block. Journal of Research in Science Teaching, 1971, 263-271.

Leonard, W.H. and Lowry, L.F. Was there really an experiment? A quantitative procedure for verifying treatments in educational research. Educational Researcher, 1979, 8 (6), 4-7.

Olsen, R.C. A comparative study of the effects of behavorial objectives on class performance and retention in physical science. Journal of Research in Science Teaching, 1973, 10, 271-277.

Petersen, R.W. & Carlson, G.R. A summary of research in science education - 1977. New York: John Wiley, 1979.

Rosenshine, B.V. Content, time, and direct instruction, In P.L. Peterson & H.J. Walberg (Eds.), Research on teaching: Concepts, findings and implications. Berkeley: McCutchan, 1979.

Santiesteban, A.J. and Koran, J.J. Instructional adjuncts and learning science from written materials. Journal of Research in Science Teaching, 1977, 14, 51-55.

Tanner, R.T. Expository-deductive versus discovery-inductive programming in physical science principles. Journal of Research in Science Teaching, 1969, 6, 136-142.

Thomas, B. and Snider, B. The effects of instructional method upon the acquisition of inquiry skills. Journal of Research in Science Teaching, 1969, 6, 377-386.

Tomera, N.T. Retention of science processes of observation and comparison in junior high school students. Science Education, 1974, 58, 185-193.

Travers, R.M. (Ed.). The second handbook of research on teaching. Chicago: Rand McNally, 1973.

Uguroglu, M.E. and Walberg, H.J. Motivation and achievement: A quantitative synthesis. American Educational Research Journal, 1979, in press.

Walberg, H.J. A psychological theory of educational productivity. In F.H. Farley and N.U. Gordon (Eds), Contemporary perspective in educational psychology. Chicago: National Society for the Study of Education, in press.

Walberg, H.J., Schiller, D. and Haertel, G. The quiet revolution in education research. Phi Delta Kappan, 1979, 61, 179-183.

Welch, W.W. and Bridgham, R.G. Physics achievment gains as a function of teaching duration. School Science and Mathematics, 1968, 68, 449-454.

Yager, R.E. Priorities for reseach in science education: A study committee report. Journal of Reseach in Science Teaching, 1978, 15, 99-107.

Yager, R.E., Engen, H.B. and Snider, B.C. Effects of the laboratory and demonstration methods upon the outcomes of instruction in secondary biology. Journal of Research in Science Teaching, 1969, 6, 76-86.

## Footnote

This research was supported by the National Science Foundation Grant No. 78-17374, Principal Investigator Herbert J. Walberg. Findings and opinions expressed in this paper are solely the responsibility of the author, who wishes to acknowledge the critical contributions of James A. Kulik, University of Michigan and Wayne W. Welch, University of Minnesota and the comments of colleagues Geneva D. Haertel, Barbara K. Kramer, and Herbert J. Walberg, who are synthesizing research on other factors influencing science learning.

Table 1

Category and Frequency of Published

Quality of Instruction Studies 1963-1978

| Independent Variable | | Independent Variable | |
|---|---|---|---|
| Category | Frequency | Category | Frequency |
| Advance Organizer | 5 | PSI vs Mastery | 1 |
| Adjunct Questions | 1 | Pretest and Overview | 1 |
| Audio-Tutorial Inst | 1 | Process Training | 4 |
| Computer Ass'd Inst | 2 | Programmed Inst | 3 |
| Cuing | 2 | Questioning Level | 1 |
| Concrete vs Formal Inst | 2 | Reward Structure | 3 |
| Cognitive Conflict | 1 | Reinforcement | 1 |
| Difficulty Conflict | 1 | Set Induction | 2 |
| Guidance in Prob. Solving | 1 | Structure in Materials | 1 |
| Group Size | 3 | Student Choice | 2 |
| Indirect/Direct Ratio | 3 | Team Teaching | 2 |
| Inductive vs Deductive | 5 | TV vs Non-TV | 1 |
| Innovative Materials | 1 | Teacher Background | 4 |
| Kenetic Structure | 6 | Teacher Characteristics | 3 |
| Labelled Drawings | 1 | Teacher Training | 4 |
| Method, Lec. vs. Disc. | 1 | Teacher/Student Moves | 2 |
| Mode of Illustration | 2 | Tea. vs Stu. Generalization | 2 |
| Non-Direct vs Direct Inst | 5 | Teacher Experience | 1 |
| Open Ended Inst | 1 | Training in Log. Operations | 8 |
| Original Sources | 1 | Type of Discussion | 2 |
| Part vs Whole Film | 1 | Verbal vs Picture Mode | 1 |
| Pacing of Instruction | 2 | | |

Note: Each study related the independent variable to a measure of science learning.

Table 2

Study Clusters and Definitions

| Cluster | Component Variables | Operational Definitions | No. of Studies[1] |
|---|---|---|---|
| Preinstructional Strategies | Advance Organizers | Identified as such by study author | 4 |
| | Behavioral Objectives | | 5 |
| | Set Induction | | 2 |
| Directness of Instruction | Direct vs. Non-Direct | Teacher or workbook controlled instruction compared to instruction allowing greater student choice in content and/or method | 7 |
| | Indirect/Direct Ratio | Used Flanders Interaction Analysis | 2 |
| Inductive vs. Deductive Strategies | Same as cluster | Sequence of instructional components in two competing treatments such that one proceeded from rule or generalization to examples while the other reversed this sequence | 9 |
| Training in Scientific Thinking | Training in Logical Operations | Training in some Piagetian task related skill or logical operation | 7 |
| | Training in Science Processes | Identified as such by study author | 2 |

Table 2 (Continued)

| Cluster | Component Variables | Operational Definitions | No. of Studies |
|---|---|---|---|
| Structure in the Verbal Content of Materials | Same as cluster | Studies comparing higher with lower structure materials using Anderson's (1971) definition of structure | 5 |
| Realism or concreteness in Adjunct Materials | Same as cluster | Studies comparing treatments with adjunct materials at different points on the concrete-symbolic continuum | 9 |

1. Table 1 frequencies were first estimates of potentially useful studies and may not agree with the final number of studies listed here.

Table 3

Cognitive Outcomes by Cluster of Studies

| Cluster | Number of Studies | Comparisons: n | Positive Total | Positive $p < .05$ | Negative $p < .05$ | Effect Sizes: n | mean | .95 conf. interval |
|---|---|---|---|---|---|---|---|---|
| Preinstructional Strategies | 10 | 11 | 10 | 7 | 0 | 9 | 1.03 | ±.68 |
| Indirect Instruction | 9 | 10 | 5 | 3 | 2 | 10 | .11 | ±.27 |
| Inductive Strategies | 9 | 10 | 1 | 0 | 1 | 7 | -.22 | ±.25 |
| Training in Scientific Thinking | 9 | 12 | 10 | 8 | 0 | 11 | .89 | ±.59 |
| Structure in Verbal Content of Materials | 5 | 6 | 6 | 4 | 0 | 6 | .74 | ±.27 |
| Concreteness in Adjunct Materials | 9 | 9 | 9 | 5 | 0 | 9 | .58 | ±.22 |
| Totals | 51 | 57 | 41 | 23 | 3 | 52 | .55 | ±.21 |

Table 4

Study Variable Subgroup Comparisons

| Variable | Subgroups 1, 2 & 3 | Subg. 1 n | Subg. 1 $\overline{ES}$ | Subg. 2 n | Subg. 2 $\overline{ES}$ | Subg. 3 n | Subg. 3 $\overline{ES}$ | F | p |
|---|---|---|---|---|---|---|---|---|---|
| Design | 1) Quasi-Exper | 9 | .43 | 29 | .54 | | | .16 | .69 |
| | 2) True Exper | | | | | | | | |
| Grade Level | 1) 6 through 9 | 23 | .60 | 15 | .38 | | | .83 | .37 |
| | 2) 10 through 12 | | | | | | | | |
| Type of Course | 1) Elective | 12 | .57 | 13 | .60 | 13 | .38 | .31 | .73 |
| | 2) Required | | | | | | | | |
| | 3) Combination | | | | | | | | |
| Student Ability Level | 1) High | 5 | 1.08 | 24 | .40 | 9 | .52 | 1.90 | .16 |
| | 2) Average | | | | | | | | |
| | 3) Low | | | | | | | | |
| Component Manipulated | 1) Teacher Behavior | 1 | .84 | 21 | .44 | 16 | .59 | .31 | .74 |
| | 2) Materials | | | | | | | | |
| | 3) Combination | | | | | | | | |
| Experimental Treatment Teacher | 1) Regular | 15 | .55 | 6 | .64 | 17 | .44 | .19 | .83 |
| | 2) Special | | | | | | | | |
| | 3) Materials Only | | | | | | | | |

Table 4 (Continued)

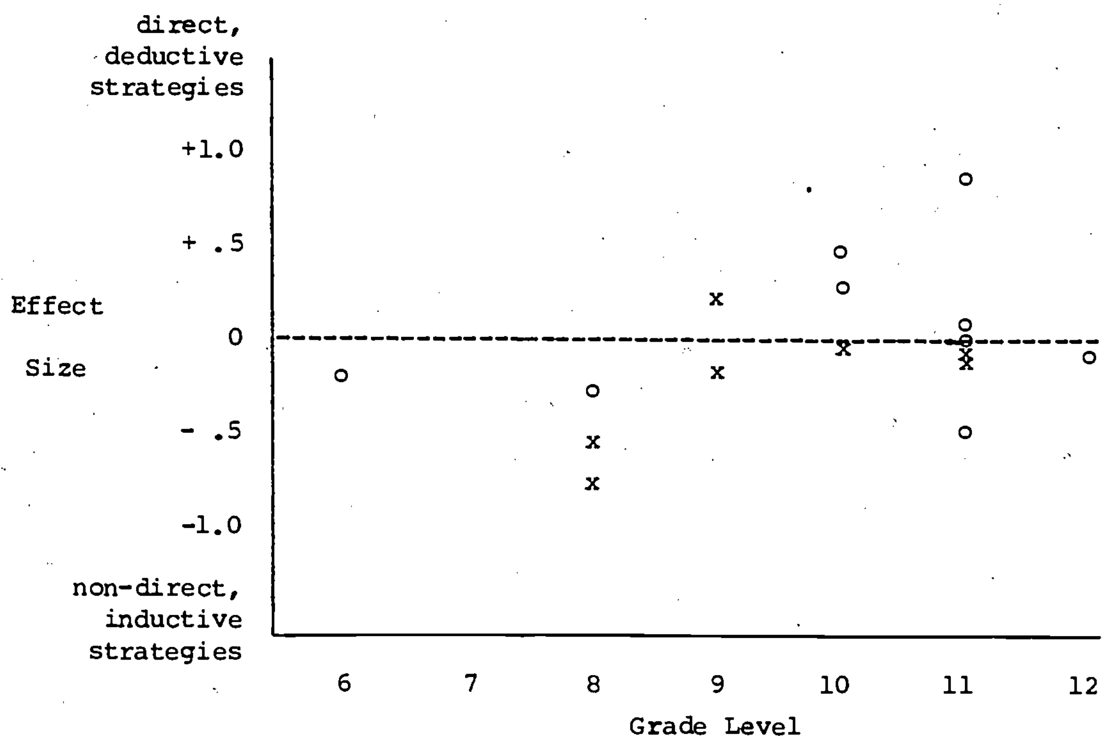
| Variable | Subgroups 1, 2 & 3 | Subg. 1 n | Subg. 1 $\overline{ES}$ | Subg. 2 n | Subg. 2 $\overline{ES}$ | Subg. 3 n | Subg. 3 $\overline{ES}$ | F | p |
|---|---|---|---|---|---|---|---|---|---|
| Focus of Instruction | 1) Non-Lab<br>2) Lab<br>3) Combination | 18 | .50 | 16 | .54 | 4 | .48 | .02 | .98 |
| Experimental Treatment to | 1) Individuals<br>2) Small Group<br>3) Class Group | 22 | .46 | 4 | .52 | 12 | .61 | .15 | .86 |
| Length of Treatment | 1) Less than 1 hr.<br>2) 1 to 10 hrs.<br>3) Greater than 10 hrs. | 11 | .66 | 12 | .43 | 15 | .47 | .31 | .74 |
| Source of Outcome Measure | 1) Local<br>2) Published | 29 | .41 | 9 | .84 | | | 2.41 | .13 |

Note: $\overline{ES}$ means mean effect size.

Figure 1

Cognitive Median Effect Size with Grade

Level of Studies in Two Clusters




Note: "o's" indicate indirectness of instruction clusters studies.

"x's" indicate inductive vs deductive cluster studies.

The correlation of Effect Size with Grade Level is .48.

Appendix

# Bibliography of Studies by Cluster

## Preinstructional Strategies

Bryant, N. and Anderson, H. A. Effects of performance objectives on achievement. Journal of Research in Science Teaching, 1972, 8, 369-375.

Johnson, C. C. and Sherman, J. E. Effects of behavorial objectives on student achievement in ISCS. Science Education, 1975, 59, 177-180

Kahle, J. B. and Rastovac, J. J. The effect of a series of advance organizers in increasing meaningful learning. Journal of Research in Science Teaching, 1976, 60, 365-371

Martin, W. J. and Bell, P. E. The use of behavorial objectives in instruction of basic vocational science students. Journal of Research in Science Teaching, 1977, 14, 1-11.

Olsen, R. C. A comparative study of the effects of behavorial objectives on class performance and retention in physical science. Journal of Research in Science Teaching, 1973, 10, 271-277.

Santiesteban, A. J. and Koran, J. J. Instructional adjuncts and learning science from written materials. Journal of Research in Science Teaching, 1977, 14, 51-55.

Schuck, R. F. The influence of set induction upon student achievement and perception of effective teaching. Journal of Research in Science Teaching, 1970, 7, 35-40.

Schuck, R. F. The influence of set induction upon pupil perception of effective teaching, achievement, and retention in units on respiration and circulation in BSCS curricula. Journal of Research in Science Teaching, 1971, 8, 51-56.

Weisberg, J. S. The use of visual advance organizers for learning earth science concepts. Journal of Research in Science Teaching, 1970, 7, 161-165.

West, L.H. T. and Fensham, P. J. Prior knowledge or advance organizers as effective variables in chemical learning. Journal of Research in Science Teaching, 1976, 13, 297-306

Indirectness of Instruction

Bernett, H. C. An investigation of relationships among biology achievement, perception of teacher style, and cognitive preferences. Journal of Research in Science Teaching, 1974, 11, 141-147.

Campbell, J. R. Cognitive and affective process development and its relation to a teacher's interaction ratio. Journal of Research in Science Teaching, 1971, 8, 317-323.

Crockett, R. K., Bartlett, K. R. and Elliot, H. G. A comparison of structured and unstructured modes of teaching science activities. Journal of Research in Science Teaching, 1976, 13, 267-274.

Humphreys, D. W. and Townsend, R. D. The effects of teacher-and-student-selected activities on the self-image and achievement of high school biology students. Science Education, 174, 58, 295-301.

Kline, A. A. A study of the relationship between self-directed and teacher-directed eighth grade students involved in an open-ended supplementary ESCP laboratory block. Journal of Research in Science Teaching, 1971, 263-271.

Montague, E. J. and Ward, R. M. The development of problem solving abilities in secondary school chemistry. Journal of Research in Science Teaching, 1968, 5, 354-357.

Penick, J. E., Schlitt, D., Bender, S., and Lewis, J. Student structured learning in biology. Science Education, 1975, 59, 13-18

Rainey, R. G. The effects of directed versus non-directed laboratory work on high school chemistry achievement. Journal of Research in Science Teaching, 1965, 3, 286-293

Wolfson, M. L. A consideration of direct and indirect teaching styles with respect to achievement and retention of learning in science classes. Journal of Research in Science Teaching, 1973, 10, 285-290.

Inductive vs. Deductive Strategies

Babikian, Y. An empirical investigation to determine the relative effectiveness of discovery, laboratory, and expository methods of teaching science concepts. Journal of Research in Science Teaching, 1971, 8, 201-209

Douglass, C. B. and Kahle, J. B. The effect of differentially sequenced individualized instructional materials on student achievement in biology. Journal of Research in Science Teaching, 1977, 14, 335-340.

Douglass, C. B. and Kahle, J. B. The effects of instructional sequence and cognitive style on the achievement of high school biology students. Journal of Research in Science Teaching, 1978, 15, 407-412.

Egelston, J. Inductive vs. traditional methods of teaching high school biology laboratory experiments. Science Education, 1973, 57, 467-477.

Hermann, G. D. and Hincksman, N. G. Inductive versus deductive approaches in teaching a lesson in chemistry. Journal of Research in Science Teaching, 1978, 15, 37-42.

Sakmyser, D. D. Comparison of inductive and deductive programmed instruction on chemical equilibrium for high school chemistry students. Journal of Research in Science Teaching, 1974, 11, 67-77.

Tanner, R. T. Expository-deductive versus discovery-inductive programming in physical science principles. Journal of Research in Science Teaching, 1969, 6, 136-142.

Thomas, B. and Snider, B. The effects of instructional method upon the acquisition of inquiry skills. Journal of Research in Science Teaching, 1969, 6, 377-386.

Voelker, A. M. Elementary school children's attainment of the concepts of physical and chemical change: a replication. Journal of Research in Science Teaching, 1975, 12, 5-14.

Training in Scientific Thinking

Bredderman, T. A. The effects of training on the development of the ability to control variables. Journal of Research in Science Teaching, 1973, 10, 189-200.

Case, R. and Fry, C. Evaluation of an attempt to teach scientific inquiry and criticism in a working class high school. Journal of Research in Science Teaching, 1973, 10, 135-142.

Cole, H. and Raven, R. Principle learning as a function of instruction on excluding irrelevant variables. Journal of Research in Science Teaching, 1969, 6, 234-241.

Hammond, J. and Raven, R. The effects of a structured learning sequence on the achievement on compensatory tasks. Journal of Research in Science Teaching, 1973, 10, 257-262.

Howe, A. C. and Mierzwa, J. Promoting the development of logical thinking in the classroom. Journal of Research in Science Teaching, 1977, 14, 467-472.

Lawson, A. E., Blake, A. J. D. and Nordland, F. H. Training effects and generalization of ability to control variables in high school biology students. Science Education, 1975, 59, 387-396.

Lawson, A. E. and Wollman, W. T. Encouraging the transition from concrete to formal cognitive functioning:an experiment. Journal of Research in Science Teaching, 1976, 13, 413-430.

Nelsen, B. J. The relationship of fifth and sixth grade students' achievement to preservice science teacher preparation. Journal of Research in Science Teaching, 1978, 15, 161-166.

Stronck, D. R. Comparative effects of three seventh-grade science programs with different laboratory materials. Science Education, 1971, 55, 125-130.

Structure in the Verbal Content of Materials

Anderson, O. R. The strength and order of responses in a sequence as related to the degree of structure in stimuli, Journal of Research in Science Teaching, 1966, 4, 192-198.

Anderson, O. R. The effects of varying struture in science content on acquisition of science knowledge. Journal of Research in Science Teaching, 1968, 5, 361-364.

Anderson, O. R. and Lee, M. T. Structure in science communications and students recall of knowledge. Science Education, 1975, 59, 127-138.

Ferraro, E., Lee, M. T. and Anderson, O. R. The effects of structure in science communications on knowledge acquisition and conceptual organization by students of varying mental maturity. Journal of Research in Science Teaching, 1977, 14, 441-447.

Simmons, E. S. The effects of kinetic structure on knowledge about and performance of psychomotor skill: teaching students to use the compound microscope. Science Education, 1977, 61, 19-28.

Realism or Concreteness of Adjunct Materials

Barufaldi, J. P. and Dietz, M. A. Effects of solid objects and two-dimensional representations of the objects on visual observation and comparison among urban children. Journal of Research in Science Teaching, 1975, 12, 127-132.

Cantu, L. L. and Herron, J. D. Concrete and formal Piagetion stages and science concept attainment. Journal of Research in Science Teaching, 1978, 15, 135-143.

Dwyer, F. M. An experiment in visual communications. Journal of Research in Science Teaching, 1969, 6, 185-195.

Holliday, W. G. The effects of verbal and adjunct pictorial-verbal information in science instruction. Journal of Research in Science Teaching, 1975, 12, 77-83.

Holliday, W. G., Brunner, L.L., and Donais, E.L. Differential cognitive and affective responses to flow diagrams in science. Journal of Research in Science Teaching, 1977, 14, 129-138.

Holliday, W. G. and Harvey, D. A. Adjunct labeled drawings in teaching physicis to junior high students. Journal of Research in Science Teaching, 1976, 13, 37-43.

Pella, M. O. and Sherman, J. A comparison of two methods of utilizing laboratory activities in teaching the course IPS. School Science and Mathematics, 1969, 69, 303-314.

Wollman, W. T. and Lawson, A. E. The influence of instruction on proportional reasoning in seventh graders. Journal of Research in Science Teaching, 1978, 15, 227-232.

Yager, R. E., Engen, H. B. and Snider, B. C. Effects of the laboratory and demonstration methods upon the outcomes of instruction in secondary biology. Journal of Research in Science Teaching, 1969, 6, 76-86.

Quantity of Instruction

Economos, J. M. Achievement gains as a function of teaching duration for Introductory Physical Science. Unpublished doctoral dissertation, University of Mississippi, 1972.

Tomera, N. T. Retention of science processes of observation and comparison in junior high school students. Science Education, 1974, 58, 185-193.

Welch, W. W. and Bridgham, R. G. Physics achievement gains as a function of teaching duration. School Science and Mathematics, 1968, 68, 449-454.

# Social-Psychological Environments and Learning: A Quantitative Synthesis

Geneva D. Haertel, Herbert J. Walberg, Edward H. Haertel

University of Illinois at Chicago Circle

Running head: Social-Psychological Environments

# Social-Psychological Environments and Learning: A Quantitative Synthesis

During the past decade and a half, educational researchers and evaluators issued or published about one hundred reports concerning student perceptions of the social-psychological dimensions of their classroom group such as cohesiveness, satisfaction, goal direction, difficulty, competitiveness, and friction. Reviews of this work (Randawa & Fu, 1973; Shulman & Tamir, 1973; Walberg, 1974, 1976, and Moos, 1979) discuss theoretical, methodological, and practical issues and conclude that such perceptions are useful as independent, mediating, and dependent variables in educational investigations in natural settings. Much of the research shows that social-psychological perceptual scales are reliable, and are sensitive to educational treatments such as curriculum, teacher training, and instructional innovations, as well as to project efforts to increase teamwork, cross-sex, cross-ethnic-group cooperation, and similar group properties. Other work reveals that such perceptions reflect and mediate teacher and student characteristics and that they provide diagnostically-valuable profiles of classroom climate and individual morale.

The focus of the present work is the predictability of end-of-course cognitive, affective, and behavioral learning from mid-course social-psychological perceptions, with and without statistical control for beginning-of-course measures, ability, or both. Even if constructive perceptions of the social environment are considered worthy ends in their own right, it is important to determine if they are positively associated with learning gains and outcomes. Consistent, positive associations indicate that it is unlikely that learning is being traded off for more constructive social-psychological morale; under certain assumptions (Walberg, 1976), such associations may indicate causal connections between social-psychological perceptions and learning.

It should, of course, be acknowledged that there are many educational, psychological, sociological, and even anthropological approaches to the measurement or operationalization of the social-psychological environment, climate, or morale of classes and schools. Behavioral psychologists, for example, in the study of groups, have often emphasized the frequency of leader and member behaviors (Bales, 1950). One sociological tradition has analyzed the socio-economic and racial-ethnic composition of classroom, peer, and school groups (Coleman, 1961); and anthropologists have studied the cultural relevance of classroom speech and other interactions to learning in ethnographic accounts (Tikunoff, Berliner, Rist, 1975). It seems premature and certainly beyond the scope of the present synthesis to analyze and integrate these somewhat disparate approaches; and the scope of the effort is therefore restricted to student ratings of their perceptions of social psychological characteristics of their classes and schools, a topic that includes a sufficient amount of quantitative information on the environment-learning relation (with statistical controls) while maintaining construct continuity of psychological constructs across the studies analyzed.

Reviews of research reported in twelve studies of ten large, independent data sets show that student perceptions of classroom climate can account for significant variance in a variety of cognitive, affective, and behavioral learner outcomes. It has not seemed possible until recently, however, to summarize the 734 correlations in these studies to determine, for example: Which perceptions are most predictive? What learnings are most predictable? And, how does the predictability vary across such factors as grade levels of students, subject matter, and methodological characteristics of the studies? Jones and Fiske (1953), Light and Smith (1971), Gage (1978), Rosenthal (1976) describe a number of quantitative techniques, which are employed here, to synthesize the quantitative findings across studies and to provide answers to such questions. (See Glass, 1978, for a critical exposition.) As in quantitative summaries of empirical works in the natural sciences, the techniques are intended to provide estimates across investigations of: the consistency of observations or coefficients such as means, correlations, and regression weights; their average magnitudes and margins of error; and their boundaries of application. The present application draws on the techniques developed by Glass (1978) for meta-analysis, by Mosteller and Tukey (1977) for obtaining appropriate error estimates when some of the data are not independent, and by ourselves for weighting independent data sources equally as well as estimating simultaneously the complete set of possible determinants of the correlation coefficients.

## Method

### Sample of Studies

A search was made of Dissertation Abstracts, Education Index, Psychological Abstracts, Social Science Citation Index, and, since much of the relevant research involved science curricula, the annual summaries of research sponsored by the National Association for Research in Science Teaching for the years 1963 through 1977. On-going, unpublished studies known by the authors or those cited in recent works were also considered for inclusion. Studies were considered that involved naturalistic classroom settings, kindergarten through twelfth grade, and that reported simple, partial, and part correlations between student perceptions of social-psychological climate of their classes and end-of-course learning.

Many different qualities of the student-perceived social-psychological environments of classrooms have been quantified. Because of the difficulties involved in determining whether or not subscales of different instruments measure the same construct, a single instrument, the Learning Environment Inventory (LEI), by Anderson and Walberg (Note 1), was designated the "anchor instrument" for the research synthesis, and correlations from various studies were categorized as involving one of the subscales on the LEI. There were several reasons for the selection of the LEI as "anchor." First, the LEI incorporates a broad range of 15 subscales measuring different aspects of the learning environment and reflecting a broad conceptualization of social-psychological dimensions found in many social collectivities such as hospitals, prisons, and workgroups, corporations and fraternities (Insel & Moos, 1974). Second, the psychometric properties of the LEI, including the reliability and factorial purety of its many subscales have been thoroughly investigated (Anderson and Walberg, Note 1). Finally, ten of the existing studies meeting all other criteria for inclusion, employ the LEI itself, or instruments derived directly from the LEI. These are among the studies listed in Table 1. The search and selection procedures yielded

---

Insert Table 1 about here

---

twelve investigations of ten data sets that report 734 correlations calculated from a combined total of 17,805 students in 823 classes (Table 1). The correlations from three studies of a single data set (Walberg, 1969a and b, and 1972) that explored predictability of learning across units of analysis and statistical-control techniques were counted as a single data source in the main regression analyses as explained in a subsequent section.

Features of Studies

Table 2 summarizes the key features of the twelve studies.

---

Insert Table 2 about here

---

Information on each study includes a description of the environment measures used, the variables controlled, learning outcomes, and a brief statement of results.

In general the environment measures employed have internal-consistency reliabilities between .41 and .86. No strong systematic relationship is evident between the reliability of the measures and the grade level at which data were collected.

The outcome measures include not only standardized achievement tasks, but affective and behavioral measures as well. Ten of the twelve included some cognitive measure of achievement, seven include affective or interest measures; and five employ a behavioral measure such as daily attendance.

Nine of the twelve studies are statistically controlled. Most studies control for the corresponding pretest; four studies control for student aptitudes such as IQ, and two control for instructional variables such as teacher attitude.

Although the focus of the present synthesis is the magnitude of the correlations for specific environment scales an` learning outcomes, the multivariate results for sets of environment scales in the last column of Table 1 may be noted. Seven studies added sets of these scales to regression equations containing ability or pretest measures or both as controls, and reported the percentage increment in accountable variance. The average incremental variance accounted for on 19 learning outcomes is 20 percent with a range of 1 to 54 percent. Thus, regressions containing control and perceptual variables account for large amounts, in some cases, nearly all, of either the total or reliable variance in learning outcomes.

Characteristics of Correlations

Information on eight characteristics was recorded for each simple, part, or partial correlation: national location of the study; grade level and number of students; unit of analysis; type of correlation; type of social-psychological perception; outcome domain; and content area of subject-matter (Table 3).

Insert Table 3 about here

The continuous variables, grade level and number of students were grouped into class intervals to calculate frequencies and one-way analyses of variance, but were left in their full continuous precision for the regression analyses. The nominal variables such as location and unit of analysis were treated as categorical in the analyses of variance and converted to sets of binary (0,1) variables in the regressions, as explained in the discussion of the results.

Most of the variables listed in Table 3 are self-explanatory. However, a few deserve explanation. Unit of analysis refers to the level of aggregation used in the data analysis, i.e., student, sub-group, class, or school. With the exception of Moos and Moos (1978), and Bardsley (1976), all investigations employed the Learning Environment Inventory (LEI) in original or simplified or shortened form to reduce the 25-minute time required for the full 105 items for the 15 scales. For example, Perkins (1976) and Talmage and Walberg (1978) employed the 45-item My Class Inventory, an adaption of the LEI for elementary school. Since the five scales on the Moos and Moos instruments and adaptions of the LEI correspond closely, it was possible to code all scales to correspond to the LEI scales.

Outcome domain refers to the type of criterion measure used. Criterion learning measures are coded as either cognitive, attitudinal or behavioral. Types of cognitive measures include conventional multiple-choice achievement tests and tests of understanding, critical thinking, and tests of formal reasoning. Attitudinal criteria include instruments such as interest measures and motivation and self-concept tests. Behavioral criterion measures include self-report activity inventories and absence rates.

The last item coded for each correlation coefficient is the outcome content area. Each correlation is coded for one of the following subjects: general science, life science, physical sciences, mathematics, social sciences, humanities, general achievement or miscellaneous subject areas which are aggregations across several of these categories.

Data Analysis

In the analyses, the magnitude of the correlation of environment scale and learning outcome is the dependent variable. The specific LEI scale, outcome domain, the sample size, unit of analysis, and the other explanatory factors are the independent variables. Several analytical techniques were used, beginning with a tabulation of the signs of the correlation coefficients by expected direction, proceeding through one-way analyses of variance and culminating in a series of multiple regression analyses. In the regression analyses, correlations were weighted to equalize the contributions of the different studies. The rationale for these procedures and details of their execution are explained in the next section, but the weighting issue deserves discussion here.

The 734 correlation coefficients are by no means statistically independent since they arise from only twelve studies and ten independent data sets. Table 1 shows, moreover, that the number of correlations taken from individual studies varies from 5 to 240. Weighting each correlation equally would give 48 times more weight to the latter study. For these reasons, special procedures were developed for the regressions to give each data set equal weight. Tukey's "Jackknife" procedure (Mosteller & Tukey, 1977; Glass,

1978) was used to obtain estimates of the regression weights and their standard errors which provided statistically valid tests of the significance of each predictor.

## Results and Discussion

### Directional Hypotheses

The 15 LEI subscales include some positive and some negative characteristics of the classroom environment. The positive subscales are: Cohesiveness, Satisfaction, Task Difficulty, Formality, Goal Direction, Democracy, Environment, and Competition. The negative scales are: Friction, Cliqueness, Speed, Apathy, Favoritism, Disorganization, and Diversity. In subsequent analyses the signs of correlations involving negative aspects of the classroom environment are reversed. Thus, the expected signs of correlations with all LEI scales, as coded for these analyses, are positive.

From social-psychological research, Walberg (1969b) derived 36 hypotheses concerning the direction of relations between selected LEI scales and learning criteria, namely that Cohesiveness, Satisfaction, Task Difficulty, Goal Direction, Democracy, Diversity, and Environment would be positively correlated with learning outcomes, and that Friction, Cliqueness, Apathy, Favoritism, Disorganization would be negatively correlated with cognitive, affective, and behavioral learning. Data assembled for the present research synthesis permitted the testing of these hypotheses by tabulating the number of studies in which correlations of each sign were found, for each scale. The prediction for each of the 12 scales gave rise to 3 testable assertions, namely that correlations between that scale and measures in each of the three outcome domains would have the predicted sign. The resulting 36 testable assertions were evaluated as follows. Each study employing a given combination of scale and outcome was examined to determine whether the preponderance of coefficients is positive, negative, or evenly split. The three Project Physics studies that explored the consistency of correlations across class and individual units of analysis and analytic techniques were combined since they arose from a single data set (Tables 1 and 2). Ties were broken by assigning even splits the values plus, then minus, then plus, and so on. A tabulation of the results shows that 31 of the 36, or 86 percent, of the signs support the hypotheses; and the binonimal probability of an even split in a sign test is less than .001. Three of the five disconfirmations concern the Diversity subscale, which shows negative relation with outcomes in all three domains, rather than the hypothesized positive relation.

## Unweighted, Univariate Analyses

Table 3 reports the means, standard deviations, and frequencies of the 734 correlations grouped according to each of the eight factors individually. These descriptive statistics and the F-tests for the corresponding correlation-weighted, one-way analyses of variance are intended to show trends, variations, and frequencies in the correlations, for each of the eight factors separately, before presenting the results of the regressions. The latter combine all eight factors in a simultaneous analysis that controls each factor for the others and points to a few strong trends that summarize much of the variation in the magnitudes of the correlations.

The F-ratio for location of study is highly significant ($F = 35.58$, $df = 3,730$, $p < .001$). Table 3 shows that the correlations from studies in India and Canada are higher than those in the United States and Australia. This finding may be explained by reference to the specific studies from which these correlations were taken. The single study in India (Walberg, Singh, and Rasher, 1977) used only extreme groups of students, nominated as most and least studious in their classes. These subgroups were the unit of analysis. The single Canadian study (Walberg and Anderson, 1972) is distinctive in that class, rather than student, was the unit of analysis. Aggregation effects may have raised correlations from these countries since analyses of collectivities usually yield stronger correlations.

The F-ratio comparing correlations by grade level is not significant ($F = 0.10$, $df = 2,731$), which may be due in part to the unequal group sizes which resulted from the collapsing of the twelve grades into only three levels (elementary, junior high, high school). In the regression analyses discussed below, the grade levels were not collapsed; and grade level proved to be a strong predictor of correlation size.

Grouping the correlations by sample size yields a significant F-ratio ($F = 23.25$, $df = 3,730$, $p < .001$). The absence of linear trend across the different sample size categories, however, may suggest that the apparent effect of sample size is attributable to variations across studies in unit of analysis, type of correlation coefficient, type of outcome measure used, or other factors. The regression analyses were employed to test these possibilities.

With the exception of the mean correlation for "subgroups," the correlations show stronger relationships with larger units of analysis. Differences among the mean correlations with different units of analysis are clearly significant ($F = 19.44$, $df = 3,730$, $p < .001$). The anomolous value for subgroups may again be explained by reference to the peculiarities of the single study using this unit of analysis, by Walberg, Singh, and Rasher (1977). The method used to select a sample in this study may have given rise to unusually high correlation coefficients. Aside from the "subgroups" anomaly, the strength of the environment-outcome relation, uncontrolled for the other seven factors, increases as larger and larger units are examined.

No significant differences were found between simple, part, and partial correlation coefficients ($F = .83$, $df = 2,731$). In some cases, partialling out ability or pretest scores or both and analyzing adjusted scores may increase precision and raise the correlation. In other cases, this increase may be more than offset by the attenuation due to the lowered reliability of the adjusted scores.

Differences among aspects of the environment measured, are highly significant ($F = 4.43$, $df = 14,719$, $p < .001$). The subscales which show the strongest relations to learner outcomes are Cohesiveness, Friction, and Satisfaction, all of which show average correlations of over .22 with outcomes. These are followed by Cliqueness, Difficulty, Apathy, Favoritism, Direction, Democracy, Disorganization, and Environment, with averages in the range from .16 to .12. The remaining four subscales, Speed, Formality, Diversity, and Competition, all show average correlations of less than .07.

The analysis of variance for outcome domain showed significant differences among the three outcome domains ($F = 19.67$, $df = 2,731$, $p < .001$). Higher correlations were observed with outcomes in the Cognitive domain than with those in either the attitudinal or behavioral domains.

The last of the eight one-way analyses of variance contrasts the eight content areas in which outcomes were related to environments ($F = 12.21$, $df = 7,726$, $p < .001$). Table 3 shows that the content areas in which outcomes are most predictable from environments are mathematics and the social sciences, followed by general science, the physical sciences, and the humanities. The category "general achievement" includes standardized test scores summed over several content areas. These indexes are all in the cognitive domain, and the relatively high mean correlation for this area may reflect primarily the exceptional reliability of such measures.

Regression Analyses and Jackknifed Estimates

Before the regression analyses, the categorial variables were replaced by sets of binary variables. Location, for example, was recoded into four variables, called "USA," "Canada," "Australia," and "India." Each of these variables was given an identifying value of either zero or one for each correlation. For a correlation computed on a sample from the United States, "USA" is 1, and "Canada", "Australia," and "India" are each 0. For an Australian study, "Australia" would be 1 and the rest 0, and so on. If the values of three of these variables are known, the fourth can always be determined (is redundant); therefore only three need be entered in the regression, and the convention was adopted to omit the last value of each categorical variable in Table 3. The continuous variables, grade and sample size, were left in their full metric precision.

In addition to recoding categorical variables, weights were introduced to equalize the contributions of the ten data sets to the estimates. Each correlation was given a weight proportional to the inverse of the number of correlations from its study. These weights were scaled so that the average weight for each coefficient was 1.00; thus the sum of the weights is the number of coefficients in the sample.

After the recoding and dropping of one variable from each binary set, 33 variables were available for the regression analysis: 3 variables for location, 1 variable for grade level (not recoded), 1 variable for sample size (not recoded), 3 variables for unit of analysis, 2 variables for correlation type, 14 variables for subscales, 2 variables for outcome domain, and 7 variables for content area. Consideration of variables representing interactions of LEI scales with sample and study characteristics could introduce still more variables. It would be quite unusual if all of these variables possessed significant power to predict the size of the correlation coefficients after controlling for all the other variables. The problem, therefore, was to decide which of the 33 variables and additional interactions should be included in the final regression equation. To screen the variables, a multi-stage procedure was employed, whereby weak or colinear predictors were successively eliminated. First, a run was made using all 33 predictors. Then those with F-ratios of less than 1.00 were eliminated, and a second run was made. Then all remaining variables with F-ratios less than 2 were dropped, and the remaining variables were used in a third run. Variables in this run which showed F-ratios of less than 4 were eliminated, and a fourth run was conducted. This run included only 18 of the original 33 variables. The regression with these 18 variables will be referred to as the reduced model. At this point, sets of product variables were introduced that measure the influence of interactions of significant LEI scales with grade level and unit of analysis. To determine which of these product variables possessed additional explanatory value, a stepwise procedure was employed, in which the 18 variables already identified were forced into the equation and product terms were then entered one at a time, using an F ratio of 4.00 as the criterion for entering each new variable. The final criterion corresponded closely to significance at the .05 level. A total of 32 variables were included in the final equation. In addition to the 18 variables in the last reduced equation, 14 cross-product terms, representing interactions, were introduced. The final equation with 32 variables will be referred to as the product model.

Conventional significance tests computed for the regression coefficients assume that the correlations are statistically independent. Since in all cases several or many correlations are taken from the same study, this assumption is not met. Accordingly, the significance of the coefficients was estimated using the "jackknife" (Mosteller & Tukey, 1977), assuming studies to be independent, but making no assumptions about the independence of two or more correlation coefficients taken from the same study. To apply the jackknife, the final regression equations for the reduced (18-predictor) and cross-product (32-predictor) models were computed ten times, each time omitting all correlations from one of the data sets. These ten regression equations, together with the original equation were then used to obtain new, robust estimates of the unstandardized regression coefficients and their standard errors. For each of these estimated b-weights, a t-ratio was computed, on nine degrees of freedom (since there were ten data sets).

Table 4 presents the original and jackknifed estimates of all

---

Insert Table 4 about here

---

regression coefficients for both the reduced and product models. Significance tests (t-ratios) are also shown. In the leftmost column, the 18 variables included in both models are listed, and the next four columns give conventional estimates and t-ratios followed by jackknifed estimates and t-ratios for the reduced (18-variable) model. The rightmost four columns give the same information on these variables for the product model. For the 14 crossproduct terms in the product model, only the jackknifed estimates and t-ratios are presented. These appear in the note at the end of the table.

Reduced Model

As shown in Table 4, jackknifing showed 10 of the 18 coefficients from the reduced equation and 15 of the 32 coefficients from the product model to be significant at the .05 level. The jackknifed estimates are similar to the original regression estimates; but the t values are somewhat lower on average, indicating that the significance levels of the original estimates are somewhat inflated as a result of the non-independence of correlations from the same study. The majority of the independent variables in the jackknifed reduced equation, however, are significantly related to the magnitude of the correlations between LEI scales and learning outcomes.

Among the variables representing location of the study, USA had a jackknifed b-weight significantly less than zero in the reduced model. Holding other things constant, correlations taken from studies conducted in the United States are estimated to be .22 less than those from studies abroad.

Grade level shows a small but persistent positive relationship to correlation size; the average correlation rises roughly .05 per year. Older high school students have longer school experience, usually attend more classes each week, and may thus be more astute raters of the class learning environment.

With other variables in the reduced equation held constant, a clear trend appears in the size of correlations as a function of the unit of analysis. Although only the variable representing class as unit-of-analysis is significant by the jackknife procedure, t-tests for student and subgroups as unit are nearly significant at the .05 level, and reveal, in the context of the other variables, a monotonic increase in the magnitude of the correlations with increasing aggregation from student to subgroup to class to school as unit of analysis.

The strength of the LEI scale-outcome relation was found to be significantly higher for seven of the scales than for the others. This is shown by the t-ratios in Table 4 for the reduced model, jackknifed estimates, for Cohesiveness, Friction, Satisfaction, Favoritism, Goal Direction, Democracy, and Environment. Correlations of these scales with learning outcomes are estimated by the regression to be from .21 to .38 higher than for the other eight scales, when all other factors are controlled.

It is notable that type of correlation (simple, part, or partial) is not significant in either the one-way analyses of variance or the regression analyses. Differences are statistically undetectable between (1) simple correlations of perceptual scales and learning outcomes and (2) part or partial correlations controlling ability, cognitive, or affective pretests, or both. One possible explanation for this finding is that the unreliability of adjusted scores may compensate for increased control for appitudes. However Anderson and Walberg (1974) show that IQ contributes little to the prediction of adjusted gains in cognitive, affective, and behavioral learning in several data sets whereas LEI scales contribute substantially. Thus, correlations may be unaffected by statistical controls because the scales in fact measure determinants of learning that are independent of aptitudes and pretests.

It is also notable that learning domain is not significant: the correlations of perceptions and cognitive outcomes do not differ significantly from those involving affective and behavioral learning outcomes. Thus, it appears that constructive aspects of class morale are equally associated with outcomes in all three domains rather than being associated with benefits in one domain sacrificed for losses in another.

The reduced model in Table 4 may be used to estimate the sizes of correlations to be expected under specific conditions in future research. This is done by adding together selected coefficients, or multiples of coefficients. The coefficient for the constant, .42, is always included. This value alone is the estimate for a correlation from a study not in the United States, not in Australia, at the kindergarten level ("grade 0"), with school as the unit of analysis, and not involving any of the twelve LEI scales for which binary variables were included in the reduced model. To estimate correlations at higher grade levels, the coefficient for grade level, .05, times the grade (1 through 12) is added in. Coefficients for binary variables are added in if the corresponding conditions obtain. For example, the estimate of a correlation at the tenth grade level, in the United States, with student as the unit of analysis, involving the satisfaction subscale would be .42 + 10 x .05 - .22 - .82 + .38, or .26. Further illustrations appear in Table 5. Least confidence can be placed in the estimates

Insert Table 5 about here

of correlations for elementary grades since only two of the ten data sets were obtained at this level (Tables 2 and 3). Extreme caution must be taken in extrapolating estimates beyond the data ranges given in Tables 2 and 3; caution is also required for interpolated estimates within sparsely sampled data regions such as those for India, elementary and junior-high grades, and part correlations (Table 3) as well as sub-spaces of these regions, which would require additional empirical investigation to enlarge the areas of confident estimation.

Product Model

Entering significant interactions of grade level and of unit of analysis with LEI scales changes the magnitude and significance of the b-weights but not their sign (Table 4). For example, the estimated b-weight for grade level changes from .05 ($t$=2.80, $p < .05$) to .06 ($t$=1.01, N.S.). The negative signs of the coefficients for products of grade level with Cohesiveness and Satisfaction (footnote to Table 4) indicate that estimates of correlations involving these two affective perceptions of class morale increase more slowly with grade level than do correlations involving other scales. Confirmations of these trends in additional empirical investigations, particularly in the elementary grades, would support the interpretation that organizational and task aspects of the social environment strengthen and affective aspects weaken relative to one another with increasing grade level.

Six interactions of unit of analysis and LEI scale are significant (Table 4 footnote). As in the case of grade level, they call for qualifications of the general relations of perceptions and learning; these interactions reveal stronger associations of some specific scales than others at certain levels of analysis. Since only one study used schools as units and another used sub-groups within classes, these interactions do not warrant much interpretation until further empirical studies are conducted. Cronbach (1976 and personal communications) and Walberg (1976) have discussed substantive and methodological issues of varying empirical relations across units of analysis. Future investigations can contribute to the understanding of these complexities if parallel analyses are conducted using the student, sub-group, class, and school as units of analysis.

Although the individual b-weights for the product terms should be interpreted cautiously, on the whole, inclusion of statistically significant product terms improved the fit of the model. This is indicated in Table 4 by the increase in the multiple correlation from .57 for the reduced model to .71 for the product model. Illustrative coefficient estimates derived from each of the two models are presented in Table 5, together with corresponding observed coefficients taken from two of the original studies included in the research synthesis. These are observed and estimated coefficients for U.S. classes, computed for grades 4 (elementary) and 11 (high school). The signs and magnitudes of the observed and estimated coefficients in Table 5 are in good agreement, given the standard errors for the two models as reported in Table 4. It should be noted that the estimates for high school samples are clearly more accurate, and more empirical work on elementary samples is in order. As would be expected, the estimates derived using the product model are in somewhat closer agreement with the observed values than those derived from the reduced model. The difference in goodness of fit of the two models is not large, however, and for most purposes either set of estimates provides a reasonable summary of the data structure as well as expected sizes of correlations for future empirical investigations.

## Conclusions

Across ten data sets from four countries and in a variety of samples, subject matters, and methodological approaches, perceptual aspects of the social-psychological environment of learning are correlated consistently in sign in their relation to cognitive, affective, and behavioral learning outcomes with or without statistical controls for ability, pretests, or both. Specifically, these learning outcomes are postively associated with perceptions of Cohesiveness, Satisfaction, Task Difficulty, Formality, Goal Direction, Democracy, and Environment and negatively associated with perceptions of Friction, Cliqueness, Apathy, Disorganization and Favoritism. As a set, these perceptions account for substantial variance in learning outcomes, beyond that accounted for by ability and pretest measures.

The correlations differ significantly in magnitude across perceptual scales, units of analysis, nations, and grade levels, as well as combinations of scales and the other factors. Although these differences require further empirical investigation, the theoretical plausibility and incremental predictive validity of the scales, as well as their utility for further research and evaluation, seem warranted. Their causal relation to learning is plausible but unproven. Educators who doubt the causal relation, however, or who believe in the inherent value of learning environment properties as ends in themselves rather than as means to standard outcome measures may not need to fear sacrificing one for the other since they appear to go together.

## Footnotes

1. The authors thank Maurice J. Eash and Harriet Talmage for institutional support. The research presented in this article was supported by the National Institute of Education (Grant No. NIE-G-78-0090) and the National Science Foundation (Grant No. NSF-78-17374); the points of view and opinions stated do not necessarily represent the official position or policy of either agency.

2. The authors thank Professor Chad Ellet at the University of Georgia for providing data from an unpublished dissertation (Perkins, 1976).

Reference Note

1. Anderson, G. J. and Walberg, H. J. The assessment of learning environments: A manual for the Learning Environment Inventory and My Class Inventory. Unpublished manuscript, 1974. (Available from Dr. Herbert J. Walberg, Office of Evaluation Research, College of Education, University of Illinois at Chicago Circle, Chicago, Ill., 60680.)

## References

Bales, R. Interaction process analysis: A method for the study of small groups. Reading, Massachusetts: Addison-Wesley, 1950.

Bardsley, W. N. Student alienation and commitment: A multivariate analysis of the effects of home and school environments. Unpublished doctoral dissertation, Australian National University, 1976.

Coleman, J. S. The adolescent society. New York: Cromwell-Collier Publishing Company, 1961.

Cronbach, L. J. Research on classrooms and schools: Formulation of questions, design, and analysis. Occasional paper, Stanford Evaluation Consortium, Stanford University, California, 1976.

Fraser, B. Evaluation of a science-based curriculum. In H. J. Walberg (Ed.), Educational environments and effects. Series on Contemporary Educational Issues. Chicago: University of Chicago Press, 1979.

Gage, N. L. The scientific basis of the art of teaching. New York: Teachers College Press, 1978.

Glass, G. V. Integrating findings: The meta-analyses of research. In L. S. Shulman (Ed.), Review of Research in Education, Vol. 5. Itasca: Peacock, 1978.

Insel, P.M. & Moos, R.H. Psychological environments: Expanding the scope of human ecology. American Psychologist, 1974, 29, 179-188.

Jones, L. V. and Fiske, D. Models for testing the significance of combined results. Psychological Bulletin, 50, (3), 1953, 375-382.

Light, R. J. and Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 421-471.

Moos, R. H. Educational climates. In H. J. Walberg (Ed.), Educational environments and effects. Series on Contemporary Educational Issues. Chicago: University of Chicago Press, 1979.

Moos, R, and Moos, B. Classroom social climate and student absences and grades. Journal of Educational Psychology, (in press).

Mosteller, F. and Tukey, J. W. Data analysis and regression: A second course in statistics. Reading: Addison-Wesley, 1977.

Perkins, M. L. A canonical correlational analysis of the relationships among school climate, teacher morale, and educationally-relevant performance of fourth grade students. Unpublished doctoral dissertation. University of Georgia, 1976.

Randhawa, B. S. and Fu, L. L. W. Assessment and effect of some classroom environment variables. American Educational Research Journal, 1973, 43, 303-321.

Rosenthal, R. Experimenter effects in Behavioral research. New York: Irvington Publishers, 1976.

Shulman, L. S. and Tamir, P. Research on teaching in the natural sciences. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand Mc-Nally, 1973.

Talmage, H. and Walberg, H. J. Naturalistic decision-oriented evaluation of a district reading program. Journal of Reading Behavior, 1978, X, 185-195.

Tikunoff, W. J., Berliner, D. E. & Rist, R. C. An ethnographic study of the forty classrooms of the Beginning Teacher Evaluation Study Known Sample, Technical Report No. 75-10-5 (San Francisco: Far West Laboratory for Educational Research and Development, 1975).

Tisher, R. P. and Power, C. N. The effects of classroom activities, pupil's perceptions and educational values in lessons where self-paced curricula are used. Research Report. Australian Advisory Committee on Research and Development in Education, 1975.

Walberg, H. J. Social environment as a mediator of classroom learning. Journal of Educational Psychology, 1969, 60, 443-448. (a)

Walberg, H. J. Predicting class learning: An approach to the class as a social system. American Educational Research Journal, 1969, 6, 529-542. (b).

Walberg, H. J. Social environment and individual learning: A test of the Bloom model. Journal of Educational Psychology, 1972, 63 (1), 69-73.

Walberg, H. J. (Ed.) Evaluating educational performance: A sourcebook of methods, instruments, and examples. Berkeley: McCutchan, 1974.

Walberg H. J. Psychology of learning environments. In L. S. Shulman (Ed.), Review of research in education, Vol. 4. Itasca, Illinois: F. E. Peacock, 1976.

Walberg, H. J. and Anderson, G. J. Classroom climate and individual learning. Journal of Educational Psychology, 1968, 59 (6), 414-419.

Walberg, H. J. and Anderson, G. J. Properties of the achieving urban class. Journal of Educational Psychology, 1972, 63, (4), 381-385.

Walberg, H. J. and Anderson, G. J. Learning environments. In H. J. Walberg (Ed.) Evaluating educational performance: A source book of methods, instruments, and examples. Berkeley, California: McCutchan, 1974.

Walberg, H. J., Singh, R., and Rasher, S. P. Predictive validity of student perception: A cross-cultural replication. American Educational Research Journal, 1977, 14, (1), 45-49.

Table 1

Statistics on Twelve Studies

| Study | Number of Correlations | Number of Classes | Number of Students |
|---|---|---|---|
| Bardsley (1976) | 7 | 30 | 374* |
| Fraser (1979) | 11 | 153* | 541 |
| Moos and Moos (1978) | 10 | 19* | 375 |
| Perkins (1976) | 5 | 108 | 3700 |
| Talmage and Walberg (1978) | 5 | 59* | 1600 |
| Tisher and Power (1975) | 240 | 20 | 315* |
| Walberg and Anderson (1968) | 22 | 76 | 2600* |
| Walberg and Anderson (1972) | 150 | 64* | 1600 |
| Walberg, Singh and Rasher (1977) | 60 | 150* | 3000 |
| Subtotal | 510 | 679 | 14,105 |
| Walberg (1969a) | 84 | 144* | 3700 |
| Walberg (1969b) | 84 | 144* | 3700 |
| Walberg (1972) | 56 | 144 | 3700* |
| Subtotal | 224 | 144 | 3700 |
| Total | 734 | 823 | 17,805 |

Note: Since three studies (Walberg, 1969 a & b, 1972) analyzed a single data set to explore predictability across units of analysis and other methodological variations; the correlations from these studies were combined in the regression analyses to give each data set equal weight.

Table 1 continued

*The units of analysis for each study are indicated with an asterisks. See text on the sub-group analysis in the Walberg, Singh, and Rasher (1977) study. See text on the use of school as unit of analysis in the Perkins (1976) study.

Table 2

Design Characteristics and Results of Twelve Studies

| Author (date) | | Environment Measures | Outcome Measures | Controls | Results |
|---|---|---|---|---|---|
| Bardsley (1976) | 374 senior high students in Australia. Student is unit of analysis | A single questionnaire item which corresponds to the Formality subscales of the Learning Environment Inventory (LEI). | Series of adjustment variables: misfeasance, self-estrangements, social powerlessness, value isolation, meaninglessness, social isolation, task powerlessness. | None | Correlation of "Rules" item with subscales are weak: misfeasance =.01; self-estrangement = .14, social powerlessness = .24, value isolation = .00, meaningfulness = .06, social isolation = .04 task powerlessness = -.07 |
| Fraser (1978) | 541 students in 20 seventh-grade general science classes, Melbourne, Australia, 10 experimental and 10 traditional curriculum; 153 sub-groups classified by sex, socio-economic status, and ability as within classes units | Modified, 55-item of the LEI containing 9 scales with internal consistencies from .50 to .80 | Seven cognitive and affective outcome measures ranging in internal consistency from .63 to .91. | Pretests, student attitudes, and instruction | In guided stepwise regression, aptitude alone produced Rs of from .48 to .76; instruction raised these by .16 to .26; and LEI raised R additionally from .17 to .47. Total R ranged from .81 to .86. |

Table 2 continued

| | | | | | |
|---|---|---|---|---|---|
| Moos and Moos (1978) | 19 representative classes in a U.S high school; classes as units | 90-item, 9-scale Classroom Environment Scale with internal consistencies from .67 to .86 | Attendance and average grade given in class | None | Absences correlated with Competition and Teacher Control; grades correlated positively with Involvement, Affiliation, and Teacher Support and negatively with Rule Clarity and Teacher Control. |
| Perkins (1976) | About 3,700 elementary school students in grade 4 from 42 U.S schools; School was the unit of analysis | My School which is identical to My Class contains 5 subscales: Cohesiveness, Competitiveness, Friction, Difficulty and Satisfaction. | Five subscales of the Iowa Test of Basic Skills: Vocabulary, Reading, Language Skills, Work Study Skills, Math Skills; and average daily attendance | Teacher attitude variance partialed out. | Positive relationship between performance on My Class and student performance on achievement tests when teachers' perception of the school environment is removed. |
| Talmage and Walberg (1978) | About 1600 elementary school students in grades 1, 2, 3, and 6; 59 classes as units | My Class, an elementary school version of the LEI with 5 scales as specified above, reliabilities from .54 to .77. | Science Research Associates Reading Test Total Scores | SRA alternate form pretest given one year earlier | Pretest-posttest r. of .87 raised to R of .93. |

Table 2 continued

| | | | | | |
|---|---|---|---|---|---|
| Tisher and Power (1975) | 315 junior high students in grade 9 from 20 classes in Australia. Student was the unit of analysis | Modified version of 15 LEI scales with internal consistencies from .53 .82. | Achievement in content areas, attitudinal measures, and satisfaction with teaching methods. | 7 selected pretests: 1) ecology achievement 2) population ecology 3) pollution 4) population 5) resources 6) interest 7) learning | Attitudes: Pre-test posttest r of .33 raised to .48, satisfaction (3 scales): .32 raised to .52, .20 raised to .47, .22 raised to .46 |
| Walberg and Anderson (1968) | 2,100 students in 76 high school physics classes in U.S; student as unit | 80-item, 18 factor analytically-derived scales on the Classroom Climate Questionnaire with internal consistencies ranging from .41 to .86. | 9 cognitive, affective, and behavioral posttests regression-adjusted for corresponding pretests ranging in internal consistency from .61 to .86. | None | 20 percent of the 162 intercorrelations significant at .05; achievement positively correlated with intimacy, negatively correlated with goal diversity and social heterogeneity; understanding negatively correlated with stratified; affect positively correlated with democratic, goal direction, formal, egalitarian, and satisfaction and negatively with stratified, friction, disorganized, social and interest heterogeneity. |

Table 2 continued

| | | | | |
|---|---|---|---|---|
| 3,700 students in national sample of 144 high school physics classes; class as unit | Learning Environment Inventory: Fourteen scales with internal consistencies from .58 to .86 for individuals and .43 to .84 for classes | Test on Understanding Science, internal consistences, .76; Welch Science Process Inventory, Physics Achievement Test, .77; Academic Interest Measure, .91; Pupil Activity Inventory, .80; Semantic Differential, .86; all given at the beginning and end of a one-year course. | IQ, pretest achievement and interest | On three cognitive criteria, median R with controls of .66 raised to .72; on three non-cognitive criteria, median R raised from .40 to .51; on achievement, .71 to .73. |
| Same as Walberg (1969a), but student as unit | Same as above | Selected pretests and posttests on understanding, achievement, interest, and activities. | Corresponding pretest | R with controls raised: from .68 to .69 for understanding, .73 to .75 for achievement, .70 to .70 (not significant) for physics interest, and .75 to .76 for voluntary physics activities |

Table 2 continued

| | | | | | |
|---|---|---|---|---|---|
| Walberg and Anderson (1972) | About 1600 students in 64 Montreal high school classes in 8 subjects; class as unit | The fifteen LEI scales. Reliabilities from .58 to .86. | Standardized Quebec High School Learning Examinations; internal consistencies range from .70 to .80. | IQ | In split-sample double cross-validations, r with controls raised to R with LEI and cross-validated: .42 to .87 and .67 in sample A and .24 to .78 and .43 in second sample. |
| Walberg, Singh, and Rasher (1977) | Somewhat less than 3,000 students in 300 studious and non-studious sub-groups of 10 students each in 83 science and 67 social studies classes in Rajasthan, India High Secondary schools; sub-group as unit | The fifteen LEI scales. Reliabilities from .58 to .86. | 100-item multiple-choice achievement tests geared to standard curriculum; general science internal consistency, .67; social studies, .81. | IQ | In general science, r of .63 raised to R. of .82; in social studies, .61 to .90. |

Table 2 continued



| | | | | | |
|---|---|---|---|---|---|
| Walberg (1969) | Same as Walberg (1969a). | Same as above | Same as above | Corresponding pretest | R of regression-residualized gain score with 14 LEI scales: .45 for cognitive criteria; .41 for non-cognitive criteria; .43 for achievement. |

Table 3

Descriptive Statistics for Correlations

Between Educational Outcomes and Learning Environments

| Factor | Mean Correlation | Standard Deviation | Frequency |
|---|---|---|---|
| Location* | | | |
| USA | .10 | .15 | 266 |
| Canada | .26 | .37 | 150 |
| Australia | .06 | .09 | 258 |
| India | .32 | .49 | 60 |
| Grade Level | | | |
| Elementary | .12 | .36 | 10 |
| Junior High | .11 | .13 | 11 |
| High School | .14 | .26 | 713 |
| Sample Size* | | | |
| 40-299 | .25 | .40 | 120 |
| 300-499 | .06 | .12 | 257 |
| 500-999 | .05 | .08 | 67 |
| 1,000-3703 | .18 | .27 | 290 |
| Unit of Analysis* | | | |
| Students | .07 | .10 | 325 |
| Subgroups | .29 | .46 | 71 |
| Classes | .17 | .29 | 333 |
| Schools | .30 | .31 | 5 |

Table 3 continued

| | | | |
|---|---|---|---|
| Type of Correlation | | | |
| Simple | .15 | .29 | 442 |
| Part | .11 | .13 | 11 |
| Partial | .12 | .20 | 281 |
| Learning Environment Scale* | | | |
| Cohesiveness | .23 | .27 | 50 |
| Friction[a] | .23 | .23 | 53 |
| Cliqueness[a] | .12 | .19 | 46 |
| Satisfaction | .22 | .21 | 54 |
| Speed[a] | .02 | .31 | 48 |
| Task Difficulty | .13 | .24 | 50 |
| Apathy[a] | .14 | .32 | 48 |
| Favoritism[a] | .16 | .16 | 46 |
| Formality | .06 | .26 | 57 |
| Goal Direction | .17 | .23 | 51 |
| Democracy | .17 | .24 | 50 |
| Disorganization[a] | .13 | .22 | 50 |
| Diversity | .02 | .73 | 47 |
| Environment | .18 | .26 | 49 |
| Competition | .06 | .38 | 35 |
| Outcome Domain* | | | |
| Cognitive | .17 | .33 | 403 |
| Attitudinal | .10 | .11 | 284 |
| Behavioral | .07 | .13 | 47 |

Table 3 continued

| | | | |
|---|---|---|---|
| Content Area* | | | |
| General Science | .12 | .25 | 133 |
| Life Sciences | .05 | .11 | 165 |
| Physical Sciences | .12 | .19 | 279 |
| Mathematics | .38 | .37 | 15 |
| Social Sciences | .34 | .50 | 60 |
| Humanities | .15 | .35 | 35 |
| General Achievement | .25 | .28 | 40 |
| Miscellaneous | .08 | .08 | 7 |

*Factors significant at the .001 level are indicated with an asterick.

[a]Signs of correlations reversed for these LEI scales.

Table 4
Conventional and Jackknifed Regression Statistics for the Two Model

| | Reduced Model | | | | Cross-Product Model | | | |
|---|---|---|---|---|---|---|---|---|
| | Conventional | | Jackknifed | | Conventional | | Jackknifed | |
| Variable | b-weight | t | b-weight | t | b-weight | t | b-weight | t |
| USA | -.20 | -6.94 | -.22 | -2.51 | -.23 | -8.72 | -.25 | -2.33 |
| Australia | -.16 | -4.34 | -.12 | -1.05 | -.18 | -5.26 | -.19 | -1.54 |
| Grade Level | .03 | 5.41 | .05 | 2.80 | .06 | 8.42 | .06 | 1.01 |
| Student as Unit | -.37 | -6.36 | -.82 | -2.22 | -.18 | -3.03 | -.28 | -1.61 |
| Subgroup as Unit | -.30 | -5.72 | -.76 | -2.09 | -.21 | -3.94 | -.42 | -1.98 |
| Classroom as Unit | -.39 | -8.95 | -.46 | -4.57 | -.26 | -5.41 | -.27 | -3.09 |
| Cohesiveness | .37 | 9.51 | .36 | 3.23 | .91 | 9.76 | .85 | 16.38 |
| Friction[a] | .37 | 9.40 | .37 | 2.64 | .62 | 6.40 | .42 | 1.19 |
| Cliqueness[a] | .20 | 3.26 | .17 | 1.72 | .12 | 2.18 | .08 | .77 |
| Satisfaction | .36 | 9.92 | .38 | 4.00 | .82 | 9.50 | .81 | 9.10 |
| Task Difficulty | .27 | 6.80 | .27 | 2.11 | .72 | 10.12 | 1.03 | 3.36 |
| Apathy[a] | .11 | 2.33 | .11 | .96 | .16 | 3.51 | .14 | 1.62 |
| Favoritism[a] | .23 | 3.86 | .21 | 2.46 | .15 | 2.86 | .12 | 1.24 |
| Formality | .16 | 4.24 | .12 | 1.39 | .01 | .20 | .04 | .60 |
| Goal Direction | .30 | 6.60 | .34 | 3.05 | .15 | 3.10 | .20 | 2.32 |

Table 4 continued

| Variable | b-weight | t | b-weight | t | b-weight | t | b-weight | t |
|---|---|---|---|---|---|---|---|---|
| Democracy | .30 | 5.93 | .32 | 3.04 | .14 | 2.79 | .17 | 2.18 |
| Disorganization[a] | .17 | 3.73 | .16 | 1.63 | .08 | 1.89 | .06 | .74 |
| Environment | 2.8 | 6.14 | .31 | 3.06 | .13 | 2.16 | .12 | 1.37 |
| Constant | .11 | | .42 | 1.19 | -.20 | | .05 | .18 |
| Standard Error | .247 | | .259 | | .214 | | .225 | |
| R | .571 | | .533 | | .709 | | .691 | |

[a]Signs of correlations for these LEI scales reversed in all analyses.

Note: For the jackknifed estimates, t-values of 2.26, 3.25, and 4.78 are significant at the .05, .01, and .001 levels, respectively. The jackknifed b-weights (and t-values) for cross-product terms are as follows: grade level by Cohesiveness -.03(3.53), Friction -.02(.64), Satisfaction -.05(3.78); Student as Unit of Analysis by Cohesiveness -.36(2.14), Task Difficulty -.98(2.99); Subgroup as Unit of Analysis by Task Difficulty -.30(3.20), Apathy -.99(12.26), Formality .39(6.66), Goal Direction .17(.84), Democracy .19(2.44), Environment .24(1.43); Class as Unit of Analysis by Cohesiveness -.33(5.95), Friction .41(1.75) and Task Difficulty -.94(3.01).

G 45

Table 5

Correlations of Learning Environment with Achievement at Two Grade Levels for U. S. Classes

| | Elementary | | | High School | | | |
|---|---|---|---|---|---|---|---|
| | Reduced Estimate | Product Estimate | Observed | Reduced Estimate | Product Estimate | Observed History | Observed Physics |
| LEI Scale | | | | | | | |
| Cohesiveness | .27 | .17 | .00 | .63 | .38 | .81 | .93 |
| Friction[a] | -.29 | -.52 | -.37 | -.65 | -.80 | -.90 | -.75 |
| Cliqueness[a] | | | | -.45 | -.27 | -.74 | .04 |
| Satisfaction | .30 | .38 | .11 | .66 | .45 | .63 | .63 |
| Task Difficulty | .19 | -.14 | -.27 | .35 | .28 | .26 | .53 |
| Apathy[a] | | | | -.39 | -.33 | -.86 | -.79 |
| Favoritism[a] | | | | -.50 | -.31 | -.59 | -.03 |
| Formality | | | | .39 | .23 | -.12 | .55 |
| Goal Direction | | | | .62 | .39 | .66 | .44 |
| Democracy | | | | .59 | .36 | .61 | .50 |
| Disorganization[a] | | | | -.44 | -.25 | -.71 | -.42 |
| Environment | | | | .59 | .55 | .80 | .88 |
| All others | -.08 | -.23 | -.51 | .28 | .19 | -.15 | -.57 |

Table 5 continued

---

[a]Signs of correlations for these LEI scales were reversed in all analyses. The original signs, however, have been restored in this table. Observed correlations were obtained from Talmage and Walberg (1978) and Walberg and Anderson (1972).

# A Methodology for Research Synthesis in Science Education

F. David Boulanger

Barbara K. Kremer

University of Illinois at Chicago Circle

Chicago, Illinois 60680

May, 1980

# A Methodology for Research Synthesis in Science Education

Over the past 50 years, science educators have periodically reviewed and organized the research literature on science learning. Most of the reviews were designed as comprehensive summarizations of the literature over a specified time period. Mallinson (1977) described these past efforts starting with A Digest of Investigations in the Teaching of Science in the Elementary and Secondary Schools (Curtis, 1926) through Mallinson's (1977) A Summary of Research in Science Education — 1975. Though valuable as comprehensive summaries, the past reviews are difficult to compare due to the absence of a common model or set of constructs defining the major categories of variables influencing science learning. This absence has meant that gaps in the research often go unnoted, since each reviewer develops a unique organization of material based on the trends and priorities of the period and the reviewer's point of view. Another limitation of these qualitative reviews is loss of the quantitative aspects of the accumulated studies. Advocating the quantitative synthesis of research, Light (1971) commented, "Little headway can be gained by pooling the words in the conclusions of a set of studies" (p.443).

The purpose of this paper is to present a model to guide reviews of research on science learning, a methodology for the quantitative synthesis of studies, and a summary of the results of an application of the model and methodology.

## A Model to Guide Reviews

A model to guide reviews of research on science learning should consist of a manageable number of constructs, reasonably comprehensive in explaining the observed variance in science learning. The set should correspond closely to past and present categories of research on factors affecting learning while allowing for the subsumption of new variables under the constructs. The set should provide for the inclusion of variables of immediate influence on learning, e. g., teacher reinforcement of student behavior, as well as variables representing important but less direct influence, e. g., parent education. It would be unrealistic, however, to expect the set to account for all predictable variance, given the multiplicity and complexity of factors that affect learning.

Ideally, a widely accepted theoretical model would provide the framework for empirical literature reviews. But such a model is not available, nor does there appear to be serious study of this problem within contemporary science education research (Peterson & Carlson, 1979, p. 506). In the absence of a theoretical model, a set of constructs might be identified through examination of general education empirical research. Variables found to be substantially related to learning could be organized into a manageable number of major construct categories.

Bloom's (1976) search and analysis of the research literature led to three major immediate factors influencing learning within the classroom: student affective entry characteristics, student cognitive entry behaviors, and the quality of instruction. This set of factors appears sufficiently comprehensive but does not make explicit how these constructs correspond to past and present categories of research nor how they incorporate out-of-classroom factors.

Walberg (1979) developed a larger though still manageable set of constructs which make explicit the subcategories implied in Bloom's set. Walberg's eight constructs are: student ability, motivation, and age or developmental level; the quality and quantity of instruction; and the home, peer, and classroom psychological environments. This list may be further revised and refined, but it provides for the major interrelated factors which the empirical literature would support as significant correlates of learning (Bloom, 1976, 1980; Comber & Keeves, 1973; Rosenshine, 1979). The list has the advantage of a close relationship with the major schools of empirical educational inquiry over the past three- quarters century, allowing the quantitative synthesis of many past studies to form estimates of the degree of association with or influence on learning by each construct.

Though certainly not the only possible choice, Walberg's eight constructs were adopted as the framework for this research synthesis with the hope that the results would support their routine use. That is, it was hypothesized that constructs important to learning in general would be important in science learning as well. If so, these constructs might routinely form the core of bivariate and multivariate studies in the future.

## A Methodology for Quantitative Synthesis

Several different approaches to the quantitative synthesis of research have been proposed in recent years. Light (1971) recommended a cluster approach wherein the data from studies of similar high quality and instrumentation are combined. Gage (1978) used a technique of converting the p-value (significance level) of comparisons of similar treatments on the same dependent variable into a form of the chi-square statistic allowing studies to be combined. Glass (1978) described a method for combining reported correlations or calculated effect sizes across related studies.

Each approach to quantitative synthesis has strengths and weaknesses. Light's ideal of using original data from closely comparable, high quality studies greatly limits the number of useful studies while increasing the time and effort requirements. Gage's use of p-values and Glass' use of correlations and effect sizes violates some assumptions of sampling and statistical comparability but provides estimates of effects and directions at a level of accuracy probably appropriate to the general quality of the data in the original studies, and at an effort level which makes the synthesis practical. Rosenthal (1978) has presented several variations on the above three approaches.

Regardless of the effort required and the precision of the methodology, the advantage of quantitative synthesis is the possibility of increased objectivity, precision, and conciseness in reporting quantitative outcomes and trends compared to a purely qualitative treatment. Objectivity is gained through the use of coding schemes which allow different raters to arrive at reliably similar characterizations of several features of a given study. Quantification also improves precision since features of a study (study-variables) can be coded at several gradations providing a finer discrimination among, and comparability across studies, than qualititative statements would allow. Finally, quantification gives a conciseness to the integration of a set of studie , yielding, where a sufficient number of studies is available, a regression equation predicting study outcomes under multiple-study conditions.

Quantitative synthesis should be veiwed, however, as supplementing rather than replacing, qualitative review. The quantification of any variable ultimately rests on qualitative descriptions of what the numbers represent and how the measurements were conducted. Research, as well as research synthesis, begins with the qualitative and moves toward quantification as greater objectivity, precision, and conciseness is sought.

Glass' approach to quantitative synthesis was chosen for this study since the information needed for coding a study can be extracted from the published study report; and correlations and effect sizes provide a measure of the strength of a relationship, not simply whether or not it is statistically significant. In fact, a strong argument can be made that effect size should replace alpha level as the most important outcome in experimental studies (Cohen & Hyman, 1980).

## A Model Guided Quantitative Synthesis

A quantitative synthesis of research in science education was conducted, guided by Walberg's eight constructs and using Glass's methodology. The remainder of this paper describes the adaptation of the methodology to this research synthesis.

The purpose of the synthesis was to develop sound approximations of the magnitude of the relationship between each construct and grade 6 through 12 student learning in science. Literature selection was restricted to the 1963 through 1978 period, a time of major curriculum reform and increase in the quality and quantity of research. The grade 6 through 12 levels were chosen to include the usual range of science course offerings in the precollege curriculum, beginning with required science in the junior high school and terminating with elective science in the senior high school. This age group is also characterized by the transition for many students from concrete to formal operational thinking (Inhelder and Piaget, 1958), an important research topic of the period.

Coding

The quantitative synthesis of research requires the development of a coding scheme. The coding scheme should summarize in numerical form the characteristics of the subjects, the setting, the independent and dependent measures, the research design and threats to validity, and the reported strength and direction of the relationship between the variables under study.

In this synthesis, the coding schemes for the eight constructs were identical with the exception that each construct had a special section for coding the independent variable. Aside from source information, the typical number of characteristics coded was approximately 40. A 90 percent agreement was readily attained between coders on a sampling of independently read studies.

To provide a sense of the degree of discrimination and detail in the coding scheme, three sections are briefly discussed here.

1) Dependent measure. Eight categories of dependent measures were chosen for coding based on a sampling of the research literature and a desire to be comprehensive. The categories were cognitive achievement, factual learning, conceptual learning, process learning, logical operations, creative or critical thinking, attitudes and interests, and lab performance. (Each kind of measure is operationally defined in the study code book.) Most often, the label given a measure by the author of the study was accepted as the proper classification, even though it is known that the many kinds of cognitive measures have similar items and a large amount of shared variance. Later analysis always required some combining of

these categories due to an insufficient number of studies with measures in any one category.

Four other characteristics of the dependent measure were coded: the type of measure (general, science, specific discipline, specific course); whether a locally constructed or a published instrument; the reliability; and a judgment about the validity (adequate or inadequate).

2) Study design characteristics and threats to validity. Glass (1978) proposed the coding of individual study design and analysis features which might have influenced study results. Once coded, the covariance of these study-variables with study findings could be examined, making full use of statistical methods. In the present synthesis, various aspects of each study's research design were coded. These design factors included the threats to experimental validity identified by Cook and Campbell (1976), and are summarized in Table 1.

3) Quantitative relationships. The value, sign and level of significance of each reported correlation was recorded. Where an experiment or quasi-experiment was involved, the effect size and direction and the level of significance of the statistical test were recorded.

Nearly all effect sizes were computed using one of the following two formulas from Glass:

$$ES = \frac{\overline{Xe} - \overline{Xc}}{Sc} \qquad ES = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\overline{Xe}$ and $\overline{Xc}$ represent experimental and control group means respectively. Sc is the standard deviation of the control group. t is the computed t-test statistic. If an F-test were used in a one-way analysis of variance to compare two groups, the F value was considered equal to $t^2$.

If only the total sample size was given it was assumed that $n_1 = n_2$, since equal n's provide a more conservative estimate of effect size than the unequal n's. Finally, in cases where one-way analysis of variance was used, homogeneity of variances was assumed setting Sc = MSw. Two-way analysis of variance tables, however, without other statistics, were insufficient for computing effect sizes.

Code sheets were devised for recording all above and additional information about each study. Several code sheets were often necessary for a single study depending on the number of effect sizes computed, correlations reported, and whether data were reported separately for different study-variables such as grade or ability level.

Analysis

Once coding is completed, the analysis of the coded data can take a variety of forms depending on the quantity of data and the researchers inclination toward the liberal application of statistical techniques. The first step is to decide how to deal with the problem of the non-independence of multiple correlations or effect sizes extracted from the same study. One solution is weighting each correlation or effect size inversely to the number of each extracted from a given study. A study contributing 10 correlations receives the same total weight as a study contributing three. A second solution, the one used in this study, is to select only the median value correlation or effect size from each study. This procedure will greatly reduce the data set, but will equalize the contribution of each study.

Many more questions can be raised about the appropriate weighting of studies. For example, should a study with an n of 30 receive the same weight as one with an n of 2000? Should a high quality true experiment receive the same weight as a low quality quasi-experiment? Answers to questions like these might be framed as hypotheses prior to analysis; an empirical answer is available in the analyses itself. Correlations from large sample studies can be compared to correlations from smaller sample studies; true experiment studies can be compared to quasi-experiment studies, and so on for any study-variable the researcher wishes to code.

Once the weighting problem has been resolved (whether to use weighted or median correlations and effect sizes), analysis involves grouping code sheets of studies with the same independent and dependent variables and associated statistic i.e., correlation or effect size; treating the statistic as a dependent variable and relating the values of this dependent variable to different study-variable conditions. Questions like the following can be addressed: What is the average correlation or effect size across studies? Does the reported correlation (dependent variable) vary in a systematic way with sample size, outcome measure, reliability, or the ability level of the subject sample? Or, is the mean correlation fairly constant across variations in study-variables. Since several independent study-variables and a single dependent variable (correlation of effect size) have been quantified, t-tests, F-tests, correlations and regressions can be conducted to characterize the relationship among the variables. Table 2 summarizes the literature selection and central tendencies of the correlations or effect sizes in this study. References to separate, more detailed reports of each synthesis are given in the Table.

The extent of the analysis conducted on the data in each construct was a function of the number of data points available. The 734 correlations and related set of study-variables in the classroom environment construct allowed extensive regression procedures wherein the contribution of each study-variable to variation in the set of correlations could be estimated. Analysis of the impact of study-variables on the 67 ability construct correlations took the form of a series of t-tests wherein each study-variable was dicotomized into high and low or two nominal categories, each containing approximately the same number of median correlations.

A study-variable of particular interest in the quality of instruction construct studies was design quality. A design quality index representing a summation of positive design features (features minimizing threats to validity, i. e., Table 1) was significantly ($p = .09$) corelated ($r = .21$) with effect size. That is, the better the study design, the greater the difference between experimental and control group means as measured by effect size. Better study design meant a greater effect favorable to the experimental group.

A study-variable which systematically influenced the ability and cognitive achievement correlation was the reliability of the two instruments. For example, the reliability of the outcome instrument was correlated with the correlation between ability and cognitive achievement at an $r$ value of .33.

## Summary

The purpose of this paper was to provide an overview of a model and methodology for quantitative research synthesis in science education and a summary of their application.

The model included a comprehensive set of constructs to guide the literature review and to help identify important groups of variables not receiving sufficient research attention. The quantitative methodology added objectivity, precision, and conciseness to the traditional quantitative review. Quantification allowed the reporting of mean correlations and effect sizes and examples of how these statistics vary with study-variables such as design quality and instrument reliability. Detailed reports were identified which the interested researcher might consult for an in-depth treatment of the quantitative integration of research summarized here.

References

Bloom B. S. Human characteristics and school learning. New York: McGraw Hill, 1976.

Bloom B. S. The new direction in educational research: Alterable variables. Phi Delta Kappan, 1980, 61, 382-385.

Boulanger, F. D. Ability and science learning: A quantitative syntheses. Journal of Research in Science Teaching, 1980a, in press.

Boulanger, F. D. Instruction and science learning: A quantitative synthesis. Journal of Research in Science Teaching, 1980b, in press.

Boulanger F. D. & Kremer, B. K. Age and development level as antecedents to science learning. Manuscript submitted for publication, 1980.

Comber, L. C. & Keeves, J. P. Science education in nineteen countries: An empirical study. New York: John Wiley, 1973.

Cook, T. D. and Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunniette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Curtis, F. D. A digest of investigations in the teaching of science in the elementary and secondary schools. Philadelphia: P. Blakison's Son and Co., Inc., 1926.

Frederick, W. C. & Walberg, H. J. Learning as a function of time. Journal of Educational Research, 1980, in press.

Gage, N. L. The scientific basis of the art of teaching. New York: Teachers College Press, 1978.

Glass, G. V. Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), Review of Research in Education (vol. 6). Itasca, Il.: F. E. Peacock, 1978.

Haertel, G. D., Walberg H. J., & Haertel, E. H. Social-Psychological Environments and Learning: A Quantitative Synthesis. A paper presented at the 1979 Annual Meeting of the American Educational Research Association, San Francisco, Cal.

Kremer, B .K. & Walberg, H J. A synthesis of social and psychological influences on science learning. Science Education, 1980, in press.

Light, R. J., & Smith, D. V. Accumulating Evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Mallinson, G. G. A Summary of research in science education - 1975. New York: John Wiley, 1977.

Peterson, R. W. & Carlson, G. R. A summary of research in science education-1977. Science Education, 1979, 4, 421-553.

Rosenshine, B.V. Content, time, and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), Research on Teaching: Concepts, Findings, and Implication. Berkeley: McCutchan, 1979.

Uguroglu, M. E. and Walberg, H. J. Motivation and achievement: A quantitative synthesis. American Educational Research Journal, 1979, 16, 375-389.

Walberg, H J. Psychological theory of educational productivity. Invited address presented at the annual meeting of the American Psychological Association, Toronto, Ontario, Canada, 1978.

Table 1

Code Categories for
Study Design Characteristics

Sample selection
1 = simple or stratified random
2 = purposive sample, e. g., extreme or specialized group
3 = matching
4 = convenience or ill-specified sample

Unit of analysis
1 = individual
2 = group

Study design (Campbell & Stanley)
1 = correlational
2 = quasi-experimental
3 = true experimental (random assignment)

Reliability of treatment implementation
1 = low; treatment and implementation poorly described and documented
2 = adequate; treatment and implementation clearly described and documented
3 = high; treatment described or documented with observational checks on implementation

Statistical power
1 = inadequate
2 = adequate, i. e., 6 or more classes; 100 or more individuals total in 2 comparison groups or in correlational group

Error rate (Given the number of comparisons or correlations, is the overall p level sufficiently low to assure a less than .05 chance occurrence of this particular relationship?)
1 = inadequate p level
2 = adequate, i. e., p less than .05

Maturation (Have factors within units rather than the treatment brought about the difference observed?)
1 = probable threat
2 = adequately minimized

History (Have external factors in the environment rather than the treatment brought about the differences observed?)
1 = probable threat
2 = adequately minimized

Selection Bias (Do pre-existing differences among the groups account for later observed differences?)
1 = probable threat
2 = adequately minimzed

Contamination, Compensation, Differential Incentive (Do untreated control groups work harder, work less, or somehow gain benefits or lose incentive due to influences from treated groups or teachers?)
1 = probable threat
2 = adequately minimized

Mortality (Do different dropout rates account for observed differences?)
1 = probable threat
2 = adequately minimized

Generalizability (Can results be generalized to other times, units, or settings with similar demographic characteristics?)
1 = probable threat
2 = adequately minimized

Table 2

Summary of Sources and Findings on the Relationship of Each Construct to Cognitive Learning in Science

| t | Measures | Detailed Report | Number of Studies | Sources Searched[a] | Data Points: Number & Kind | Mean Corr. or Effect Size | Comment |
|---|---|---|---|---|---|---|---|
| | Chronological age (within grade) | Boulanger & Kremer, 1980 | 3 | JRST, SE, DA, ERIC | 3 median corr. | $\bar{r}$ = - .01 | Age is a sign. pos. predictor only if multi-grade level data. |
| ental | Piagetian stage or logical operations (within grade) | " | 6 | " | 13 median corr. | $\bar{r}$ = .40 | This correlation peaked at .69 in grade 9. |
| | IQ, gen. aptitude, prior achieve., quant. - spacial | Boulanger (1980a) | 34 | JRST, SE, SSM, ERIC | 62 median corr. | $\bar{r}$ = .48 | This construct gave the highest and most reliable mean corr. |
| ion | Self-concept, need-achiev., persistence test anxiety | Kremer & Walberg, 1980 | 5 | JRST, SE, SSM, DA, ERIC, SSCI | 5 median corr. | $\bar{r}$ = .37 | Higher corr. were obtained with standardized over locally made scales. |
| iron- | Parent SES; science equip. and documents in home; parent involvement, education | " | 13 | " | 6 median corr. | $\bar{r}$ = .24 | Parent education and aspiration and involvement with child's science best predictors (r = .36) |
| iron- | Rapport with peers, peer partic. in science | " | 5 | " | 2 median corr. | $\bar{r}$ = .25 | Studies too few and too diverse to note trend in best predictors. |
| of tion | Use of adv., organizers, beh., obj., concrete materials, higher structure, indirect and inductive strategies, training in thinking | Boulanger (1980b) | 51 | JRST, SE SSM, ERIC (pub. only) | 52 median effect sizes | $\overline{ES}$ = .55; equivalent r = .25 - .30 | Significant effects due to: behav.obj., concrete materials, higher structure, training in thinking. |
| y of tion | Class periods spent on teaching the content | " | 3 | JRST, SE, SSM, ERIC | 3 corr. | $\bar{r}$ = .19 | None of three studies gave sign. corr. |
| om ment | Student perception of several classroom Social variables | Haertel, Walberg & Haertel, 1979 | 12 | ERIC, DA, EI, DA, SSCI | 7 mean corr. based on 353 raw corr. | $\bar{r}$ = .19 | Sign. predictors were: cohesiveness, low friction, satisfaction, low favoritism, goal direction, democracy, and material environment. |

ym code: JRST, Journal of Research in Science Teaching; SE, Science Education; School Science and Mathematics; DA, Dissertation Abstracts; ERIC, Educational urce Information Center (Science) bibliographies and computer search; SSCI, al Science Citation Index; EI, Education Index. Pub. only means only published studies included.

# Toward A Synthesis of Research Findings on Sex Differences in Science Learning

Barbara K. Kremer

University of Illinois at Chicago Circle

Running head: Sex Differences and Science Learning

Toward A Synthesis of Research Findings
on Sex Differences in Science Learning

That far fewer women than men pursue careers in mathematics and science and reputedly demonstrate far lower scores on tests of aptitude and achievement in these areas has, until recently, been accepted as a natural consequence of innate sex differences in aptitudes for those fields. Thorndike (1973) noted that in none of the countries he surveyed (including the U.S.A), did girls do as well as boys in science. Differences were observed on the order of half a standard deviation.

Investigations as recent as those of Stafford (1972), and Page (1976) have suggested sex-linked, hereditary hypotheses in explaining differences in male-female achievement, particularly in math. Reviewers of sex difference research have, however, indicated the inconclusiveness of findings favoring biological explanations of sex-related cognitive differences (Maccoby & Jacklin, 1974; Fennema, 1977; Sherman, 1977; Vandenberg & Kuse, 1979). An alternative hypothesis rivalling alleged biological influences, stresses the importance of sociocultural factors on male-female performance.

Some have concluded that social values and behaviors more often shown by girls tend to be those not associated with the successful development of intellectual achievement (Maccoby & Jacklin, 1974; Fox, 1977). Nash (1979) discusses evidence that individuals' gender-identity influences their motivation to successfully perform on cognitive tasks that have been sex-typed as male or female. Recent work by Kremer and Walberg (1979) strongly suggests the importance of social psychological factors such as home environment, and student motivation in science learning and achievement. Fennema and Sherman (1977) have demonstrated that much of the difference in mathematics achievement between males and females is attributable to course-taking history. These researchers also found that mathematics is viewed as a male domain by many girls, thus resulting in their avoidance of math (Sherman & Fennema 1977).

While considerable attention has recently been devoted to sex differences in mathematics achievement, relatively little attention has been given to this topic in science. Concerning sex differences and women's achievement in science, the questions are still largely those of "if, when, and where" sex differences exist in performance in this broad intellectual area. While the percentage of doctorates in the sciences currently being awarded to women is increasing and has once again reached the level of the 1920's , there have been significant shifts in this trend over intervening decades, and women continue to be under-represented (Vetter, 1978). The uniqueness of the perspective women bring to the sciences, and their potential, high productivity in science has been noted (Astin, 1978). Much stands to be lost, in the face of lagging industrial and technological productivity if this potential is not developed.

Women's participation in science is dependent upon the quality and the effectiveness of the science education they receive. More needs to be known about women's achievement in science, and the social, psychological, and even biological factors that influence it. If one assumes that important cognitive differences exist between the sexes, then understanding the nature and extent of these differences is important for determining what type of intervention, if any, would be most effective. An objective base for examining issue related to sex differences in science achievement is needed if recommendations for educational policy are to be made, and promising directions for research in this important area are to be identified.

## Toward an Objective Base

A review of related literature has revealed the following issues that need be addressed in the development of an objective base for identifying research priorities in sex differences and science learning, and drawing implications for educational policy. These are: male-female achievement in different domains of science learning; male-female achievement in science when cognitive, instructional, and attitudinal factors are controlled; age-related trends in male-female science achievement; and the variance of observed differences in male-female achievement with the chronological time of investigation.

### Domains of Science Learning

In what domains of science learning (for example, factual learning, scientific processes, attitudes to science) do male-female differences occur? Research in math education has shown differential achievement for males and females in some areas of mathematics learning, but not in others. Could the same be true for science? If so, can differential achievement in science learning best be understood in terms of biological and genetic hypotheses, or by social psychological explanations?

Cognitive, Instructional, & Attitudinal Constrols

If cognitive, instructional, and/or attitudinal factors are controlled, are observed sex-related differences in science learning accounted for? Fennema and Sherman (1977) have shown, for example, that studies of particular abilities such as mathematics will be biased if course-taking is not similar for males and females in the study samples. Moreover, studies based on large random samples of secondary school students, may be comparing a more heterogenous group of females with a more homogeneous, intellectually motivated group of males if males are more likely than females to drop out of school.

Age-related Trends

Are there significant, age-related trends in sex differences observed in science learning? Terman and Tyler (1954) report evidence for increasing sex differentiation with age in the areas of abilities, interests, preferences, and responses to personality inventories. Petersen and Wittig (1979) have suggested that observed differences between the sexes increase with age as socialization effects accumulate, and that puberty, is likely to be a critical time for the intensification of socialization effects. Conclusions about the existence of sex differences, then, may well depend upon the age mix in the best of studies in which sex differences are examined.

Chronological Time of Investigation

Does the frequency and m gnitude of reported sex-related differences in science learning appear to vary with the chronological time of investigation? Are reported sex differences in performance in science of greater magnitude in older, or more recent studies?

Maccoby and Jacklin (1974) noted: "As sex role behaviors have been less rigidly defined and enforced, sex-related differences have decreased and in many instances they have not been demonstrated." One might therefore expect to see sex-related differences of a much smaller magnitude or lesser frequency in more recent, as opposed to older studies.

## A Methodology

Probably the best known work on the psychology of sex differences is that of Maccoby and Jacklin. Among the most noteworthy contributions of this extensive work is their systematic, and analytic synthesis of research. Often, what is considered "truth" is shown upon closer analysis to be based on inadequate reporting, or the failure of researchers to control for significant variables. Where popular beliefs are supported, new complexities are often revealed. A synthesis of research findings on sex differences in science learning employing quantitative meta-analysis techniques (Glass, 1978).

### Rationale

Numerous studies currently exist which report the results of male-female comparisons on measures of science learning and achievement. A recent literature search revealed over 150 studies of science learning reporting cross-sex comparisons. An integration of research reporting gender differences on measures of science learning is needed at this time for a number of reasons.

The purpose of research synthesis is to determine what existing research proves about the relationship of one variable, or class of variables to another; in this case, sex differences in science learning. While extensive reveiws of sex differences in cognition have been conducted (Wittig & Petersen, 1979);

it is often difficult to relate basic psychological research to educational policy, even studies of cognitive processes basic to science learning. While national surveys of science achievement based on random samples have been conducted, and report results by gender, the policy implications to be drawn from these studies are sometimes limited. They are limited by the lack of control of important variables mediating science achievement (e.g., previous instruction and attitude), and the exclusive focus on one (or possibly two) outcomes. Even the application of secondary analysis techniques to these data bases cannot go beyond such limitations. What is needed therefore, is a systematic integration of science education research incorporating findings across numerous studies representing diverse samples, with outcome measures reflecting several domains of science learning.

Extensive syntheses of science education research have been conducted (Walberg, Boulanger, Kremer, & Haertel, 1980). However, much of the research reviewed under the social, instructional, and ability constructs defined in the guiding model, does not routinely report male-female comparisons. Furthermore, numerous investigations reporting sex differences were not included in these reviews as they did not fall within the boundaries established for the selection of literature. A synthesis of findings on sex differences in science education research is therefore needed to address the concerns of science educators and policy-makers in this important area.

Methods of Research Synthesis

Syntheses of empirical research generally employ one, or a combination of methods: narrative reviews of the literature, box scores or tallies of significant findings (Light and Smith, 1971), and quantitative, statistical techniques as exemplified by meta-analysis (Glass, 1978) and the joint probability method (Rosenthal, 1978). The use of box scores for integrating research findings typically involves determining for each study, whether or not a statistically significant difference was found; and if so, its direction (i.e., whether the treatment or control group was favored). Studies are then tallied according to whether significant differences are reported, and the direction of significant findings is noted. In the field of sex differences research, the work of Maccoby and Jacklin best exemplifies the use of box scores for research integration.

The joint probability method of research integration involves combining, or pooling the exact one-tailed probabilities of each comparison reported. Methods for combining probabilities are discussed, and illustrated by Jones and Fiske (1953), and by Rosenthal (1978). A still more powerful quantitative method however, is the meta-analysis technique developed by Glass (1978).

Meta-analysis is based upon the derivation of an effect-size representing a normalized measure of the difference between two comparison groups on a measured outcome. The effect size expresses the magnitude of group differences on a common scale, so that findings from studies employing different measures, and methods, are rendered comparable. In contrast to methods employing box scores, and combined probabilities, the effect size calculated in a meta-analysis has the advantage of providing an estimate of the over-all size of an effect, in addition to its significance.

A Case for the Meta-Analysis of Sex Differences

Quantitative methods for synthesizing research, entail more objective processes for summarizing individual studies, and allow for more concise means of displaying and interpreting results than more conventional reviews. Meta-analysis in particular, allows for the application of a broad range of analysis techniques, from frequency distributions to multivariate methods. Moreover, meta-analysis permits the integration of primary studies representing diverse samples and outcome categories. This is a distinct advantage in the formulation of well-rounded policy statements, and the definition of research priorities.

The meta-analysis of research represents a particularly appropriate technique for synthesizing studies on sex differences, since it results in a statistical statement about the magnitude of differences between groups. This is especially important since previous syntheses of research on sex differences have been criticized for overemphasis on null hypotheses, and failure to note the magnitude of differences (Block, 1975). Jacklin (1979) has recently called upon researchers to go beyond the necessary first step of sorting findings by statistical significance (as exemplified by the work of Maccoby & Jacklin, 1974), and employ techniques estimating the size of observed differences.

## Conclusion

A quantitative synthesis of research findings on sex differences delineating both the frequency, and magnitude of observed differences in science learning is warranted. It is warranted on the basis that important issues concerning the participation, and education of women in the sciences have yet to be resolved, and appropriate methodologies for the synthesis of research have been currently developed. Of concern are questions regarding male-female achievement in different domains of science learning, particularly when critical instructional and attitudinal factors are controlled, age-related trends in male-female science achievement, and the observed variance of male-female achievement with the chronological time of the investigation. The systematic integration of existing research constitutes a necessary first step in the formulation of meaningful hypotheses for further study, and directing the concerns of policy-makers.

References

Astin, H. Factors affecting women's scholarly productivity. In H. Astin and W. Hirsch (Eds.), The higher education of women: Essays in honor of Rosemary Park. N. Y.: Praeger Publications, 1978.

Block, J. Issues, problems, and pitfalls in assessing sex differences: A critical review of the Psychology of Sex Differences. Merrill-Palmer Quarterly, 1976, 22, 285-308.

Fennema, E. Influence of selected cognitive, affective, and educational variables on sex-related differences in mathematics leraning and studying. In J. Shoemaker (Ed.), Women and mathematics: Research perspectives for change, (N.I.E. Papers in Education and Work: No. 8). Washington, D. C.: Education and Work Group, The National Institute of Education, U. S. Department of Health, Education, and Welfare, 1977.

Fennema, E., & Sherman, J. Sex-related differences in mathematics achievement, spatial visualization, and affective factors. American Educational Research Journal, 1977, 14, 51-71.

Fox, L. The effects of sex role socialization on mathematics participation and achievement. In J. Shoemaker (Ed.), Women and mathematics: Research perspectives for change, (N.I.E. Papers in Education and Work: No. 8). Washington, D. C.: Education and Work Group, The National Institute of Education, U.S. Department of Health, Education and Welfare, 1977.

Glass, G. Integrating findings: The meta-analysis of research. In L. Shulman (Ed.), Review of research in education, Vol. 6. Itasca, Illinois: F. E. Peacock, 1978.

Jones, L. V., & Fiske, D. W. Models for testing the significance of combined results Psychological Bulletin, 1953, 50, 375-382.

Kremer, B., & Walberg, H. J. A synthesis of social and psychological influences on science learning. Science Education, accepted.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Maccoby, E., & Jacklin, C. The psychology of sex differences. Stanford, CA: The Stanford University Press, 1974.

Nash, S. C. Sex role as a mediator of intellectual functioning. In M. A. Wittig & A. C. Petersen (Eds.), Sex-related differences in cognitive functioning: Developmental issues. N. Y.: Academic Press, 1979.

Page, E. A historical step beyond Terman. In D. Keating (Ed.), Intellectual talent: Research and development. Baltimore, MD: The Johns Hopkins University Press, 1976.

Rosenthal, R. Combining results of independent studies. Psychological Bulletin 1978, 85, 185-193.

Sherman, J. Effects of biological factors in sex-related differences in mathematics achievement. In J. Shoemaker (ed.), Women and mathematics: Research perspectives for change, (N.I.E. papers in Education and Work: No. 8). Washington, D.C.: Eduation and Work Group, The National Institute of Education, U. S. Department of Health, Education, and Welfare, 1977.

Sherman, J. & Fennema, C. The study of mathematics by high school girls and boys: Related variables. American Educational Research Journal, 1977, 14, 159-168.

Stafford, R. Hereditary and environemntal components of quantitative reasoning. Review of Education Research, 1972, 42, 183-201.

Thorndike, R. The problems of empirical comparative education. In W. Coffman (Ed.), Frontiers of educational measurement and information systems. Boston: Houghton-Mifflin Co., 1973.

Vandenberg, S. G., & Kuse, A, R. Spatial ability: A critical review of the sex-linked major gene hypothesis. In M. A. Wittig & A. C. Petersen (Eds.), Sex-related differences in cognitive functioning: Developmental issues. N. Y.: Academic Press, 1979.

Vetter, B. Data on women in scientific research. Paper presented at the meeting of the American Association for the Advancement of Science, Washington, D. C., February, 1978.

Walberg, H. J., Boulanger, F. D., Kremer, B. K., & Haertel, G. D. A meta-analysis of productive factors in science learning in grades 6 through 12. In preparation.

Wittig, M. A., & Petersen, A. C. Sex-related differences in cognitive functioning: Developmental issues. N. Y.: Academic Press, 1979.

# Science Curriculum Effects in High School: A Quantitative Synthesis

Thomas Weinstein, F. David Boulanger, and Herbert J. Walberg

University of Illinois at Chicago Circle

Running Head: Curriculum Effects

Abstract

To assess the impact of the innovative pre-college science curricula of the past 20 years on achievement, a search was conducted using the computer-assisted Bibliographic Retrieval System(BRS), the ERIC Annual Summaries of Research in Science Education and Dissertation Abstracts International. A total of 197 comparison-effect sizes were obtained from 33 studies representing 19,149 junior and senior high school students in the United States, Great Britain, and Israel. Study-weighted analysis yielded an overall mean effect size of .308 significantly favorable to the innovative curricula ($t(25) = 2.183$, $p<.05$). Student performance in innovative curricula averages at the 62nd percentile relative to the control norm. Tabulation of signed comparisons indicated that 64 out of 81 unweighted outcomes were favorable to the innovative curricula. Separate analyses for test content bias, methodological rigor, type of learning, and student characteristics showed no significant differences across these categories.

Beginning in the late 1950's[1] and continuing to the present day the American taxpayer has supported scientists and educators in pre-college curriculum development. Creation of innovative courses in science for grades 7 through 12 received special attention in the 1959-1973 period, accounting for approximately two-thirds of the National Science Foundation's (NSF) curriculum development expenditures for that period (NSF, 1975). NSF allocated $92 million to development and implementation of 19 such projects. Associated with the effort were extensive teacher-training summer and year-long institutes in the use of the new programs and numerous evaluation studies of the classroom impact of the innovative programs, usually in comparison to "traditional" curricula.

A debate about the effectiveness of the new programs began with the first implementations in the early 1960's especially the "new math". By the mid-70's sufficient studies had accumulated that a summative judgement appeared possible. Walker and Schaffarzick (1974) conducted a partial search of the literature and located 26 studies which compared students exposed to different curricula in the same subject on some measure of school achievement. Using statistical significance as the criterion for counting a comparison, they reported a general trend supporting the hypothesis that a treatment, whether innovative or traditional, yields the higher score on tests biased in its favor. On tests biased toward the innovative program, the innovative group performed better on 44 out of 45 significant comparisons. On tests biased toward the traditional treatment, the traditional group performed better on 9 out of 14 comparisons. Where test bias was neutral or could not be determined, both treatments were equally effective. Unfortunately, out of 98 comparisons, 32 were not considered in the count since significant differences were not found by the original investigators.

Walker and Schaffarzick acknowledged both the limited extent of their search and the inadequacy of statistical significance as a criterion for counting comparisons. Objective quantitative techniques for synthesizing research, moreover, that weight studies equally and also compare the effects in studies categorized by validity, subject matter, and other characteristics were not widely known in education.

Cohen and Hyman (1979) addressed this overreliance on alpha for decision making in experimental studies and argued that effect sizes should be emphasized instead. Use of effect size in planning and reporting studies would also aid research synthesis by providing a measure of the size of the difference between groups, quite apart from its statistical significance.

Glass (1978) has described a methodology for the use of effect size in research synthesis. Glass' technique can provide better estimates of effects than simple counts of comparisons to ascertain the cumulative meaning of a body of research such as curriculum-evaluation studies.

Bredderman (1978) used Glass' approach to quantitatively synthesize the results of over 60 evaluations of nationally developed elementary school (K-6) science curricula. Using this technique, Bredderman was able to estimate average effect sizes for each kind of program compared to traditional treatment controls, and the degree of impact of the programs on different kinds of students. One of Bredderman's major findings was that curricular effects on student achievement were in harmony with curricular objectives and content, a finding basically in agreement with the less well founded conclusion of Walker and Schaffarzick.

Welch (1979), drawing on the results of 82 studies, counted studies supporting certain generalizations about the effects of the innovative curricula on K-12 students. Welch's rough count of studies also gave results essentially in agreement with Walker and Schaffarzick.

The purpose of the present study was to apply Glass' technique in the quantitative synthesis of the secondary school, grades 7 through 12, curriculum evaluation studies of the 1963-1978 period. Only the nationally developed innovative program evaluations would be included. A thorough literature search combined with quantitative techniques that include all treatment comparisons would provide a new and critical look at the Walker and Schaffarzick conclusion.

Among the techniques employed that permit more confident conclusions than previous comparison counts are estimates of the size and significance of the average effect size, and the dependence of the effect size, which in this case, represents a comparison of performance under innovative and traditional curricula, on the subject matter, grade level, type of outcome measure, and methodological qualities of the study such as experimental and instrumentation validity. In addition, "vote counts" of positive and negative contrasts of the two types of curricula are reported to afford comparisons with the effect size method as well as with the results of previous syntheses.

## Method

### Search and Selection

Studies were identified through a search of the computer-assisted Bibliographic Retrieval System (BRS), which provides access to Dissertation Abstracts International, and the ERIC database of published and unpublished articles. The ERIC annual summaries of science education research were also consulted. The BRS search was conducted using the descriptors "curriculum development," "innovative/innovation," "science education in courses," "secondary school science/ elementary school science," and "1963-1978." A manual scan of the two major journals Science Education and Journal of Research in Science Teaching for the years 1963-1978 supplemented the ERIC summaries. Studies that quantitatively compared traditional and nationally developed innovative science curricula on student learning outcomes, grades 6 through 12, were selected.

Thirty-three studies representing 19,149 students in the United States, Great Britain, and Israel were chosen for investigation. Among these, 13 curricula are included, eight at the senior-high school level and five at the junior-high school level. To investigate the Walker and Schaffarzick hypothesis, 9 of 23 of their sources were included in this investigation. Those omitted did not meet either subject matter (science) or grade level (6 or higher) criteria or both.

Coding

To allow quantitative synthesis of both study characteristics and outcomes, the following study-variables were coded for each comparison: study origin and source; subjects and setting, i.e. grade level, gender, ethnicity, academic achievement level, community SES and urban-rural character; subject matter of treatment; treatment characteristics such as group size, elective or required course participation, regular or special teacher, lab or non-lab focus, reliability of implementation, length and quality of control group access to content; study design and nine categories of threats to validity; sample size; dependent measure type, origin, reliability and innovative, traditional or neutral bias including an indication of the source of information on bias, i.e., author description or independent inspection of the test; and outcome statistics, i.e., direction of effect, level of significance and effect size.

The coded threats to validity were reliability of treatment; statistical power; error rate; maturation; history; selection bias; contamination; compensation or differential incentive; mortality; and generalizability. They were categorized either 1) potential threat or 2) adequately minimized. An overall index of design quality was taken from the sum of these ratings.

The five dependent outcome categories were: 1) conceptual learning, e.g., Concept Attainment Test (Cunningham, 1970); Taxonomy Test (Herron, 1966) based on the comprehension, analysis and application levels of Bloom's (1956) Taxonomy; and standardized achievement tests; 2) inquiry skills, e.g. tests of controlling variables, formulating hypotheses, critical thinking, and logical operations; 3) attitudinal development, e.g. any measure of attitude, interest, or opinion toward science or science related concerns; 4) laboratory performance, including observation, investigation and manipulative skills with actual appratus; and 5) concrete skills,i.e. classification of properties represented by pictorial stimuli. Unlike inquiry skills, concrete skills require only observation and classification of directly perceived objects or pictures. Inquiry skills require some form of hypothetical-deductive reasoning as in Piagetian formal operations.

In coding the dependent measure, it was usually not difficult to distinguish between tests containing content favorable to the traditional versus the innovative programs. The Test on Understanding Science (TOUS), however, while clearly not traditional in content, differs markedly from tests designed by most investigators of the new curricula. Such tests require the student to apply the inquiry skills gained in the innovative program, e.g.,a transparency test in BSCS biology (Mascolo, 1969). TOUS, while designed to measure knowledge of the scientific point of view, does not require that inquiry skills be applied while taking the test. Consequently, although it is a non-traditional test, TOUS was grouped separately and coded neutral.

The formula for calculating effect size was almost always one of the following (Glass, 1978):

$$ES = \frac{\bar{x}_e - \bar{x}_c}{s_c} \quad \text{or} \quad ES = t\sqrt{\frac{1}{n_e} + \frac{1}{n_c}}$$

where $\bar{x}_e$ and $\bar{x}_c$ represent experimental and control group means respectively; and $s_c$ is the standard deviation of the control group. Where applicable, t is computed from the t-test statistic. When F was the result of a two group comparison, t was considered equal to $\sqrt{F}$. In cases of one way analysis of variance, homogeneity of variance was assumed, setting $s_c = \sqrt{MSw}$. All effect sizes favoring the innovative curricula were given a positive sign, those favoring the traditional curricula a negative sign.

Weighting Procedure

The number of effect sizes computed from each study varied as a result of both the quantity of comparisons and the quality of supplied data. Some studies presented means and standard deviations for each of several test categories. Other studies failed to give one or both of these statistics. In consequence, the number of effect sizes per study ranged from one to 33. To give equal weight to each study rather than to each comparison, each effect size was assigned a weight equal to the reciprocal of the number of effect sizes in its study. Each of the 33 effect sizes from one study received a weight of 1/33. This procedure weights each study equally and yields a smaller number of independent degrees of freedom than a count of unweighted comparisons which are not statistically independent. Unweighted signs were used in one instance, however, to allow comparison with Walker and Schaffarzick's data.

Data Analysis

A visual analysis of the set of effect sizes was made by plotting a stem-and-leaf diagram, displayed in Table 1.

Insert Table 1 about here.

To obtain an overall measure of the impact of the innovative versus the traditional curricula, a mean effect size was computed. To check for a systematic influence of any coded study-variable on effect sizes, a one way analysis of variance was conducted on each study-variable, with effect size as the dependent variable. The categories of the study-variables were either the actual coded categories or a collapsed version of the coded categories. Grade level, for example, was converted to two categories: grades 6-9 and 10-12; while separate categories of chemistry curricula were compared. The chemistry programs were: CHEMS (e.g. Hardy, 1970; Herron, 1966; Heath and Stickell, 1963; Pye and Anderson, 1967; and Rainey, 1964), CBA (Heath and Stickell, 1963; and Pye and Anderson, 1967, Nuffield (e.g. Kempa and Dube, 1974; and Meyer, 1970), and MCA (e.g. Charen, 1963). Table 2 lists all study-variable categories, category means, and F-test results.

Insert Table 2 about here.

A study-variable of special interest was outcome test bias. In addition to the analysis of variance (see Table 2) on the test bias variable, a tabulation ("vote count") of test bias against the number of effect sizes favorable to each curriculum type was made and is presented in Table 3. Unlike Walker and Schaffarzick, the analysis also included nonsignificant effect sizes in the tabulation.

Insert Table 3 about here.

## Results and Discussion

### Overall Effect

The weighted mean of the effect sizes is .308; and the standard deviation is .717. A t-test ($t(25)=2.183$; $p<.05$) indicated that this mean is significantly different from zero and favorable to the innovative curricula. Converting the results to percentiles, and placing students taking traditional courses at an average 50th percentile, students taking innovative courses scored on average at the 62nd perentile.

### Distribution of Effect Sizes

Stem-and-leaf diagrams (Table 1) give an indication of distribution and magnitude of effect sizes, weighted and unweighted. Stems (on the left of the vertical line) are broken down into intervals of .2; leaves represent the first decimal place (tenths) of the effect size. The -.0 -- .0 interval includes 26 unweighted effect sizes (ten falling in the negative .0 range, 16 falling in the positive .0 range -- see diagram in the left) and four weighted effect sizes (two positive and two negative -- see diagram on the right). Both stem-and-leaf diagrams indicate a predominance of positive datapoints; 104 out of 151 of the comparison-weighted effect sizes are positive. There is a difference of 46 between the number of sign comparisons (197) and the number of computed effect sizes (151) as a result of studies which were lacking information necessary to compute effect sizes. The 26 study-weighted effect sizes (on the right), 19 positive and 7 negative, are also based only on studies for which there was enough data to compute effect sizes. Peaks at .1 and .2 are consistent in both diagrams. The relatively larger number of points above .3 than below -.3 account for a mean of about .3 in both cases.

Influence of Study Variables

The overall result does not depend on test bias (Table 2). The high weight frequency innovative and neutral tests show a clear superiority of innovative over traditional curricula. The large number of neutral tests in Table 2 reflects the number of tests designed to favor neither the innovative nor traditional curricula in content and therefore classified as neutral. For example, Wasik (1971) analyzed the items of the College Entrance Examination Board (CEEB) Physics Achievement Test into categories of the Taxonomy (Bloom, 1956) and found evidence to support its neutrality with respect to both PSSC and non-PSSC students. Cunningham (1970) designed the neutral Concept Attainment Test based on "Refraction," a topic covered by both PSSC and non-PSSC students in his sample.

Table 2 also shows that neither subject matter nor type of dependent outcome significantly affected the innovative-traditional effect size mean. Principal subject characteristics such as grade level, gender, and academic achievement level also did not have any influence on the effect size mean.

Among design quality features, "unit of analysis" yielded individual and group means of .23 and .88, respectively ($F = 2.27$, $p < .14$). An increase in effect size resulting from group means is to be expected because individual subjects often are a source of variation. However, the relatively small frequency for "group" (N = 3) should be taken into consideration. When those studies involving TOUS (which happened to report group means) were excluded from the analysis, the group mean was .65 ($F = .854$, $p < .37$). Any inferences regarding "unit of analysis" are qualified by the presence of TOUS in some studies.

Data Comparisons

It may be recalled that for reasons pertaining to selection criteria, not all of Walker and Schaffarzick's sources were included in the present investigation. The 23 studies used by those investigators yielded a total of 98 comparisons. From the 197 raw comparison-effect sizes obtained in the present analysis, a subset of nine studies overlap between the two investigations. From the 14 studies omitted from the present analysis, Walker and Schaffarzick reported 44 comparisons out of 53 significantly favorable to the innovative curricula (Table 3). This applies primarily to their findings for elementary students and mathematics curricula.

The top of Table 3 provides vote counts of signs of Walker and Schaffarzick's and the present data. Although this allows for comparison of all comparisons favorable to each treatment, the weighted tabulation at the bottom of Table 3 is a more accurate reflection of the results of the present study since it gives equal weight to each independent study. Either procedure (including Walker and Schaffarzick's data) yields a ratio of approximately four to one in favor of combined outcomes favorable to the innovative curricula.

By the vote count method, Walker and Schaffarzick found stronger evidence than the present study for the superiority of the innovative curricula on the innovative tests (44 out of 45 significant outcomes for their data; 29 out of 37 for the present data). But the present data, both unweighted and weighted, shows the superiority of students taking innovative programs on neutral tests (31 out of 38 unweighted outcomes; 6 out of 7 weighted outcomes) and on traditional tests in the unweighted data (4 out of 6).

## Conclusion

Although great national interest in science curricula by the general public and professional educators may have abated in the 1970s, the post-Sputnic (1958) curricula produced beneficial effects on science learning that extended across science subjects in secondary schools, types of students, various types of cognitive and affective outcomes, and the experimental rigor of the research. Past reviews showed the percentage of positive results; but the present analysis shows a moderate 12 point percentile advantage on all learning measures of average student performance in the innovative courses.

Contrary to Walker and Schaffarzick, who used earlier methods of research synthesis and concluded that performance merely reflects content exposure, the present results suggest that students in secondary-school science courses score moderately better (Effect Size = .308) than students in traditional courses on both innovative and neutral tests and negligibly lower (Effect Size = -.04) on tests favoring traditional science content.

References

Bloom, B. S. (Ed.) Taxonomy of educational objectives: The classification of educational goals. Handbook 1. Cognitive Domain. New York: McKay, 1956.

Bredderman, T. A meta-analysis of elementary school science progress curricula evaluation studies. A paper presented at the National Association for Research in Science Teaching Annual Meeting, Atlanta, 1979.

Charen, G. The effect of open-ended experiments in chemistry on the achievement of certain objectives of science teaching. Journal of Research in Science Teaching, 1963, 1, 184-190.

Cohen, S. A., & Hyman, J. S. How come so many hypotheses in educational research are supported? (A modest proposal.) Educational Research, 1979, 8 (11); 12-16.

Cunningham, J. B. The measurement of concept attainment: A comparative study of modern and traditional high school physics courses. (Doctoral dissertation, Syracuse University, 1970). Dissertation Abstracts International, 1971, 32, 268A. (University Microfilms No. 71-18, 476)

Glass, G. V. Integrating findings: The meta-analysis of research. In L. S. Shulman (ed.), Review of Research in Education, 6, Itasca, Illinois: F.E. Peacock, 1978.

Hardy, C. A. CHEM study and traditional chemistry: An experimental analysis. Science Education, 1970, 54, 273-276.

Heath, R. W. & Stickell, D. W. CHEM and CBA effects on achievement in chemistry. Science Teacher, 1963, 30, 45-46.

Herron, J. D. Evaluation and the new curricula. Journal of Research in Science Teaching, 1966, 4, 159-170.

Kempa, R. F., & Dube, G. E. Science interest and attitude traits in students subsequent to the study of chemistry at the ordinary level of the general certificate of education. Journal of Research in Science Teaching 1974, 11, 361-370.

Mascolo, G. R. Reactions of pupils to Nuffield Science Teaching Project Trial materials in England at the ordinary level of the Général Certificate of Education. Journal of Research in Science Teaching, 1970, 7, 283-302.

Meyer, G. R. Reactions of pupils to Nuffield Science Teaching Project Trial Materials in England at the ordinary level of the General Certificate of Education. Journal of Research in Science Teaching, 1970, 7, 283-302.

Pre-College science curriculum activities of the National Science Foundation, Report of the Science Curriculum Review Team, Volume II-Appendix. Washington, D. C.: The National Science Foundation, 1975.

Pye, E. L., & Anderson, K. H. Test achievements of chemistry students. The Science Teacher, 1967, 34, 30-32.

Walker, D. F., & Schaffarzick, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-111.

Wasik, J. L. A comparison of cognitive performance of PSSC and non-PSSC physics students. Journal of Research in Science Teaching, 1971, 8, 85-90.

Welch, W. W. Twenty years of science curriculum development - a look back. In D. C. Berliner (ed.), Review of Research in Education, 7, Itasca, Illinois: F. E.Peacock, 1979.

Footnote

This research was supported by the National Science Foundation Grant No. 79-17259, Principal Investigator Herbert J. Walberg. Findings and and opinions expressed in this paper are solely the responsibility of the authors.

Table 1

Stem-and-Leaf Diagrams of Effect Sizes

| | Comparison-Effect Sizes | Study-Weighted Effect Sizes |
|---|---|---|
| ≥ 1.5 | 1.7, 1.7, 2.4, 2.7, 2.8, 3.2, 3.5, 4.5 | |
| 1.3 - 1.4 | 3 3 3 4 | |
| 1.1 - 1.2 | 1 2 2 | 2 |
| 0.9 - 1.0 | 9 9 0 0 0 | 0 |
| 0.7 - 0.8 | 7 7 7 7 8 8 | 7 |
| 0.5 - 0.6 | 5 5 5 5 5 5 6 6 6 6 6 6 | 5 5 6 |
| 0.3 - 0.4 | 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 | 3 3 4 |
| 0.1 - 0.2 | 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 | 1 1 1 1 2 2 2 2 |
| -0.0 -(0.0) | 0 0 0 0 0 0 0 0 0 0 / 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | 0 0 / 0 0 |
| -0.2-(-0.1) | 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 | 2 2 1 |
| -0.4-(-0.3) | 4 4 3 3 3 3 | 3 |
| -0.6-(-0.5) | 6 5 5 5 5 5 | 5 |
| -0.8-(-0.7) | 8 8 | |
| -1.0-(-0.9) | 0 | |
| -1.2-(-1.1) | 1 | |
| ≤-1.3 | -1.7, -1.9, -2.7 | |

Table 2

Descriptive Statistics for Independent Variables*

| Study Variable | Mean Effect Size | Standard Deviation | Frequency (Sum of Cases) | F | Prob. |
|---|---|---|---|---|---|
| Totals | .307 | .717 | 26 | | |
| Dependent Measure | | | | | |
| Test Bias | | | | | |
| Innovative | .36 | .69 | 11 | .18 | .84 |
| Traditional | -.04 | .00 | 1 | | |
| Neutral | .29 | .78 | 14 | | |
| Test Bias Information Source | | | | | |
| Independent | .19 | .18 | 7 | .23 | .87 |
| Test Description (Author) | .33 | .86 | 17 | | |
| Lacking Information | .35 | .58 | 2 | | |
| Test Type | | | | | |
| General | .33 | .48 | 5 | .42 | .74 |
| Discipline | .08 | .73 | 7 | | |
| Course/Curriculum | .54 | .75 | 6 | | |
| Science | .29 | .85 | 8 | | |
| Origin | | | | | |
| Local | .28 | .70 | 11 | .03 | .86 |
| Published | .33 | .76 | 15 | | |
| Reliability | | | | | |
| $\leq$.80 | .72 | .00 | 1 | .93 | .34 |
| $\geq$.80 | .29 | .72 | 25 | | |

Table 2
(con't)
page 2 of 7

| | Mean Effect Size | Standard Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Validity | | | | | |
| Adequate | .32 | .74 | 24 | .15 | .70 |
| Inadequate | .09 | .24 | 2 | | |
| Study Source | | | | | |
| Referred Journal | .33 | .74 | 23 | .24 | .63 |
| Dissertation | .11 | .47 | 3 | | |
| Subject Matter | | | | | |
| Biology | .41 | .66 | 6 | .42 | .83 |
| Chemistry | .001 | .63 | 6 | | |
| Physics | .52 | 1.03 | 6 | | |
| General Science | .37 | .39 | 4 | | |
| Physical Science | .13 | .00 | 1 | | |
| Integrated, Unified Science | -.01 | .73 | 1 | | |
| Chemistry Curricula | | | | | |
| Chems | .14 | .35 | 3 | .76 | .69 |
| CBA | -1.45 | .00 | 0 | | |
| Nutfield | -.01 | .81 | 1 | | |
| MCA | .08 | .00 | 1 | | |
| Outcome Measure | | | | | |
| Conceptual | .39 | .87 | 13 | .19 | .96 |
| Inquiry | .21 | .41 | 6 | | |
| Attitude | .16 | .62 | 4 | | |
| Lab Performance | .59 | .00 | 1 | | |
| C[illegible]crete Skills | .16 | 1.29 | 2 | | |

Table 2
(con't)
page 3 of 7

| Subjects and Setting | Mean Effect Size | Standard Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Location | | | | | |
| USA | .32 | .75 | 20 | .53 | .6 |
| Great Britain | -.02 | .71 | 3 | | |
| Israel | .59 | .51 | 3 | | |
| Grade Level | | | | | |
| Below 9 | .20 | .51 | 7 | .22 | .64 |
| 9 - 12 | .35 | .79 | 19 | | |
| Gender | | | | | |
| Male | -.07 | .47 | 3 | .58 | .57 |
| Female | .10 | 1.00 | 2 | | |
| Mixed | .38 | .74 | 21 | | |
| Ethnicity | | | | | |
| White (mixed) | .25 | .68 | 23 | .61 | .62 |
| Other | .72 | .00 | 1 | | |
| Not reported | .78 | 1.42 | 2 | | |
| Academic Achievement Level | | | | | |
| High | .03 | .83 | 4 | .86 | .44 |
| Medium | .41 | .71 | 20 | | |
| Low | -.14 | .47 | 2 | | |
| SES | | | | | |
| Middle Class | -.11 | .33 | 2 | .31 | .87 |
| Upper Middle Class | .39 | .55 | 2 | | |
| Mixed | .61 | .96 | 3 | | |
| Not reported | .30 | .76 | 18 | | |

Table 2
(con't)
page 4 of 7

| Community | Mean Effect Size | Standard Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Urban | .30 | .40 | 5 | .23 | .95 |
| Suburban | .10 | .45 | 5 | | |
| Rural | .28 | .53 | 3 | | |
| Mixed | .25 | .90 | 7 | | |
| Not reported | .53 | 1.07 | 6 | | |
| Sample Size | | | | | |
| less than or equal to 50 | .64 | 1.52 | 3 | .36 | .90 |
| 51 - 100 | .56 | .79 | 5 | | |
| 101 - 200 | .17 | .48 | 5 | | |
| 201 - 500 | .22 | .66 | 6 | | |
| 501 - 750 | .15 | .31 | 4 | | |
| > 750 | .25 | .27 | 2 | | |
| Treatment Characteristics | | | | | |
| Experimental/Control Group Size | | | | | |
| Comparable | .38 | .90 | 9 | .09 | .92 |
| Different (+5) | .27 | .63 | 16 | | |
| Experimental Group Participation | | | | | |
| Elective | .23 | .63 | 15 | .30 | .75 |
| Required | .34 | .91 | 8 | | |
| Either (ie Biology) | .58 | .71 | 3 | | |

Table 2
(con't)
page 5 of 7

| Experimental Teacher | Mean Effect Size | Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Regular | .24 | .78 | 19 | .72 | .41 |
| Special | .50 | .52 | 7 | | |
| Control Teacher | | | | | |
| Comparable to Experimental | .32 | .75 | 24 | .08 | .78 |
| Different | .17 | .18 | 2 | | |
| Focus of Instruction | | | | | |
| Non-lab | .51 | .45 | 3 | .22 | .81 |
| Lecture and lab | .29 | .78 | 21 | | |
| Lab only | .08 | .43 | 2 | | |
| Quality of Instruction | | | | | |
| Curriculum, Course | .28 | .78 | 20 | .10 | .76 |
| Teacher Behavior and Material | .39 | .50 | 6 | | |
| Measure of Variable | | | | | |
| Self Report | .08 | .43 | 2 | .22 | .89 |
| Expert Report | .61 | .00 | 1 | | |
| Pre-determined in structure of materials | .31 | .67 | 17 | | |
| Cannot be determined | .33 | 1.01 | 6 | | |

Table 2
(con't)
page 6 of 7

| Length of Treatment | Mean Effect Size | Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Less than 1 hour | 3.27 | .00 | 0 | .42 | .74 |
| 1 week (11-50 hours) | .35 | .58 | 2 | | |
| Course (10 weeks or more) | .29 | .72 | 23 | | |
| **Control Group Access to Treatment** | | | | | |
| None | .30 | .73 | 24 | .42 | .66 |
| Comparable | .0 | .00 | 1 | | |
| **Sample Selection** | | | | | |
| Simple Random | .44 | .91 | 8 | 1.28 | .31 |
| Purposive (extreme) | -.55 | .00 | 1 | | |
| Matching | .31 | .67 | 12 | | |
| Convenience | .23 | .31 | 5 | | |
| **Unit of Analysis** | | | | | |
| Individual | .23 | .53 | 23 | 2.27 | .14 |
| Group | .88 | 1.68 | 3 | | |
| **Design** | | | | | |
| Quasi-experimental | .31 | .73 | 24 | .58 | .57 |
| True-experimental | .61 | .00 | 1 | | |

Table 2
(con't)
page 7 of 7

| Quality of design** | Mean Effect Size | Deviation | Frequency | F | Prob. |
|---|---|---|---|---|---|
| Average | .30 | .83 | 18 | .006 | .94 |
| High | .32 | .39 | 8 | | |

* Weighting procedure rounds reported statistics to the nearest integer. All tests are run with fractional figures included.

** Computed from sum of potential threats: high quality indicates threats were minimized.

Table 3

Innovative-Traditional Comparisons

Present Data

| | Result of Comparison | | | |
|---|---|---|---|---|
| Test Bias | I>T | T>I | I=T | Total |
| Favors innovative | 29 | 8 | 26 | 63 |
| Favors traditional | 4 | 2 | 7 | 13 |
| Neutral | 31 | 7 | 65 | 103 |
| Totals | 64 | 17 | 98 | 179 |

Walker and Schaffarzick Data

| | Result of Comparison | | | |
|---|---|---|---|---|
| Test Bias | I>T | T>I | I=T | Total |
| Favors innovative | 44 | 1 | 7 | 52 |
| Favors traditional | 5 | 9 | 16 | 30 |
| Neutral | 4 | 3 | 9 | 16 |
| Totals | 53 | 13 | 32 | 98 |

Present Data - Study Weighted

| | Result of Comparison | | | |
|---|---|---|---|---|
| Test Bias | I>T | T>I | I=T | Total |
| Favors innovative | 7 | 1 | 5 | 13 |
| Favors traditional | 0 | 1 | 2 | 3 |
| Neutral | 6 | 1 | 8 | 15 |
| Totals | 13 | 3 | 15 | 31 |

Note: I = Innovative Curricula; T = Traditional Curricula

Appendix

Studies Used in the Meta-analysis

of Curricula

Aikenhead, G. S. Course evaluation II: Interpretation of student performance on evaluative tests. Journal of Research in Science Teaching, 1974, 11, 23-30.

Bowyer, J. B., & Linn, M. C. Effectiveness of the Science Curriculum Improvement Study in teaching scientific literacy. Journal of Research in Science Teaching, 1978, 15, 209-219.

Brakken, E. Intellectual factors in PSSC and conventional high school physics. Journal of Research in Science Teaching, 1965, 3, 19-25.

Bredderman, T. Elementary school science experience and the ability to combine and control variables. Science Education, 1974, 58, 457-469.

Charen, G. The effect of open-ended experiments in chemistry on the achievement of certain objectives of science teaching. Journal of Research in Science Teaching, 1963, 1, 184-190.

Crumb, G. H. Understanding of science in high school physics. Journal of Research in Science Teaching, 1965, 3, 246-250.

Cunningham, J. B. The measurement of concept attainment: A comparative study of modern and traditional high school physics courses. (Doctoral dissertation, Syracuse University, 1970). Dissertation Abstracts International 1971, 32, 268A. (University Microfilms No. 71-18, 476)

Durst, W. N. The ninth-grade physical science programs: an appraisal of achievement, understanding and vocational interest developed through three different physical science curriculums in Lincoln schools. (Doctoral dissertation, University of Nebraska, 1970). Dissertation Abstracts International, 1970,) 31, 1507A. (University Microfilms No. 70-17, 719).

George, K. D. The effect of BSCS and conventional biology on critical thinking. Journal of Research in Science Teaching, 1965, 3, 293-299.

Granger, C. K. & Yager, R. E. Type of high school biology program and its effect on student attitude and achievement in college life science. Journal of Research in Science Teaching, 1970, 7, 383-389.

Hardy, C. A. CHEM study and traditional chemistry: An experimental analysis. Science Education, 1970, 54, 273-276.

Heath, R. W. Curriculum, cognition and educational measurement. Educational and Psychological Measurement 1964, 24, 239-253.

Heath, R. W. & Stickell, D. W. CHEM and CBA effects on achievement in chemistry. Science Teacher, 1963, 30, 45-46.

Herron, J. D. Evaluation and the new curricula. Journal of Research in Science Teaching, 1966, 4, 159-170.

Howe, A. C. & Butts, D. P. The effect of instruction on the acquisition of conservation of volume. Journal of Research in Science Teaching, 1970, 7. 371-375.

Johnson, R. T., Ryan, F. L. & Schroeder, H. Inquiry and the development of positive attitudes. Science Education, 1974, 58, 51-56.

Jungwirth, E. Content-Learning in a process-oriented curriculum: Some aspects of BSCS Biology in Israel. Science Education, 1971, 55, 85-96.

Kempa, R. F., & Dube, G. E. Science interest and attitude traits in students subsequent to the study of chemistry at the ordinary level of the general certificate of education. Journal of Research in Science Teaching 1974, 11, 361-370.

Koelsche, C. L., Newberry, S. L. A study of the relationship between certain variables and the science interests of children. Journal of Research in Science Teaching, 1971, 8, 237-241.

Lisonbee, L. & Fullerton, B. J. The comparative effects of BSCS and traditional biology on student achievement. School Science and Mathematics, 1964, 64, 594-598.

Mascolo, R. Performance in conceptualizing: Relationship between conceptual framework and skills of inquiry. Journal of Research in Science Teaching, 1969, 6, 29-35.

Meyer, G. R. Reactions of pupils to Nuffield Science Teaching Project Trial Materials in England at the ordinary level of the General Certificate of Education. Journal of Research in Science Teaching, 1970, 7, 283-302.

Pye, E. L., & Anderson, K. H. Test achievements of chemistry students. The Science Teacher, 1967, 34, 30-32.

Rainey, R. G. A comparison of the CHEM Study Curriculum and a conventional approach in teaching high school chemistry. School Science and Mathematics, 1964, 64, 539-544.

Ryman, D. Teaching methods, intelligence and gender factors in pupil achievement on a classification task. Journal of Research in Science Teaching, 1977, 14, 401-409.

Scott, N. C. The strategy of inquiry and styles of categorization. Journal of Research in Science Teaching, 1966, 4, 143-153.

Stallings, C. S., & Snyder, W. R. The comparison of the inquiry behavior of ISÇS and non-ISCS science students as measured by the TAB Science Test. Journal of Research in Science Teaching, 1977, 14, 39-44.

Tamir, P. Factors which influence student achievement in high school biology, Journal of Research in Science Teaching, 1976, 13, 539-545.

Tamir, P., & Glassman. A practical examination for BSCS students: A progress report, Journal of Research in Science Teaching, 1971, 8, 307-315.

Tamir, P., & Jungwirth, E. Students' growth as a result of studying BSCS biology for several years. Journal of Research in Science Teaching, 1975, 12, 263-279.

Trent, J. The attainment of the concept "understanding science" using contrasting physics courses. Journal of Research in Science Teaching, 1965, 3, 224-229.

Wasik, J. L. A comparison of cognitive performance of PSSC and non-PSSC physics students. Journal of Research in Science Teaching, 1971, 8, 85-90.

Wright, E. K. Influence of the Science Curriculum Improvement Study on attitudes and process skills of seventh-grade students. (Doctoral dissertation, University of Northern Colorado, 1976.) Dissertation Abstracts International 1977, 37, 4258H. (University Microfilms No. 76-29, 796).

Working Draft
HJW 10-30-1978

# Reflections on Research Synthesis

Herbert J. Walberg
University of Illinois at Chicago Circle

The purpose of this paper, as suggested by the title, is to reflect on recent writings on research-synthesis techniques as well our own experience in integrating research findings across studies. These reflections draw largely on the writings of N. L. Gage and Gene V Glass as well as our ongoing work in analyzing studies in the areas of open education, quality of instruction, reading instruction methods, learning environments, and home environments.

This paper is a working draft for a small, informal conference that is to assist in the planning of synthesis work for the National Science Foundation in the specific areas of quality and quantity of instruction, ability and motivation, home and peer environments, and the social-psychological environment of the class--all in relation to educational outcomes. Therefore, the ideas brought out here should be considered preliminary for the reactions of the two discussants at the conference and two or three other people conducting research syntheses.

Reactions to Glass' Chapter

Gene Glass properly notes at the beginning of his present review chapter that conventional or traditional narrative reviews of research may have been adequate in the past when only five or ten studies were available for analysis. However, the recent growth in educational and psychological research in the last several decades has made it necessary to use more advanced statistical techniques for synthesizing the findings from the many original studies now available. The full range of statistical techniques starting from elementary frequency distributions all the way through multiple regression analyses can usefully serve to synthesize findings. It is immensely difficult in most areas of educational and psychological areas of research these days for an investigator or reviewer to fully understand the meaning of the results unless they are somehow condensed, preferably by objective statistical techniques.

Glass also points out that although the techniques of Light and Smith are valuable in synthesizing original data from studies, they may be somewhat impractical, particularly in education, since it is often difficult to gain the original research data. Both the independent and dependent variables from most educational research studies are measured on uncomparable variables or scales. In addition, many investigators are unable or unwilling to share their original data. And when they do, the scales cannot be effectively compared; for example, even percentiles and grade equivalents make for questionable comparisons because the norming of different instruments from different publishers are based on different sub-populations in different years, all of which introduce bias and error into the analysis.

In his section on problems of access to data, Glass makes a number of excellent points on ERIC, dissertations, microfilms, and other techniques and materials that can be employed in attempting to gather studies for syntheses. Both backward and forward citations can be helpful in identifying a population of studies for review. A point needing emphasis is that it is becoming commonplace for reviewers to dismiss, on certain selected methodological grounds, whole bodies of research literature. Very often quite extensive reviews of studies wind up with two conclusions: 1) that more research needs to be done, and 2) that the prior studies are so weak that nothing can be concluded from a corpus of literature. Although it is possible to argue this case, I believe it is basically unconstructive for building a science of education. <u>The essence of science is the accumulation and replication of evidence</u>. Many hours of work have gone into these studies and reviewers who conclude that nothing can be made of them tend not only to dismiss others' work but the integrity of the field of educational research as well.

I believe that a more positive approach must be urged upon those who wish to synthesize research. The techniques of meta-analysis have already shown that it is quite possible to make definite conclusions across studies when the synthesis techniques are powerful and sensitive. Glass and Smith's reviews of class size and psychotherapy, and Robert Horowitz and Penelope Peterson's reviews of open education have already shown consistent results across hundreds of comparisons that suggest substantive conclusions with definite policy and practical implications. Such research syntheses can also indicate which types of methodological virtues and flaws seem to be the most decisive in determing the outcomes in question. Lastly, this work has shown which particular areas of research within a given topic have been infrequently studied and hence can point to the most decisive kinds of studies that can be done in future work.

Although Glass makes a number of constructive and practical suggestions for doing high quality synthesis in educational research, an additional point needs to be made. It is possible to set up such high standards for literature selection and search that the extent of the investigation goes far beyond the reaches of a particular investigator's budget, time, and energies. The investigator who begins a research synthesis needs to think through very carefully the trade-offs between the scope of the literature and the various analytic techniques.

Just as no single, empirical study can ever be done perfectly it seems unlikely that any meta-analysis of a significant topic in education can be done without some imperfections.

Therefore it is important to plan the scope of the literature and the selection of studies as well as the analytic techniques early so that the investigation does not get out of hand. My personal experience in conducting and advising on large scale primary studies indicates to me that too often investigators are overly ambitious in their efforts and instead of bringing to completion, and publishing studies of a modest scope, they often set such high standards and vast scope as not to be able to finish. Too often, it goes unrecognized in educational research that replication is the essence of science. Often three or four modest studies if they agree in their results can be much more creditable than a single, very large study, especially if the large study expends too much of the effort and budget on data collection rather than analysis, reflections, and writing. It can be added that replication by independent investigators should be as important in syntheses as it is in primary research.

One idea of Glass was also used by Robert Rosenthal and can be valuable in cutting down the size of a research synthesis so that it can be manageable and completed within a given time schedule. In some areas it is possible to find hundreds of studies. When they cannot all be reviewed, it makes perfectly good sense to use the standard techniques of statistical sampling. A simple random sample can be taken, or every third or tenth study might be selected. In addition, stratified sampling may be helpful, that is, an investigator can group the studies into several sets and take a random sample within each set. This is another idea that is just as appropriate in research syntheses as it is in primary research studies.

Another technique that may be used to cut down research syntheses to manageable size is to establish some selection criteria from the very beginning. Although, of course, it is desirable to gain a very large population of studies to cover all areas, the investigator can consider the important policy questions or substantive interests in setting up criteria for admissible studies. For example, one could select studies that have been done after a particular date, or one might confine scope of the syntheses to elementary and secondary schools. James Kulick confined his syntheses of Personal Systems of Instruction to studies in higher education. One might also decide to confine the studies to those which have examined the effects on particular outcomes. Glass and Smith, for example, have done one analysis of the effects of class size on student achievement measures. In subsequent and separate syntheses, they plan to analyze the effects of class size on additudinal outcomes.

Both Glass, and Light and Smith, and Gage are critical of the voting method. This method pertains to those reviews that count the number of favorable effects, mixed and nonsignificant effects, and unfavorable affects for some particular educational treatment. I share many of their reservations about the voting method because, as Glass points out, "to know that televised instruction beats traditional instruction in 25 of 30 studies is not to know whether TV wins by a nose or a walk-away." And, as Glass further points out, one ought to average measures of the strength of the effects or relationships among the variables rather than simply tabulate their sign and possibly significance.

On the other hand, if one has a choice of the voting method or no summary at all, it is perfectly clear that knowing how many studies showed favorable and unfavorable results is of greater relative value. In addition, the voting method and more complicated procedures that take into consideration the probability or strength of the findings will often come to the same conclusions. The voting method has the advantage of being readily understood by educational practitioners, and it is a simple method to use on the part of educational researchers. It may not uncover subtle, interactive, small effects; but if there are consistent effects across a series of studies they will certainly be revealed by the voting method.

I believe that the techniques of analyzing correlations and effect sizes are far superior to the voting method. But having half a loaf in a field that badly needs syntheses it is better than no loaf at all. Moreover, the new techniques of research syntheses are quite complicated to use and to understand, even on the part of well trained investigators. It is also true that quantitative research syntheses have received already some initial skepticism over the beginning efforts. Since there are seeds of doubt as to which particular analytic or summary techniques should be used, it may be advisable in the next five years as we gain experience in using the techniques to use both the simpler techniques to encourage understanding on the part of those who will digest the findings as well as using the more complex techniques such as regression of effect sizes which are certainly more powerful and sensitive to complicated effects in the data.

Another point that deserves amplification in the Glass review is the overemphasis that researchers and reviewers have given to statistical significance. It is obvious that many educational researchers have used students inappropriately as the units of analysis so that the significance levels are inappropriately high. Moreover, it should be much more impressive to us that an effect is consistently positive across a series of studies or investigators and laboratories than that several of the studies happen to be statistically significant.

Glass praises Gage's recent informative book on integrating studies on teaching but critizes Gage's advocacy of Pearson's Chi-square test for integrating probabilities across studies on the grounds that the number of studies will be so large and encompass so many subjects that no hypotheses will be routinely rejected. There are complicated statistical arguments on this matter that still require solution. However, it may be some time before our statistical colleagues settle some of these issues. Therefore, it would seem reasonable in the mean time for educational research synthesists to try using several of the techniques simultaneously. In addition to statistical issues, how convenient the procedures are to employ and how easily educational practitioners can understand them require consideration. Hopefully, before settling on one method simultaneous analyses will indicate the same results and thus will satisfy advocates of the several rival techniques and allow us all to concentrate on substantive and policy implications.

Another valuable point made by Glass is that a variety of simple and complicated techniques can be used in doing research syntheses. Tables, graphs and simple descriptive measures of location and spread will enable readers to comprehend the idea of syntheses across studies. In addition, multiple regression can parsimoniously summarize in one equation all of the results of the syntheses. Such a combination of simple and complicated techniques was recently used by Margaret Uguroglu in her analysis of the relation of motivation and achievement. The correlations between these two variables were tabulated separately for different grade levels, for different subject matter areas, and for different motivation constructs. The reader is allowed to see where the gaps in the data are and the general tendencies with respect to the major variables. After digesting these findings the reader is presented with a multiple regression equation that neatly summarizes the results with several coefficients, showing that the size of the correlation depends principally on the age of the students and the reliability of the measure of motivation. Going from the simple to the complex is a good pedagogical technique that enables readers to gain an understanding of the univariate dependencies before going on to a more complex multivariate synthesis.

Glass and Rosenthal independently arrived at a statistic called the "effect size", which is simply the mean of an experimental group minus the mean of a control group divided by the within-groups standard deviation or the control-group standard deviation. Glass and Smith used the control group standard deviation in their meta-analysis of psychotherapy but used the combined within-group standard deviation in their meta-analysis of class size. The control group standard deviation may have two advantages: 1) it enables the investigator to determine where the experimental group lies in the metric of an untreated control, and 2) it can be readily copied from reports that give separate standard deviations for experimental and control groups. On the other hand, the within-groups standard deviation makes use of the complete variation within both groups and may be a more stable measure.

Glass presents a number of statistical techniques for calculating effect sizes when only the F or T ratios are given. These techniques will be valuable for including the maximum number of studies; but in the interest of efficiency it would also be possible, although perhaps less desirable, for an investigator to eliminate a priori those studies that did not have full statistics as one of the selection criteria.

In areas of research which are basically correlational rather than experimental or quasi-experimental, it is generally advisable to analyze correlations. For example, as discussed by Glass, White found that six hundred, thirty-six available correlations of socio-economic status and achievement averaged .25 with a standard deviation of about .20. The correlation diminished as students got older; the correlation devreased from about .25 in the primary grades to about .15 late in high school. Socio-economic status also correlated higher with verbal mathematic achievement than other outcomes. Glass points out that there is no good reason to transform the correlations to Fisher's Z since it will seldom make much practical difference. Glass also gives a series of guidelines for converting t's and point bi-serial correlations and contingency tables statistics to Pearson correlations.

In the closing sections of his review Glass mentions the problems of differential weighting of studies either by the number of comparisons or the number of students on which a particular comparison is made. These weighting problems can prove highly complex. A small study of one hundred students may make as many as 20 comparisons if, for example, the comparisons are broken down by sex and outcome measure and independent variables of various kinds. On the other hand a study of one thousand may report only the means of an experimental and control group. In simply analyzing the average effects over all these comparisons one would be weighting the smaller study 20 times as much as the single study. Glass makes a point, to which I would agree, that we usually do not have the luxury of throwing out the smaller studies since when various classifications are done the cell sizes for comparisons may be too small. Therefore, it will generally be necessary to analyze all the comparisons.

It is possible to deal with this problem through a weighted regression analysis which is available on some computer programs. Another and possibly easier alternative is to perform what economists call a sensitivity test. That is to eliminate one or several studies at a time from the analysis to determine how the overall results are affected. Still another possibility is to plot either the effect sizes or correlations or the residuals from a regression on a variable such as the numbers of cases on which comparisons are being made. By examining such scatter plots unusual results can usually be detected. If the results do not appear to be determined by sample size or other characteristics it can be safely concluded that the results are not dependent on aberrations of sample sizes or other variables.

Reactions to Glass and Smith's Analysis of Class Size

In Glass and Smith's analysis, a number of examples of research synthesis techniques make more concrete the comments that Glass has made earlier and that I, in turn, have commented upon. These points deserve emulation in future research syntheses. Glass and Smith begin by noting that prior reviews have been haphazard and over-selective in reviewing the literature. Moreover, the reviews are narrow and discursive, use crude classifications, and over-emphasize statistical significance. In contrast Glass and Smith's analysis shows very definite and significant beneficial effects of small class sizes.

Glass and Smith uncovered some 80 studies which exceed by 50% the number in the largest prior review, but they estimate that they perhaps found only about half the studies that might be found using still more exhaustive search procedures. Hundreds of dissertations were scanned but only 30 seemed worth purchasing and 16 were actually useful. That only 80 studies of perhaps 160 that may be eventually uncovered were found in a fairly exhaustive research shows that there are diminishing returns in attempting to find additional literature. The fact that Glass and Smith went back some seventy years to uncover fugitive materials and ordered dissertations and unpublished studies indicates that such diminishing returns are likely to occur.

One could always recommend that the additional 80 studies should have been sought out; but, as I have emphasized above, it may be impractical to do so. In fact, contrary to Glass and Smith, it might be argued that dissertations and unpublished reports should have been excluded. The published literature is more readily accessible, and is likely to be refereed and of higher caliber. The effort to get dissertations or unpublished reports may require three times more effort than a published paper. Renouncing unpublished material might have made it possible for Glass and Smith to review the effect of class size on not only achievement but also affective outcomes with the same amount of time, energy, and budget. Either strategy would be defensible, but it should be emphasized that a trade-off among the different areas of effort is required in making a research synthesis.

Glass and Smith describe several studies in detail; these illustrate, as they point out, the characteristics and texture of the research literature that is reviewed. To simply report statistics, and particularly numbers from the tables, would be inadequate; the reader needs a qualitative feel for a few illustrative studies to understand the statistical results.

In planning the coding of studies, Glass and Smith identify characteristics that may interact with class size in determining achievement levels. First they read a few studies, then talked with experts, and finally made their best guess as to which characteristics of the studies should be coded. Modifications could later be made in the coding; but if one changes the coding, all the studies that have been done up to that point need to be re-coded.

Glass and Smith used five broad categories in categorizing the studies:

(1) study identification,

(2) method of instruction,

(3) classroom demographics,

(4) study conditions, and

(5) outcome variables.

In all they included 25 specific continuous and qualitative variables under these five categories. Including these specific variables makes it possible for the analyst to determine whether the relation between class size (or any other variable being investigated) and achievement is dependent on the characteristics of the study or the characteristics of the populations being investigated, such as elementary and secondary schools. In our work, we have used more detailed and exhaustive categories than did Glass and Smith. For example, instead of simply using three or four experimental design categories, we have used the threats to validity from the Cook-Campbell chapter which is much more extensive. This only goes to emphasize the various kinds of trade-offs of energies that can be planned.

On further reflection about our own work, I believe the Glass and Smith more simple characterization of experimental designs is more practical. It only requires one variable with several levels to record. On the other hand, we are using approximately 14 variables, each with about three levels. Such detailed coding of methodological characteristics has the disadvantage of requiring more time but permits the options during the analysis of either grouping or not grouping methodological characteristics.

On page 12 of their review, Glass and Smith note that the within-group standard deviation was used in their analysis. We noted, however, in their analysis of psychotherapy that the control-group standard deviation was used. As commented on earlier, the trade-off between these two different metrics each have advantages. On page 18, Glass and Smith report their results in percentile metrics rather than Z-scores. In any final tabulation of the results it is possible to present these in either one of the terms. Most educators will understand the percentile results better and hence the Glass and Smith report on class size will be readily comprehensible to practicing educators. One of the many fine characteristics of the Glass and Smith report is the extensive use of concrete information, particularly numerical information throughout the report. For example, on page 19, it is stated that 77 studies were reviewed, 725 effect sizes were calculated, and that these were based on some 900,000 students over a period of 70 years in research in 12 countries.

Beginning on page 20, Glass and Smith provide a series of univariate tabulations which make clear how many times various samples and measures have been researched in the class-size literature; they are unable to identify from these figures the under-studied and over-studied areas. It might be added that it would be useful in the Glass and Smith report not only to present the frequencies in which samples have been investigated but also the average effect size for the various cells so that the reader can see the univariate dependencies before going on to controlled comparisons and grand means across all studies.

Glass and Smith perform what economists call sensitivity analyses. They note, for example, that the relationship between class size and achievement is stronger in those studies that have randomly assigned students to class sizes in strict experimental design terms than in correlational studies. These results give strong reason for imputing causality, since it discredits rival hypotheses such as the co-determination of size and achievement by educational spending or community social class. Glass and Smith point out whether experimental controls mediate findings is an empirical, not an a priori question. In their review of psychotherapy, it was not found that experimental rigor determined the strength of the relationships but in the present study of class-size it did.

On page 42, Glass and Smith note stronger effects on elementary students. It appears that the age or grade levels of students should be included in most meta-analyses of educational effects because in reviews we have been examining the age level has usually mediated the relationships between the independent and dependent variables. For example, Uguroglu found stronger relationships between motivation and achievement in older secondary samples than in elementary school samples. White found that the correlation between social class and achievement was higher in younger samples.

In conclusion, the fine work of Glass and Smith is likely to become a classic synthesis in educational research. I believe it essentially settles the class-size question after so many years of uncertainty and controversy and points confidently to the benefits of smaller classes. Ten more syntheses of this quality will make educational research a science of results rather than of mere methods.

Comments on Gage's Book

Gage has contributed a number of useful insights for investigators who are about to begin research syntheses. Although the book centers on the quality of instruction, his insights have implications for other substantive areas.

On page 26, Gage makes a central point which should be considered in all research syntheses. Nine prior narrative reviews of the effects of teaching on learning conclude that educational research has not identified the consistent replicable features of teaching that are related to student outcomes. Gage points out, however, that these conclusions may be due more to the faults of the reviewers than to the totality of original research itself. Reviewers have made a great number of errors in attempting to synthesize the research. Many studies of teaching, for example, are based on limited numbers of teachers. Therefore, the results may not be statistically significant. On the other hand, to return to an earlier point made here, "replication is the essence of science." It is not two or three significant relationships that are important, but rather consistency of the direction and the magnitude of the effects across many investigations.

Gage critizes Duncan and Biddle's exhaustive review, titled The Study of Teaching. Not only did Duncan and Biddle err in over-emphasizing statistical significance, they were not explicit in stating how studies were categorized as showing positive or negative effects of a particular teaching technique. Just as explicit procedures are necessary in primary research, it is important that explicit objective procedures be followed in reviewing research as well. Duncan and Biddle, despite the great length of their book, do not describe exactly how a determination was made of what particular variable had favorable affects on student outcomes. They claim to use subjective clinical procedures, but these procedures are not spelled out. One has no way of knowing exactly how they were accomplished or how a person could repeat the procedures as a check on the reviewers.. Such a review must be an argument basically from authority rather than categorized evidence. Such arguments are not in the domain of science.

It might be added that the Duncan and Biddle review is more in the nature of advice to practitioners in some respects than a report to scientists. There is an inherent conflict of interest between the practitioner and the scientist that occasionally plagues education. The scientist wants to know exactly how the results were obtained; whereas, perhaps practitioners might be satisfied with conclusions and advice. This conflict seems to be a major difficulty in many research reviews where conclusions gain more prominence than procedures of coming to the conclusions and the nature of the evidence.

On page 38, Gage makes the excellent point that some teaching variables do not vary over a broad range. It is well known that if a variable does not vary it cannot co-vary effectively or with statistical significance with another variable. Gage wisely recommends considering the variability of the independent variables of teaching when considering their relationship to student outcomes.

Another interesting point made by Gage is the need for separating the results of different educational outcomes. In addition to citing the work of Kulick and McKeachie, he cites evidence that higher level thought questions seem to produce lower levels of achievement among students. The results suggest that lower-order questions produce more factual achievement and higher level questions produce better results for higher cognitive levels of achievement. If the results are mixed together, the analysis will be insensitive to an important distinction that applies in the research data.

An additional example is Horowitz's box score tabulation of open-education effects later confirmed by Peterson's tabulation of effect sizes. Both of these reviews indicate that open education overall seems to lead to slightly lower factual achievement on standardized examinations but strongly higher levels of performance on tests of creativity and independence and various affective outcomes. This work confirms Gage's point that the results from different outcome variables need to be tabulated separately.

Comments on Uguroglu's Synthesis

The Uguroglu-Walberg synthesis of the relationship between motivation and achievement suggests a number of points that can be mentioned here. One of the first points to be brought out in the Uguroglu paper is the chestnut that correlation does not imply causation. Simply tabulating hundreds of correlations between motivation and achievement does not establish, for example, whether motivation causes educational achievement or achievement causes higher levels of motivation or whether both factors are caused simultaneously by other variables. Nevertheless, a general estimate of the correlation between the two and showing how the correlation varies across various samples and types of motivation is useful in establishing what Blalock as called "an inventory of causes and effects". In this particular instance, if the correlation is found to be consistently positive, motivation ought to be entered into experimental and survey designs that hope to elicit the causal dependency of achievement on the production factors in education.

The Uguroglu paper illustrates a dilemma of educational psychology. Since the James-Lange theory of emotion there have been many theories of motivation. There has been much arm chair speculation and voluminous writings in the field. Nevertheless empirical work in education and psychology rarely fits a particularly psychological theory or tests one theory against another in their power to explain empirical results. Therefore there is a great gap between theories and empirical work, and it is usually difficult to establish the constructs being investigated from the empirical works. For example, Shavelson mentions 22 review articles on self-concept alone that show 17 different conceptual categories. Since self-concept is only one sub-construct of motivation, one can see that the total number of constructs and sub-constructs can be quite large and beyond the limits of synthesis. It will be difficult to find several studies of each of the sub-constructs and therefore difficult to establish the relation or correlation of each sub-construct with different types of educational outcomes.

On page 4, Uguroglu gives a one or two sentence overview of each of five major views of the field of motivation, which will be useful for readers who want to get particular perspectives beyond the summary on empirical relationships between motivation and educational outcomes. On pages 5 and 6 Uguroglu introduces the idea of replication in meta-analysis by taking from the works of Benjamin Bloom a calibration sample of 122 correlations. Working in an explicit framework she also searched through a Psychological Abstracts Reading Research Quarterly for a three year period for more recent studies: The empirical analysis can ask if the overall correlation estimated by Bloom is replicated in the validation sample. Uguroglu tabulates the correlations by sample size, grade level, sex of the sample, reliability of the motivation measure, nationality, and characteristics of the motivation and outcomes measures. It appears that the age of the sample and the characteristics of the measures, including their reliabilities are likely to turn up to be significant determinations of the correlations between education production factors and educational outcomes; these characteristics then should be included in future meta-analyses.

On page 8, Uguroglu presents stem-and-leaf diagrams. Each value of these show each individual data point in the total sample. This gives readers a concrete feeling for the range of the data and aberrent data points. The stem-and-leaf diagrams for the calibration and validation samples show the interesting distributional properties of two. The calibration sample reveals a slight tendency toward bi-modality which peaks at about .30 and .51. The validation sample is more normally distributed but has a few negative correlations based on the younger children in the primary grades. The validation sample also has two outlying correlations, .98 and -.31. Stem-and-leaf diagrams are a useful way to introduce statistical analysis of correlations to readers who may not be familiar with the idea.

Pages 9 and 10 discuss the dependency of the correlation of achievement and motivation on characteristics of the samples and the measures employed. These can be understood as one-way analyses of variance. The tables present the average correlation for the cell, the standard deviation of the correlation, and the number of correlations on which the mean and standard deviation are based. The first of these tables for example, shows that the linkage between motivation and achievement is higher in the older samples and it is quite low in the very young samples, in fact in some cases negative. Relatively simple tabulations introduce gradually the idea of the comparisons across the independent variables.

The regression control results, however, offer a much more parsimonious accounting for the significant trends in the data. Experimenting with various forms of the regression equation makes it possible to find that smaller sets of variables account for just about as much variance as the entire 25 variables that first entered the equation. Moreover, it can be concluded from these regressions with some degree of confidence that significant variables in the regression are controlled for one another, and even if the variables excluded from the equation were to be entered those that are in the final equation would still be significant. The regression yields a parsimonious set of potent, unique determinants of the relationship between motivation and educational outcomes.

It is also possible to exclude unusual studies such as those with the two outlying observations that were mentioned above. In this case, excluding the outlying studies made very little difference in the regression weights. On the other hand the very large sample size from the Coleman report suggested that the larger the sample, the smaller the correlation. However, omitting the Coleman report in the analysis suggests that this trend is not consistent in the other studies. Therefore, it is attributable only to the Coleman study because of its large magnitude.

The results further suggest that one of the complicating but significant results is attributable to the accidents of only one or two studies, particularly a large study that contains many correlations in mathematics achievement. Uguroglu expresses skepticism that these results would nec essarily be confirmed by other studies

On page 15, the analyses suggest that more reliable measures paradoxically are less closely associated with achievement than more reliable measures which may also be uncovered in other meta-analyses. This strange finding probably stems from the tendency of more internally consistent tests to have narrowed factorial content; higher internal consistency yields lower external consistency, that is, correlations with external criteria. Narrowing the scope of prediction instrument diminishes the relationship of the very criteria it is intended to predict.

## Conclusions

In ruminating about several research syntheses by others as well as our own experience it appears to me that the techniques of Gage, Glass, Light, Smith, and others will accelerate progress in educational research. Our most precious resource in formulating educational policy is the true experiment with random assignment to conditions in natural settings of learning, but these are comparatively rare in educational research. Nevertheless, we are able to draw on areas of research that have employed correlational or quasi-experimental designs. We cannot conclude from the correlational relationships established from these that certain production factors actually cause achievement but if they are supported by plausibility as well as empirical conformation they have to be suspected as possible causes just as the linkage between cigarette smoking and lung cancer should suggest caustion about smoking. Thus tabulation and analysis of correlational relationships, if nothing else, can produce inventories of causes and effects that ought to be taken into consideration in future work. The strongest correlates suggest those factors that ought to be investigated with experiments.

If in certain situations experiments cannot be done, then the investigator is obliged to fall back on correlational studies. Correlational studies that only take into consideration one or two possible causal determinants should be far less convincing than those studies that include the complete set of consistent correlates of outcome measures.

It seems clear to us even at this preliminary stage of the study of educational productivity that the following plausible, consistent correlates of outcomes need to be considered: student ability (including prior achievement) and motivation, the quality and quantity of instruction, the home, school, and peer environment. Including these factors, even in experiments, can prove valuable because these factors are consistent, potent covariates. By including them in regression equations one can get a more precise estimate of the weight of the factor of interest, for example, quality of instruction controlled for all the other factors. By having a consistent model that includes most or all of these factors in subsequent research, the replicability of a more fully specified equation rather than simply the relationship of one independent variable to one dependent variable can be more solidly established.

It is even more important to include these factors in correlational studies because they do not randomly assign the chief variable for investigation. It is well known, for example, that the home environment, that is, the intensity and amount of educationally-stimulating interaction between the parent and the child is a potent correlate of achievement and indeed with achievement gain; so that it would be important to include this variable in research on quality of instruction. Children who are receiving higher quality of instruction may also have more stimulating home environment. Children who are stimulated at home can in fact evoke higher quality of teaching in their classroom. Individual children can demand the sort of attention from the teacher and also children in schools in stimulating neighborhoods can evoke better teaching on the part of the faculty. This suggests causation from the home environment to both achievement and to quality of teaching.

One way of investigating these affects is to then include all possible causes in the equation. Those that survive screening regression techniques and make a unique contribution to the explanation of educational outcomes can have greater creditability. However, advanced techniques of econometric analysis such as two stage, least squares regressions are even more powerful in sorting out reverse cause and third cause phenomena in educational data sets, particularly in panel data in which multiple units are measured on multiple occasions.

We may be at square one with respect to what needs to be done to develop an equation for estimating educational productivity, but the research synthesis of prior literature to develop an inventory of possible causes and effects will be a major step forward. Subsequent research which includes the potent constructs and sub-constructs which are identified in the research synthesis can take into consideration a more complete set of possible causes. This kind of research can rapidly accelerate the accumlation of knowledge about the causes of educational achievement and other outcomes as well as develop a more adequate scientific and practical theory of educational productivity.

# UICC-NSF META-ANALYSIS PROJECT

## CODE BOOK

Barbara K. Kremer and F. David Boulanger

Coding Schemes for
Individual Study Characteristics and
Meta - Analysis Statistics

The purpose of a coding scheme is to provide a quantitative, computer-retrivable summarization of the key characteristics of each comparison and each correlation in each study included in the meta-analysis. Since a single comparison or single correlation is the unit around which the coding scheme is constructed, there will be often be several code sheets for one study.

Each coding scheme has three parts:

(1) General Characteristics of the study

(2) Specific Characteristics of the Construct under Study

(3) Methodological Characteristics and Meta-Analysis Statistics

The form of the General Characteristics section is identical in each coding scheme. Among other things, it includes the identification of the dependent variable to which the effect-size or correlation reported at the end of the coding scheme is related.

The Specific Characteristics of the Construct section has eight different forms, one corresponding to each construct considered in the meta-analysis project, namely: Maturation, Ability, Age or Developmental Level, Quantity of Instruction, Quality of Instruction, Home Environment, Peer Environment, and Classroom Environment. A given comparison or correlation extracted from a given study will be coded according to the independent or predictor variable into one of these constructs. Most studies will report variables relevant to only one construct.

The form of the Methodological Characteristics section is, like the Descriptive Characteristics section, identical in each coding scheme. The methodological flaws recorded here will form one criterion for the selection of studies to be included in various parts of the later statistical analysis. For example, a comparison of outcomes between true and quasi-experiments in a particular construct might be of interest; or it may be desirable to exclude all studies with certain flaws.

The last entry in the Methodological section is the correlation or effect-size that relates the independent or predictor variable in the Specific Characteristics section to the dependent variable in the General Characteristics section. As noted earlier, one study may be coded in near identical manner on several code sheets with only the two variables and the correlation or effect-size differing among sheets.

Code sheets for each construct are attached at the end of the Code Book.

Section I: General Characteristics of the Study

Section III: Methodological Characteristics and
Meta-Analysis Statistics

(each code sheet will contain each of these sections)

I. General Study Characteristics and Dependent Variable

| COLS. | Study Identification |
|---|---|
| 1- 4 | Sheet Number (four digits) |
| 5-16 | Author, last name, comma, additional last names |
| 17-18 | Year of study (last two digits) |
| 19-21 | Number of study (three digits) |
| 22 | Country of origin<br>1 = U.S. and Canada<br>2 = Britain<br>3 = Australia<br>4 = Other English-speaking countries (e.g., as in Africa)<br>5 = Non-English speaking countries |

Source of Reference

| | |
|---|---|
| 23 | 1 = refereed journal<br>2 = ERIC (not dissertation)<br>3 = dissertation or thesis abstract<br>4 = unpublished research report |

COLS. | Science Learning Outcomes

24-25

01 = Cognitive Achievement, General e.g., Standarized Achievement Test or any test with some mix of cognitive fact, concept, process, logical operation.

02 = Factual i.e., identification or recall of specific information previously learned.

03 = Conceptual i.e., generalization of a concept to a new situation. Not factual. Not identified by the author as process or logical operation.

04 = Process i.e., identified by the author as process outcomes and, on inspection, not factual category.

05 = Logical operations in Piaget's theory i.e., identified as logical operations and, on inspection, not factual category.

06 = Attitudes and interests toward science, scientists, science careers, science instruction.

07 = Critical thinking or creative applications. Identified by the author as critical or creative thinking and, on inspection, not factual category.

08 = Lab skills or performance test.

| COLS. | Dependent Measure |
|---|---|
| 26 | 1 = General |
| | 2 = Discipline specific |
| | 3 = Curriculum or course specific |
| 27 | 1 = locally developed test |
| | 2 = published test |
| | Reliability of Outcome Measure (leave blank where not given) |
| 28-29 | internal consistency (enter value) |
| 30-31 | interobserver reliability (enter value) |
| 32-33 | stability - test-retest (enter value) |
| 34-35 | alternate - forms (enter value) |
| 36 | 1 = adequate consideration of outcome measure validity. (Does the dependent measure represent a reasonable approximation of the outcome variable under consideration, without "teaching to the test"?) |
| | 2 = inadequate consideration of outcome measure validity |

| COLS. | Grade Level of Subjects (enter "median" grade if more than one considered or lower grade of two) |
|---|---|
| 37-38 | 00 = Kindergarten/preschool |
| | 01 = Grade 1 |
| | 02 = Grade 2 |
| | 03 = Grade 3 |
| | 04 = Grade 4 |
| | 05 = Grade 5 |
| | 06 = Grade 6 |
| | 07 = Grade 7 |
| | 08 = Grade 8 |
| | 09 = Grade 9 |
| | 10 = Grade 10 |
| | 11 = Grade 11 |
| | 12 = Grade 12 |
| | 13 = College of Adult |
| | Sex of Subjects |
| 39 | 1 = male |
| | 2 = female |
| | 3 = mixed sex sample |
| | Ethnicity of Subjects |
| 40 | 1 = Black |
| | 2 = White |
| | 3 = Latino |
| | 4 = Oriental |
| | 5 = Mixed ethnic sample |
| | 6 = Other ethnic, including foreign studies |
| | 7 = not specified |
| | Academic achievement level of subjects and/or academic aptitude (IQ). Assume medium unless otherwise specified. |
| 41 | 1 = high |
| | 2 = medium (as specified by verbal statement or 90-110 on intelligence measure) |
| | 3 = low |

COLS. Subjects' SES

42 1 = poor, disadvantaged

2 = middle class (including working, and lower middle class)

3 = upper middle class

4 = upper class

5 = mixed SES sample

6 = not specified

Community-type

43 1 = urban

2 = suburban

3 = rural

4 = mixed sample with regard to community

5 = not specified

Disciplinary Focus of the Study

44-45 01 = Biology

02 = Chemistry

03 = Physics

04 = General Science

05 = Earth Science

06 = Life Science

07 = Physical Science

08 = Integrated or Unified Science

09 = Environmental Science

10 = Behavioral Science

Curricular focus of Study

46 1 = Nationally funded curriculum project (BSCS, HPP, ISCS, S-APA, ESCP, etc.)

2 = Conventional, traditional, locally developed, unspecified.

Consideration of Production Factors in Study

47 Classroom environment

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

COLS.

48 Ability

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

49 Motivation

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

50 Quality of Instruction

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

51 Quantity of Instruction

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

52 Home

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

53 Peer

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

COLS.

54 Age/Developmental Level

1 = omitted

2 = measured and employed in analysis (includes use of measure in stratification, blocking, covariation)

3 = exemplary

## III. Methodological Characteristics and Meta-Analysis Statistics

COLS.

1- 4 Sheet Number (Enter)

5 Sample selection

1 = simple or stratified random

2 = purposive (a priori) sample (Extreme or specialized group)

3 = matching

4 = convenience or ill-specified sample

6 Unit of analysis

1 = individual

2 = group

7 Study design (Campbell & Stanley)

1 = correlational

2 = quasi - experimental

3 = true experimental (random assignment)

8 Reliability of treatment implementation

1 = low; treatment and implementation poorly described and documented

2 = adequate; treatment and implementation clearly described and documented

3 = high; treatment described or documented with observational checks on implementation

9 Statistical power

1 = inadequate

2 = adequate, i.e., 6 or more classes; 100 or more individuals total in 2 comparison groups or in correlational group

10 Error rate (Given the number of comparisons or correlations, is the p level sufficiently low to assure a less than .05 chance occurance of this relationship?)

1 = inadequate p level

2 = adequate, i.e., p less than .05

| COLS. | |
|---|---|
| 11 | Maturation (have factors within units rather than the treatment brought about the differences observed?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 12 | History (Have external factors in the environment rather than the treatment brought about the differences observed?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 13 | Selection Bias (Do pre-existing differences among the groups account for later observed differences?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 14 | Contamination, Compensation, Differential Incentive (Do untreated control groups work harder, work less, or somehow gain benefits or lose incentive due to influences from treated groups or teachers?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 15 | Mortality (Do different dropout rates account for observed differences?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 16 | Generalizability (Can results be generalized to other times, units, or settings with similar demographic characteristics?)<br>1 = probable threat<br>2 = adequately minimized<br>3 = information not provided |
| 17-20 | Correlation; if positive, leave sign space blank, if negative, write minus sign (-) in sign space. |

COLS.

| | |
|---|---|
| 21 | Type of Correlation<br>1 = partial<br>2 = part<br>3 = zero-order |
| 22-27 | Effect size<br>Enter 99.999 if not computable<br>Enter effect size following formular or other approach recommended by Glass<br>$\frac{\bar{X}exp - \bar{X}con}{SDcon}$ |
| 28 | Direction of effect size<br>1 = significantly (p< .05) favors control<br>2 = favors control, not significantly<br>3 = favors experimental treatment group, not significantly<br>4 = significantly (p< .05) favors experimental treatment |
| 29-32 | Level of significance or p value<br>Enter .999 if not specified<br>Enter p-value if available, otherwise enter alpha level met |
| 33-36 | Sample size enter, right justify (for effect size, sum of sample in groups compared) |

Eight Versions of Section II: Specific Characteristics of Constructs under Study

(each code sheet will contain only one version corresponding to one construct)

# II. Home Construct

| COLS. | Home Factors under Study |
|---|---|
| | Standard socioeconomic characteristics |
| 55-56 | 01 = Parent education |
| | 02 = Parent income |
| | 03 = Parent occupation level |
| | 04 = Housing value (of specific house or apartment) |
| | 05 = Neighborhood or community SES |
| | Family constellation |
| | 06 = # of children in the family |
| | 07 = Adult-child ratio in the home |
| | 08 = Birth order of student |
| | 09 = Single parent homes |
| | 10 = Crowding ratio (# of family members, rooms in house or apartment) |
| | 11 = # of persons living in the home |
| | 12 = Presence of science-related equipment and documents in the home |
| | 13 = Gender differences: sex-role stereo-typing |
| | 14 = Ethnic comparisons (within societies) exclude cross-nation |
| | 15 = Parental aspirations for child and attitudes to education |
| | 16 = Parent involvement in the school and the child's schoolwork (Keeves). |
| | 17 = Generalized SES - Judgment criteria may not be specified |
| | 18 = Multiple index SES |
| | Presence of Home Variable in Study |
| 57 | 1 = independent variable |
| | 2 = mediating or covariate variable |
| | Method of Collecting Home Information |
| 58 | 1 = Parents' questionnaire |
| | 2 = Students' questionnaire |
| | 3 = Home interview with parents |
| | 4 = Parent interviews outside the home (e.g., the school) |

Method of Collecting Home Information (continued)

5 = School records, archives

6 = Not reported

7 = Teacher or other staff rating of SES

8 = Teacher or other staff rating of home support and stimulation

9 = Multiple methods used

Validity of Home Measure

59 1 = adequate

2 = inadequate

3 = exemplary

| COLS. | Peer Variables under Study |
|---|---|
| 55 | Peer grouping |
| | 1 = Within classes (during instruction, e.g., individuals vs. group work) |
| | 2 = Between classes (tracking) |
| | 3 = School activities (athletics, extra-curricular) |
| | 4 = Outside of school (e.g., sociological characteristics of peer groupings) |
| | Participation/Interaction |
| | 5 = Degree of Participation/Interaction |
| | 6 = Quality or style of participation/interaction |
| | Subject Placement in Peer Groupings |
| 56 | 1 = Assigned |
| | 2 = Choice within requirement |
| | 3 = Free choice |
| | 4 = Intact groups |
| | 9 = variable under study |
| 57-58 | Enter categories compared |
| | Bases for Placement or Criteria for Group Membership |
| 59 | 1 = Ability |
| | 2 = Interest |
| | 3 = Psychological Characteristics (creativity, field dependence, independence) |
| | 4 = Peer acceptance |
| | 5 = Course or curricular enrollment |
| | 6 = Teacher judgment |
| | 7 = Arbitrary or unclear |
| | 8 = Combination of the above |
| | 9 = Uncategorized |
| | Position of Peer Variable in Study |
| 60 | 1 = independent variable |
| | 2 = mediating or covariate |

| COLS. | Type of Peer Measure |
|---|---|
| 61 | 1 = Observer report |
| | 2 = Self-report |
| | 3 = Standardized scale or instrument |
| | 4 = Teacher report |
| | 5 = Combination of the above |
| | Validity of Peer Measure |
| 62 | 1 = adequate |
| | 2 = inadequate |

## II. Motivation Construct

| COLS. | |
|---|---|
| | Motivation Variable Under Study |
| 55-56 | 01 = academic, n-achievement |
| | 02 = persistence |
| | 03 = intrinsic motivation |
| | 04 = locus of control |
| | 05 = self-concept (personal) |
| | 06 = continuing motivation, interest in academic study outside of school |
| | 07 = feedback/academic evaluation |
| | 08 = test anxiety |
| | 09 = attribution of causality |
| | 10 = perceived ability/success |
| | 11 = risk-taking |
| | 12 = academic self-concept or concept of ability |
| | Position of Motivation Variable in Study |
| 57 | 1 = independent |
| | 2 = mediating or covariate |
| | Motivation Measure |
| 58 | 1 = standardized scale or instrument |
| | 2 = local instrument or scoring technique |
| | 3 = observations |
| | 4 = other |
| | Reliability of Motivation Measure |
| 59-60 | (enter reliability value; whether reported or estimated) |
| | Motivation Level of Subjects |
| 61 | 1 = low motivation sample |
| | 2 = high motivation sample |
| | 3 = mixed sample; high vs. low motivation group |
| | 4 = no control for motivation, in sample; or convenience sample |
| | 5 = no information on sampling |

| COLS. | <u>Orientation of Study</u> |
|---|---|
| 62 | 1 = interventionist/experimental (focus of investigation is on increasing or otherwise controlling motivation) |
| | 2 = non-interventionist studies (including descriptive/correlational investigations) |
| | <u>Interventionist Manipulation</u> (leave blank if study non-interventionist in character) |
| 63 | 1 = Task, materials |
| | 2 = Teacher behavior |
| | 3 = Classroom environment (opern vs. closed, co-op vs. comp., matching instruction) |
| | 4 = Other |
| | <u>Validity of Motivation Measure</u> |
| 64 | 1 = Adequate consideration of independent measure validity. (Does the independent measure represent a reasonably approximation of the variable under consideration) |
| | 2 = inadequate consideration of motivation measure validity |

## II. Ability Construct

<u>COLS.</u>

55 Ability variable under study

1 = <u>General ability, aptitutde</u> (intelligence, mental maturity, general or subject specific aptitude, culture free measures of ability)

2 = <u>Pretested knowledge or skill</u> specific to the particular treatment or criterion measure; cognitive entry behavior. Includes the case where the same process or achievement measure is given pre and post.

3 = <u>Past achievement</u> (GPA, grades, general or subject area achievement)

4*= <u>Past rate of learning</u> (efficiency of learning, speed on treatment or criterion related tasks)

5*= <u>Cognitive style</u> (field dependence, cognitive preference, work style)

6*= Creativity or creative thinking

7 = Verbal aptitude

8 = Quantitative aptitude

9 = Mechanical - Spatial reasoning

56 Position of ability variable in study. If 5 or 6, list dependent variable in columns 24, 25 under general characteristics of the study.

1 = blocking variable

2 = covariate

3 = independent

4 = mediating

5 = covariate and dependent

6 = independent and dependent

57 Ability measure

1 = standardized scale or instrument

2 = local instrument or scoring technique

3 = research instrument not yet standardized

4 = observations, ratings

5 = not reported

58 Reliability of ability measure

1 = reported in study

2 = estimated

*Category deleted for lack of studies.

COLS

59-60 Reliability value, whether reported or estimated.

61 Estimated or reported general ability level of subjects on general ability, past achievement, or past rate of learning.

1 = low ability (below -1 SD)

2 = below average (-1 SD to mean)

3 = average ability (-1 SD to +1 SD)

4 = above average (mean to +1 SD)

5 = high ability (above +1 SD)

6 = information on sample insufficient to make an estimate

62 Character of study

1 = non-intervention, correlational

2 = interventionist (quality of instruction)

3 = interventionist (quantity of instruction)

4 = interventionist (motivational)

5 = interventionist (classroom environment)

(2–5: Dependent variable related treatment between test and criterion)

63 Time lapse between test (predictor) and criterion.

1 = Concurrent (less than 1 week)

2 = 1 week to 4 weeks inclusive

3 = greater than 1 month to 6 months inclusive

4 = more than 6 months

## II. Age/Developmental Level Construct

COLS.

55 Age/Developmental Level variable under study

1 = Chronological age, year in school

2 = Piaget stage

3 = Piaget logical operations associated with stages

4 = Kolberg moral stage

5 = Kolberg moral judgments associated with moral stages

6 = Havighurst's stages

7 = Erickson's stages

56 Age/Level measure

1 = Scored imitation of content and method of presentation found in original source

2 = Novel tasks, individually administered, based on the original theory

3 = Group demonstration with individual responses

4 = Group administered paper and pencil test

57 Reliability

1 = reported in study

2 = estimated

58-59 Value of reliability correlation

60 Method of validation

1 = Assumed validity based on identification of content and method with original source

2 = Validation by panel of expert judges

3 = Correlation with results of method advocated by original source (e.g. Piaget)

4 = Construct validity (includes 1, 2 & 3)

61-62 Value of validity correlations, if any

COLS.

63 Position of age/level variable in study. If 5 or 6, list dependent under general characteristics of the study.

1 = blocking variable
2 = covariate
3 = independent variable
4 = mediating variable
5 = covariate and dependent
6 = independent and dependent

64=65 Reported developmental level of subjects

1 = concrete operational
2 = formal operational
3 = full range from concrete to formal
4 = preconventional moral stage
5 = conventional moral stage
6 = post-conventional moral stage
7 = 4 and 5
8 = 5 and 6
9 = 4, 5 and 6
10 = Not reported
11 =

66 Character of study

1 = non-interventionist, correlational
2 = interventionist (quality of instruction)
3 = interventionist (quantity of instruction)
4 = interventionist (motivation)
5 = interventionist (classroom environment)

67 Years difference between groups compared. (Needed if effect size/year is to be calculated) When computing effect size, older group is "experimental group." If same age, more formal group is "experimental group."

## II. Quality of Instruction Construct

Cols.

55 Experimental treatment applied to:

1= individuals
2= small group (2-6)
3= class size group (7-40)
4= large group (more than 90)

56 Control treatment applied to:

1= comparable size group
2= different size (more than ±5)

57 Experimental participation (if course is elective, participation in any part is considered elective unless otherwise specified)

1= elective (eg. high school physics and chemistry)
2= required (eg. most junior high science)
3= both elective and required options or unknown (eg. high school biology)

58 Control participation

1= comparable to experimental
2= different from experimental

59 Experimental group teachers

1= regular teacher (s)
2= special teacher (s)
3= materials only under study (no live teacher as part of independent variable)

60 Control group teacher

1= comparable to experimental
2= different
3= materials only under study

Cols.

61 Focus of instructional treatment. Primarily:

1= non-laboratory (students not working with apparatus)
2= both laboratory and non-laboratory instruction
3= laboratory only
4= other

62 Quality of instruction component under study:

1= curriculum, course or other global comparison
2= teacher behavior and materials (more controlled, better defined, usually shorter in duration than number 1)
3= teacher behavior only (basically same materials in all treatments)
4= materials only (no teacher actively involved, eg. CAI, TV A-T, programmed instruction)

63-64 Quality of instruction variable under study (incomplete listing):

Preinstructional strategy

1= advance organizer vs none or placebo
2= statement of objectives
4= set induction

Directness of instruction

15= direct (experimental) vs non-direct(control) instruction; teacher directed (exp) vs student self-directed (cont) instruction. In correlational studies, this is "teacher-directives" (explaining, lecturing, directing)

16= Indirect/direct ration(Flanders) Lower ID group is experimental group. "teacher-directness" is degree of <u>not</u> using discussion.

Instruction in processes and logical operations

25= training in processes of science
26= training in logical operations (reasoning patterns) a la Piaget

Strucure in verbal content of materials

24= kinetic structure(High= experimental group)

Inductive vs deductive strategies

20= inductive(control) vs deductive (experimental); inquiry (cont) vs expository (exp)

21= logical i.e. inductive and/or deductive vs random sequencing

22= expository ( lecture-discussion) vs laboratory (control)

70= inductive, inquiry based curriculum (many curriculum projects of 60's) vs traditional curriculum.

65 Method of obtaining observations/measures of variable under study:

1= self report
2= expert rating
3= student rating
4= expert co[illegible]
5= specialized [illegible]ing without classroom verification
6= predetermined in structure of materials
7= cannot be determined
8= both 5 and 6

66 Interobserver agreement

1= simple percent
2= other method (eg. Scott's coefficient)
3= not reported

67-68 Enter percent agreement value (leave blank if not reported)

69 Length of treatment

1= less than or equal to one hour
2= greater than one, less than ten hours
3= 10 to 50 hours
4= a course (10 weeks or more, about one hour per weekday)
5= cannot be determined or estimated

70 Covariates partialed out of effect size

1= none
2= ability (IQ, aptitude)
3= pretested knowledge and/or achievement
4= sociological variables (SES, classroom environment,etc)
5= psychological variables(motivation, personality)
6= 2 and 3
7= three or more of the above

Cols.

71 Control group access to treatment content

1= none
2= not comparable
3= comparable ( approximately equivalent)

## II. Quantity of Instruction Construct

| COLS. | |
|---|---|
| | Instructional group one values |
| 55-56 | Minutes per session |
| 57-58 | Sessions per week |
| 59-60 | Number of weeks |
| 61-62 | Number of years |
| 63-64 | Reported estimate or observed percent of time on task |
| | |
| | Instructional group two mean values |
| 65-66 | Minutes per session |
| 67-68 | Sessions per week |
| 69-70 | Number of weeks |
| 71-72 | Number of years |
| 73-74 | Reported estimate or observed percent of time on task |
| | |
| | Quantity of Instruction Variable Under Study |
| 75 | 1 = minutes per session |
| | 2 = number of minutes |
| | 3 = sessions per week |
| | 4 = number of sessions |
| | 5 = number of weeks |
| | 6 = number of years |
| 76 | Position of quantity variable in study. If 4 or 5, list dependent variable in column under general characteristics of the study |
| | 1 = covariate |
| | 2 = independent |
| | 3 = mediating |
| | 4 = covariate and dependent |
| | 5 = independent and dependent |
| | |
| 77 | Method of measuring quantity |
| | 1 = student self-report |
| | 2 = teacher report |
| | 3 = trained observer |
| | |
| 78 | Character of study |
| | 1 = non-interventionist, correlational |
| | 2 = interventionist, experimental |

## II. Social Environment of the Classroom

| COLS. | |
|---|---|
| | **Environment Measure** |
| 55 | 1= Learning Environment Inventory (LEI) |
| | 2= Modified LEI |
| | 3= My Class |
| | 4= Classroom Environment Scale (CES) |
| | 5= Learning Environment Inventory (1966 version) |
| | **Prior Achievement Controls (by subject area)** |
| 56 | 1= General Science |
| | 2= Life Science |
| | 3= Physcial Science |
| | 4= Mathematics |
| | 5= Social Science |
| | 6= Humanities |
| | 7= General Achievement |
| | 8= Attitude toward subject matter |
| | 9= Miscellaneous |
| | **Learning Environment Inventory Scale** |
| 57-58 | 01= Cohesiveness |
| | 02= Friction |
| | 03= Cliqueness |
| | 04= Satisfaction |
| | 05= Speed |
| | 06= Difficulty |
| | 07= Apathy |
| | 08= Favoritism |
| | 09= Formality |
| | 10= Goal Direction |
| | 11= Democracy |
| | 12= Disorganization |
| | 13= Diversity |
| | 14= Environment |
| | 15= Competition |
| | **Learning Outcome Domain** |
| 59 | 1= Cognitive |
| | 2= Attitudinal |
| | 3= Behavioral |

COLS.

| COLS. | |
|---|---|
| | Learning Outcome Content Area |
| 60 | 1= General Science<br>2= Life Science<br>3= Physical Science<br>4= Mathematics<br>5= Social Science<br>6= Humanities<br>7= General Achievement<br>8= Attitude toward Subject Matter<br>9= Miscellaneous |
| | Number of Classes in Study |
| 61-63 | (enter number of classes) |
| | Unit of Analysis |
| 64 | 1= Individual Student<br>2= Subgroups of Students<br>3= Classes<br>4= Schools |
| | Reliability of Social Environment Measure |
| 65-66 | (enter relaibility value; whether reported or estimated) |
| 67-68 | Reliability of Learning Outcome Measure<br>(enter reliability value; whether reported or estimated) |