DOCUMENT RESUME

ED 090 294 TM 003 579

AUTHOR	Hillman, Jason; Gowin, D. Bob
TITLE	Development of Course Content Materials For Training Research and Research Related Personnel to Appraise Research Critically. Final Report.
INSTITUTION	National Center for Educational Research and Development (DHEW/OE), Washington, D.C. Regional Research Program.
BUREAU NO	BR-0-9048
PUB DATE	Jun 73
CONTRACT	OEC-0-70-4775 (520)
NOTE	192p.
AVAILABLE FROM	Available from Prentice Hall Inc., Englewood Cliffs, N.J. as "Appraising Educational Research: A Case Study Approach". Publisher's price: $4.95
EDRS PRICE	MF-$0.75 HC-$9.00 PLUS POSTAGE
DESCRIPTORS	*Analytical Criticism; College Students; Content Analysis; Critical Reading; *Educational Researchers; Evaluation; *Evaluative Thinking; Graduate Students; *Instructional Materials; Interpretive Reading; Literature Reviews; Research Methodology; Research Problems; Research Skills; Technical Reports; Theoretical Criticism; *Training

ABSTRACT

A description of the development of the print materials to improve the ability of learners to appraise critically educational research is provided in this report. The completed materials consist of the following: an introductory statement about the nature of criticism, a statement about the contents of the materials and suggestions for use, and nine case studies. Most cases consist of a research article, special notes intended to make the article more comprehensible, orienting questions to guide the learner, a "model" appraisal (answers to the orienting questions), learner responses, and the product developers' replies to these responses. Specifically described in this report is the selection of case study materials, conduct of the two stages of field testing, evaluation of the drafts of the materials by student users, and bases for revision of materials. Reproduction of this document has been made from the best copy available. (Author)

Final Report

Project No. 0-9048
Contract No. OEC-0-70-4775(520)

Jason Millman and D. Bob Gowin
Department of Education
Cornell University
Ithaca, New York 14850

DEVELOPMENT OF COURSE CONTENT MATERIALS FOR TRAINING RESEARCH AND RESEARCH RELATED PERSONNEL TO APPRAISE RESEARCH CRITICALLY

June 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development
(Regional Research Program)

AUTHORS' ABSTRACT

A description of the development of the print materials to improve the ability of learners to appraise critically educational research is provided in this report. The completed materials consist of the following: an introductory statement about the nature of criticism, a statement about the contents of the materials and suggestions for use, and nine case studies. Most cases consist of a research article, special notes intended to make the article more comprehendible, orienting questions to guide the learner, a "model" appraisal (answers to the orienting questions), learner responses, and the product developers' replies to these responses.

Specifically described in this report is the selection of case study materials, conduct of the two stages of field testing, evaluation of the drafts of the materials by student users, and bases for revision of materials.

Final Report

Project No. 0-9048
Contract No. OEC-0-70-4775(520)

DEVELOPMENT OF COURSE CONTENT MATERIALS FOR TRAINING RESEARCH AND RESEARCH RELATED PERSONNEL TO APPRAISE RESEARCH CRITICALLY

Jason Millman and D. Bob Gowin

Department of Education
Cornell University

Ithaca, New York

June 1973

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

PREFACE

We wish to acknowledge the assistance of the several institutions and individuals named herein who cooperated in the development and field testing of these instructional materials. We further wish to thank Prentice-Hall for their willingness to make these materials commercially available at low cost. Finally, a special thanks goes to Ms. Dorothy Pasternack who served as technical editor and administrative assistant during most of the lifetime of the project.

Jason Millman and D. Bob Gowin
Ithaca, New York

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Explicit critical appraisal of research products has been a missing element in the sequence of efforts which transforms unknowns into knowns and knowns into practical usefulness. It was our purpose to develop materials to help train research and research related personnel to appraise research critically.

Our strategy was to produce materials suitable for use either in conjunction with a research methods course or separately and of interest and value to learners regardless of their level of sophistication and field of interest within education. The articles chosen as case studies, therefore, require neither statistical sophistication nor expertise in the substantive field in order to be comprehensible. Further, the materials were given a degree of responsiveness by printing and responding to learner responses most often encountered during field tryouts.

The procedures by which the materials were developed are described in the Methods section of this report. Some student evaluations are provided under Results. The actual materials have been reproduced in Appendices I through X.

METHOD

Surveying Existing Guides. Over 30 published sets of guides for evaluating the research of others were identified. Although may of these sets provided excellent lists of aspects to consider in evaluating the research of others (e.g., educational significance), they had the common failing of not indicating with actual examples the criteria for judging whether an example of educational research actually contained to a satisfactory degree the characteristics deemed important. Thus, the decision was made to follow the procedures used in this project; namely, to guide the learner through a critical evaluation of specific examples of educational research.

Selecting Articles to be Critically Analyzed. It had been our hope to use articles which were deemed as important examples of educational research by at least one of several audiences of educational research products. On October 1, 1970 we solicited nominations of such research work from the following groups of individuals:

(a) 18 directors of reading programs at the graduate level selected, using a systematic sampling plan, from a list published in: Robert M. Wilson, Colleges and Universities Offering Programs in Reading, The Journal of the Reading Specialist, December, 1967, pp. 66-87.

(b) 38 members of the American School Counselor Association who did not have a university or college address, selected, using a systematic sampling plan, from the 1967-68 Directory of Members published by the American School Counselor Association.

(c) 66 superintendents of schools selected, using a systematic sampling plan, from the 1969 Roster of Members published by the American Association of School Administrators.

(d) 65 individuals listed as "Additional Members through April 26, 1968" of the American Educational Research Association Special Interest Group titled: Professors of Educational Research.

Only 15 suggestions of significant research were obtained from the 187 individuals listed in (a) through (d) above. Most of these suggestions were not used because they were not seen as research, they were too lengthy for the instructional use we had in mind, or they were judged as having so little value that the ensuing critique would be markedly unbalanced in the negative direction.

In addition, the editors of the seven books of Readings on Educational Research which comprise the American Educational Research Association (AERA) published series (except the editor of the readings on research methodology), were written and requested to send us the table of contents of their individual books. It was felt that such lists would represent significant educational research as viewed by the committee of research scholars charged with the responsibility of selecting the entries for each of these books of readings.

Three tables of contents of the AERA books of readings were obtained and one article was selected from these lists. Because of our desire to have the final selections represent a cross-section of areas of research as well as of research methodologies, we decided against making multiple selections from any given book of readings.

The remaining articles were selected by the Project Directors. A list of the 10 articles for which initial drafts of training materials were written is presented below:

1. Edwin M. Bridges, Wayne J. Doyle and David J. Mahan, "Effects of Hierarchical Differentiation on Group Productivity, Efficiency, and Risk Taking." Administrative Science Quarterly, September 1968, pp. 305-319.

2. Joanne Reynolds Bronars, "Tampering with Nature in Elementary School Science." The Educational Forum, November 1968, pp. 71-75.

3. Dolores Durkin, "Children's Concepts of Justice: A Comparison with the Piaget Data." Child Development, 1969, pp. 59-67.

4. David Elkind, Joann Deblinger and David Adler, "Motivation and Creativity: The Context Effect." American Educational Research Journal, May 1970, pp. 351-357.

5. J. Richard Hackman, Nancy Wiggins and Alan R. Bass, "Prediction of Long-Term Success in Doctoral Work in Psychology." Educational and Psychological Measurement, Summer 1970, pp. 365-374.

6. Florence R. Harris, Margaret K. Johnston, C. Susan Kelley, and Montrose M. Wolf, "Effects of Positive Reinforcement on Regressed Crawling of a Nursery School Child." Journal of Educational Psychology, February 1964, pp. 35-41.

7. Francis P. Hunkins, "The Influence of Analysis and Evaluation Questions on Achievement in Sixth Grade Social Studies." Educational Leadership, January 1968, pp. 326-332.

8. Charles H. Josephson, "Do Grades Stimulate Students to Failure?" Chicago Schools Journal, December 1961, pp. 122-127.

9. Eleanor Kaplan, "'Head Start' Experience and the Development of Skills and Abilities in Kindergarten Children." Graduate Research in Education and Related Disciplines, April 1966, pp. 4-28.

10. Henry R. Weinstock and Charles M. Peccolo, "Do Students' Ideas and Attitudes Survive Practice Teaching?" Elementary School Journal, January 1970, pp. 210-218.

Obtaining Copyright Release. The appraisal materials would have little value if the original article being critiqued could not be made available at the same time. The Project Directors were unsuccessful in locating Eleanor Kaplan. However, her article had not been copyrighted. Dolores Durkin and Henry Weinstock refused to give permission to have their articles reproduced. Permission to reproduce the other seven articles was received.

Because of the severe time constraints necessitated by the requirement to schedule two field tryouts within a single semester, the Project Directors commenced preparation of instructional materials to accompany the Durkin and Weinstock articles before negotiating for a copyright release. Had the Project Directors realized that such releases would not be forthcoming, they would have substituted other articles at the outset.

Nevertheless, the Weinstock article was dropped because it did not seem to work out well. The Project Directors persisted in refining the materials related to the Durkin article because technically only the journal's permission (which was granted) was required and because nine critiques were contracted with the Federal government.

Obtaining Reviews of Experts. Approximately two dozen scholars were invited to prepare a critical appraisal of one of the articles which matched their area of expertise. The names of the individuals who prepared such reviews and the articles they reviewed are listed below:

BRIDGES ARTICLE	:	Dr. W. W. Charters, Jr., University of Oregon Dr. Robert Ennis, University of Illinois
BRONARS ARTICLE	:	Dr. John Easley, University of Illinois Dr. Jonas Soltis, Columbia University
DURKIN ARTICLE	:	Dr. Alfred Baldwin, Cornell University Dr. Brian Crittenden, Ontario Institute for Studies in Education
ELKIND ARTICLE	:	Dr. J. P. Guilford, Beverly Hills, California Dr. Kenneth Strike, University of Wisconsin
HARRIS ARTICLE	:	Dr. Alberta Siegal, Stanford University Dr. Harold Stevenson, University of Minnesota
HACKMAN ARTICLE	:	Dr. Leonard Krimerman, University of Connecticut
HUNKINS ARTICLE	:	Dr. Kenneth Hopkins, University of Colorado Dr. William Lowe, University of Rochester
JOSEPHSON ARTICLE	:	Dr. David Farr, State University of New York at Buffalo Dr. John Milholland, University of Michigan
KAPLAN ARTICLE	:	Dr. Gene Glass, University of Colorado
WEINSTOCK ARTICLE	:	Dr. George Newsome, Jr., University of Georgia Dr. William Gephart, Phi Delta Kappa

Preparing Initial Drafts of the Instructional Materials. Armed with the reviews of the experts as well as reviews written by students at Cornell University, the Project Directors prepared initial drafts of instructional materials to accompany each of the ten articles.

No single format was adopted but rather the materials were developed in a way which seemed most appropriate for the article in question. In this first draft phase, students were given instructions to first read the article to be evaluated. In most cases these initial instructions were followed by a section of "Special Notes" which explained terms or procedures likely to be unfamiliar to education students (e.g., use of various statistical measures, definition of unusual terms, explanation of previous relevant literature). Following these "Special Notes" the materials requested students to complete a written assignment which required focusing critically upon the article. In some cases the assignment required an overall general appraisal in which students were to cite strengths as well as weaknesses; in some cases specific questions about the article were devised which students were then to answer; in a few cases both kinds of assessments were requested. Of course in all these cases, the assignments were developed with an eye to having the student consider and evaluate all aspects of a given research paper.

"Model" answers to the various requests for general appraisal or sets of specific questions was the last section of these written materials included in the first draft. These answers were made fairly detailed with the idea that students would compare their answers to the "model" answers provided and thereby obtain valuable instruction in how to approach educational research papers.

Securing Field Tryout Populations. It was felt essential for a high quality product that the materials be field tested at least twice; once after the initial draft and once after a revised draft. At the time that the AERA Special Interest Group was solicited for suggestions of research articles, a request was made to use students in their research methods classes as a tryout population. Fourteen responses were received. Since this number was felt to be inadequate for both tryout phases, an additional 12 instructors willing to participate in the testing of the materials were recruited at the annual National Symposium of Professors of Educational Research held in November 1970 in St. Louis. An additional three instructors who were personal friends of a Project Director agreed to cooperate. All materials were also field tested in classes at Cornell University.

Conducting First Field Testing. During February and March of 1971 the initial drafts of the materials were sent to twelve institutions. The names of the institutions and the number of

student responses returned are listed below:

		Returned
BRIDGES	: C. W. Post	17
	St. Louis University	13
	University of New Mexico	5
BRONARS	: Stanford University	0
	University of Wisconsin	0
	University of Maryland	31
	Kansas State	8
	University of Nevada at Reno	0
DURKIN	: University of Maryland	18
	Kansas State University	8
	University of Nevada at Reno	0
	Stanford University	2
ELKIND	: University of Washington	14
	University of Maryland	27
	University of Nevada at Reno	2
	Stanford University	1
HACKMAN	: C. W. Post	37
	Purdue University	11
	University of Wisconsin	0
HARRIS	: C. W. Post	36
	St. Louis University	23
	University of Wisconsin	2
HUNKINS	: C. W. Post	10
	St. Louis University	8
	University of New Mexico	7
	University of Wisconsin	1
JOSEPHSON	: St. Louis University	13
	Towson State	30
	University of Wisconsin	1
KAPLAN	: St. Louis University	8
	Towson State	32
	Purdue University	9
WEINSTOCK	: Towson State	32

The instructors were asked to withhold distribution of the "model" answers until written responses of their students were received. The reason for this request was to obtain reactions of stu-

dents which were uncontaminated by these "model" answers and which could serve as data upon which to base future revisions of the materials.

Preparing Second Drafts of the Instructional Materials. Student responses to the initial drafts were carefully reviewed before developing second drafts of the materials. These responses were most helpful in indicating (a) where there were ambiguities in the questions and "model" answers; (b) where it seemed necessary to cue students more specifically to the desired focus; and (c) where it seemed wise to offer further supplementary explanation in certain areas. Not only were the materials revised with an eye to clarification and amplification, but in the cases of the Bridges and Elkind articles, additional sections were developed, Student Responses to Question and Our Replies, to handle the many points which student responses indicated needed explanation.

Conducting Second Field Testing. During April and May of 1971* the second drafts of the materials were sent to the institutions listed below:

			Returned
BRIDGES	:	University of Colorado	9
		Catholic University of America	8
		Montclair State College	0
		Arizona State University	9
BRONARS	:	Catholic University of America	9
		University of Northern Iowa	28
		George Washington University	15
		Arizona State University	10
		University of Southwestern Louisiana	2
DURKIN	:	William and Mary	22
		Catholic University of America	0
		University of Southwestern Louisiana	3
		George Washington University	15
		University of Bridgeport	0
ELKIND	:	Eastern Kentucky University	14
		University of Northern Iowa	59
		University of Bridgeport	0
		Arizona State University	15
		University of Southwestern Louisiana	2
HACKMAN	:	University of Colorado	20
		Montclair State College	11
		University of Louisville	21
		Arizona State University	0
		University of Southwestern Louisiana	1

*The University of Southwestern Louisiana and Texas Tech were sent the materials during July 1971.

		Returned
HARRIS :	University of Southern California	0
	University of Wisconsin at Milwaukee	16
	Ohio State	11
	University of New Mexico	0
	University of Southwestern Louisiana	2
HUNKINS :	University of Georgia	13
	Pennsylvania State University	2
	Creighton University	16
	University of Wisconsin at Milwaukee	11
	Ohio State	11
	University of Southwestern Louisiana	2
KAPLAN :	University of Southwestern Louisiana	2
	Texas Tech	1
	Eastern Kentucky University	14
	University of Northern Iowa	28
	George Washington University	16
JOSEPHSON :	University of Southern California	0
	Creighton University	13
	University of Wisconsin at Milwaukee	5
	Ohio State	16
	University of Southwestern Louisiana	2
WEINSTOCK :	University of Southern California	0
	University of Wisconsin at Milwaukee	6
	Ohio State	13
	Texas Tech	1
	University of Southwestern Louisiana	1

In this second draft phase students were asked not only to complete the written assignments, but also to evaluate our instructional materials as well. Specifically, they were asked to compare their answers to the "model" answers provided, indicating where they felt these answers were ambiguous, incomplete or in error. They were further asked for a general evaluation of the article and its accompanying materials. In this regard, students were asked to indicate whether they thought the article and the materials were interesting, dealing with a topic important to the field; whether the time spent was worthwhile; and whether the materials were indeed self instructional. Naturally the responses varied from paper to paper and a discussion of these responses will be included in the results section of this report.

Soliciting Authors' Responses to the Materials. A copy

of the second draft materials developed for his or her paper was sent for comments to eight senior authors who could be located. Responses were received from all eight.

Soliciting Experts' Comments on the Materials. A copy of the second draft materials developed for each article was also sent to the appropriate experts who had prepared an initial review of these articles. Each expert was asked to read the instructional materials and to indicate where he felt we had made serious errors or omissions in our interpretation of the article. Virtually no criticism was received.

Summer Field Test. In addition to the other schools listed in this report, two copies of each of the papers and instructional materials were sent to the University of Southwestern Louisiana and Texas Technical College during July 1971. A total of 19 papers were returned.

Preparing Third Drafts. Using all the materials gathered during the first 10 months of this project, final drafts of the materials were developed during the fall of 1971. In at least three of the case studies, the authors' own reactions were incorporated as direct quotes in the model answers. No third draft of the Weinstock article was prepared.

Securing a Commerical Publisher. During 1972, negotiation with Prentice-Hall to publish the materials was successfully completed.

Preparing the Final Draft. During February 1973, two orienting chapters were written to facilitate use of all the articles as a collection. Based on comments received by Prentice-Hall editor, Gene V Glass, minor changes were made to the orienting chapters and to all third draft materials with the exception of the Durkin article. It is expected that the two orienting chapters and eight case studies (Durkin and Weinstock articles excluded) will be published as a paperback during spring 1974.

RESULTS

Enclosed in Appendices I through X are the third drafts of the instructional materials to accompany nine case studies (Weinstock and Peccolo article excluded) and the two orienting chapters. These, of course, are the principal results of this instructional material development project, they are the terminal contracted product. The responses of students to earlier drafts of the materials were the primary data which stimulated revision of the materials.

Content Revisions of the Materials. The revisions in content made from the first to last drafts of the materials can be categorized as follows.

1. Questions were designed to provide more adequately cues for the desired response. Early in the development effort, it became obvious to the Project Directors (i.e., project developers) that the users of the materials were interpreting the questions differently from initial intention. Students would complain that their answers differed from the model appraisals, not because they weren't able to say the kinds of things given as answers, but because they didn't realize what was wanted.

The most frequent change to questions was to add to questions which could be answered solely by "yes" or "no" such phrases as, "explain why you answered as you did," "state reasons for your answer," and "why?". In many cases, the questions were made longer in order to communicate more clearly the intended direction the answers should take. Words were defined, limits set, orienting statements made, and cautions about possible misinterpretations provided.

2. Reasons for dogmatic-sounding statements were provided. Frequently the model answers contained statements which to the Project Directors were simple statements of fact but which were challenged by the student readers as dogmatic. In such cases, reasons were added to support the claim being made.

3. Inferences about feelings were reduced. Earlier versions of the materials contained many inferences regarding the feelings of the research investigators. Student readers were critical of the Project Directors' willingness to state how other people, namely the research investigators, felt. Students were also critical of some of the personal opinions of the Project Directors. Although these feelings were not

always eliminated, reasons were provided in support of the convictions being expressed.

4. Greater explanation was provided. The student answers to earlier drafts of the materials highlighted sections of the materials where greater explanation was needed. Hard-to-understand concepts were clarified.

5. Questions and sections in the Model Answers were eliminated. Reactions such as "dumb question" convinced the Project Directors that some material was less important than others. Further, other questions and comments were so hard or technical that few students were able to answer them or understand the explanations given. In most of such instances, the material was eliminated.

6. Mistakes were corrected. Errors in the earlier drafts ranged from simple typographical mistakes to a few real blunders on the part of the Project Directors. In addition to revisions resulting from such clear-cut errors, a very large number of changes were made not because one wording was right and the other wrong, but because one wording was more appropriate than the other. For example, in the Elkind et al. article, the issue was raised whether the research instruments were valid measures of the construct, "creativity". The original wording was, "...creativity is not really the dependent variable,..." and this was changed to "...creativity is not really measured by the tests,..."

7. Different expressions were used to express the same ideas. Very often the student answers were better expressed than those provided in the Model. In such cases, such wording was substituted. Further, comments of the research investigators themselves were sometimes substituted for similar comments made by the Project Directors because users of the materials expressed an interest in knowing how the Research Investigators felt.

8. Comments were qualified. Blanket statements were often altered for the sake of accuracy. For example, "There is no..." was changed to "We know of no..." and "should" to "might".

9. Student response sections were created. Frequently occurring or interesting comments of the student users were added to the model appraisals together with replies from the Project Directors. These sections were very well received by later users.

Finally, it should be pointed out that the eight points listed on pages 14-16 in Appendix I resulted directly from comments made by the users of the initial drafts of the materials.

General Evaluation of the Materials by Students. During the second field trials, students were asked the following:

> Please give us a general evaluation of these materials. Specifically, comment on:
>
> (a) how interesting you found the article;
> (b) whether you felt the time you spent working on these materials was worthwhile;
> (c) whether you can think of other aspects of the study you wished we had commented upon;
> (d) whether you think we were too hard or too easy on the investigator (can you give specifics?); and
> (e) any other comments you might wish to make.

A summary of the students' reactions follow.

Presented in Table 1 is a tally of the perceived interest-producing quality of the articles used as case study material. As can be seen in Table 1, most respondents found the articles interesting or very interesting. Less than 10% rated the articles as not interesting. This trend appears for each of the nine articles.

The 31 responses coded, "Irrelevant Comment" dealt with other than the interest-producing qualities of the articles. Examples of comments included in this category are: "enjoyed reading it," "thought provoking," "too confusing," and "very informative."*

The assessment materials were judged to be worthwhile or very worthwhile by 207 respondents as shown in Table 2.

*Responses of students from George Washington University to the Bronars and Durkin article were inadvertently omitted from Tables 1-3.

Table 1

Student Reactions to How Interesting They Found the Article

Senior Author	Response Category: Very Interesting	Interesting	Not Interesting	Irrelevant Comment	No Comment
Bridges	6	11	4	3	2
Bronars	17	16	2	2	12
Durkin	11	9	3	0	2
Elkind	23	29	8	5	25
Hackman	12	22	7	3	9
Harris	6	16	2	3	2
Hunkins	10	20	6	5	14
Josephson	14	11	1	4	6
Kaplan	20	25	4	6	6
Total	119	159	37	31	78

Although some differences exist, the balance of worthwhile over not worthwhile assessments is maintained for each article. The materials related to the article by Kaplan were especially well received.

Table 2

Student Reactions of the Worthwhileness of the Materials (Second Draft)

Senior Author	Very Worthwhile	Worthwhile	Not Worthwhile	Irrelevant Comment	No Comment
Bridges	2	11	5	3	5
Bronars	4	17	7	11	10
Durkin	4	8	4	8	1
Elkind	5	34	13	15	23
Hackman	2	21	11	6	13
Harris	2	11	6	8	2
Hunkins	5	20	7	6	17
Josephson	4	15	5	3	9
Kaplan	7	35	2	8	9
Total	35	172	60	68	89

Comments listed as "irrelevant" to the worthwhileness characteristic included the following: "did not mind doing this," "it produced food for thought," "took too long but was a good expose," "it was informative," "the article was worthwhile," "did not understand the purpose," "I enjoyed the experience," and "the experience was certainly educational."

It should be kept in mind that different colleges are represented for the several articles. The reception that a particular article received seemed to depend in part on how it was introduced by the instructor. As will be indicated shortly, several students did not understand the purpose of the exercises and resented the time it took away from "required" course work.

The third general evaluation question asked whether the student could think of other aspects of the study they wished we had commented upon. Some of the answers to these questions were useful in preparing the third drafts of the materials. Since they dealt with content specific to the individual article, they are not summarized here.

Table 3 contains a summary of responses to the query

Table 3

Student Reactions to the Fairness of the Criticism (Second Draft)

Senior Author	Too Easy	Too Hard	Neither	Irrelevant*	No Comment
Bridges	1	1	9	4	11
Bronars	6	5	13	6(2)	19
Durkin	0	7	7	7(4)	4
Elkind	3	10	18	29(10)	30
Hackman	6	3	16	12(5)	16
Harris	5	4	11	4	5
Hunkins	2	6	17	6(2)	24
Josephson	5	1	14	2	14
Kaplan	5	11	20	11(5)	14
Total	33	48	125	81(28)	137

*Values in parentheses indicate the number of responses coded irrelevant which suggested the assessment was either not too hard or the criticisms were well supported.

whether we were too hard or too easy on the investigators. Although the question was worded in a forced-choice format, we were pleased 125 respondents answered neither too easy nor too hard and that those who picked a direction were roughly split between the too easy and too hard poles.

There were some differences among articles in ways which we could have predicted. Our assessment was seen as too easy for the Josephson article, and this was somewhat intentional for two reasons. As the anticipated lead case study in the published collection, we wanted to impress upon readers that all assessments need not be negative. Further, although the outward appearance of the article was that it was terribly naive (there were some glaring weaknesses), we wished to stress some basic strengths which we expected would be overlooked.

The Durkin review was seen as too harsh. This reaction was not unexpected either, because although the article looked sophisticated on the surface we had some serious questions about the educational significance of the research being reported. It is possible that the author herself felt that we were being too harsh because she refused to give approval for us to reprint the article.

Comments coded "irrelevant" to the question included, "analysis good," "approach highly professional," and "I generally agree with the comments." As indicated in the footnote of Table 3, 28 students answered that we were not too hard or that our criticisms were well supported. Some readers of this report may prefer to consider these 28 responses under the heading "neither".

The last general evaluation question merely asked the respondents to make any other comments they wished. These comments could be grouped into several categories.

First, there were many noncontent related negative objections. Many students saw the assignment as an infringement on their time, especially scarce since the school year was almost over. Other students were not clear on the purpose of the assignment, and were we to do it over, we would have given much more explanation on this matter. Further, the article, the special notes, the questions, the model answers, and the student responses and answers were usually distributed separately. A large number of students reacted negatively to this paper shuffling chore. In the published version, all materials will be bound together.

A second general category of responses were negative content related comments dealing most often with the long amount of time needed to study and critique carefully a particular article. Other negative comments dealt with the article itself (not in the field of interest of the person), questioned the importance of the appraisal activity, or stated the materials were too hard or confusing.

A third group of comments were positive in which improved skills and quality of the materials were most frequently mentioned. Of the 226 comments, 35% were in this category; 50% in categories one of two; 7% in both the positive and one of the two negative categories; and 8% in none of the three categories.

CONCLUSIONS

A set of nine case studies and two orienting chapters were prepared. The materials underwent marked revisions as a result of the field tryouts. As gleaned from the student reactions, the articles were judged interesting and the appraisal materials worthwhile. The Project Directors are left with three salient impressions about the development effort and the materials.

First, producing the appraisals was very hard work. It taxed the scholarship of the Project Directors greatly. Contrary to expectation, it was found that little of the work could be delegated and, consequently, the Project Directors had to assume responsibilities previously thought would be assumed by bright graduate students.

Second, general principles of research appraisal did not emerge. Each article seemed to generate its own unique points of criticism. The Project Directors are convinced now more than ever that checklists for research appraisal must, by their very nature, be too shallow to provide the depth of assessment evident in the present materials. The important tools for the successful critic would appear to be strategies for handling the appraisal task and subject matter and methodology content. The true value of the case studies may be both in reinforcing requisite "habits of workmanship"--strategies like reading carefully, perceiving the compromises in design, searching for significance, etc.--and in providing concepts and facts for the learner to use.

Third, trainers of research workers familiar with the materials have found them appropriate for their own teaching. For example, students at Cornell who now have college teaching positions are employing the materials in their classes; the Prentice-Hall editor who provided the expert review has adopted the materials for his class.

There remains the question whether the instructional materials actually improve the appraisal skills of the users and, in the long run, has impact on educational knowledge, policy and practice. Although not part of the present contract, the Project Directors hope that such terminal evaluations will be forthcoming. By putting these materials in the public domain, both as part of this report and in a commercially distributed version, others will have an opportunity to evaluate the product in terms of their own concerns and standards.

Millman
Gowin
0-9048
OEC-0-70-4775

Appendix I

Preface

A surprising fact to us is that the tradition of critical appraisal is so largely missing in the context of educational research. Very little good criticism of educational research occurs. Why?

Perhaps it is a matter of assumed senatorial courtesy or that the best criticism of other research is simply doing a superior piece of research, or that since all in education are dedicated good people one should not be critical, or that a sharp-eyed critic is a dangerous fellow because he will embarass a naive or foolish empiricist, or possibly that research is done to achieve tenure, promotion, increase in slary, prestige, esteem, more grants, etc. These reasons and others much gossiped about at research meetings do not really concern us. We think both research and criticism are matters that intelligent students can be expertly trained to do well, and we see no reason not to try to improve present practices. One need not be afraid of criticism.

Very little good material is available as instruction in criticism. We do think that good criticism is needed in education and this fact has led us to put forth the effort recorded here.

This book has a rare characteristic; the nine critiques which comprise the main contribution have been extensively field tested. That is, the critiques have been developed and modified on the basis of comments supplied by over 800 students from 27 colleges and universities. Indispensible to us were the reactions of subject matter experts and of students participating in the successive field tryouts. (Point A - See p. 3). The cycle of field test, modify, field test, modify...permitted the materials to achieve a level of quality not possible otherwise.

Acknowledgment of the help of several groups are in order. Specifically, we owe much to:

1. the students and cooperating professors of the following institutions: Arizona State University, C. W. Post, Catholic University of America, Cornell University, Creighton University, Eastern Kentucky University, George Washington University, Kansas State University, Montclair State College, Ohio State University, Pennsylvania State University, Purdue University, St. Louis University, Stanford University, Towson State, University of Colorado, University of Georgia, University of Louisville, University of Maryland, University of Nevada at Reno, University of New Mexico, University of Northern Iowa, University of S. W. Louisiana, University of Washington, University of Wisconsin at Madison, University of Wisconsin at Milwaukee, and William and Mary.

2. the following scholars who supplied us with an initial reaction to one of the articles: David Farr and John Millholland (Chapter 3), Gene Glass (Chapter 4), Leonard Krimerman (Chapter 5), Kenneth Hopkins and William Lowe (Chapter 6), Alfred Baldwin and Brian Crittenden (Chapter 7), Alberta Siegal and Harold Stevenson (Chapter 8), J. P. Guilford and Kenneth Strike (Chapter 9), John Easley and Jonas Soltis (Chapter 10), and W. W. Charters, Jr. and Robert Ennis (Chapter 11).[1]

3. the following professionally-minded investigators who offered constructive reactions to our critiques of or further information about their research articles: Edwin M. Bridges, Joanne Reynolds Bronars, Wayne Doyle, David Elkind, J. Richard Hackman, Charles H. Josephson and David Hahan.[1]

4. the publishers and investigators who were willing to grant permission to reproduce their articles despite the presence of negative comments.

5. the United States Office of Education which provided the financial support for development of these materials.

6. Prentice Hall for making it possible for these training materials to be disseminated.

[1]Acknowledgment of the assistance of the scholars and investigators names above should not be construed to mean that they approve of all aspects of our appraisals.

CHAPTER 1
The Nature of Criticism

There is a sense in which the critical appraisal of empirical research papers is also an act of research. It is an act of research because the critic reviews each of the aspects of the research paper very much in the same way as the original author considers aspects of the research paper. The key elements in the pattern of inquiry are the same for both the doing of research and the doing of criticism. In each case one must take a look at these elements: the nature of the problem, the phenomena of interest, the telling question, the key concepts, the methods of work, the knowledge claims and other products of the research effort, and the value or significance of the research.

The act of critical appraisal is a process of analysis, of breaking down and taking apart, what was produced by an act of synthesis by the original author(s). There is another pair of eyes, another mind, another point of view about the research. Specific training in criticism will in the long run enhance the fertility of actual research.

Each element in the pattern of inquiry requires the investigator to select, arrange, modify, and interpret. This process requires judgments. For example, the selection of a phenomena of interest and from that the setting up of a problem involve a judgment that these aspects of the world of experience are worth inquiring into, implicitly rejecting other concerns that might be worked on. The precise form of the telling question is a judgment that this question and not some other question will enable the researcher to find out something of importance. The use of one set of key concepts to ask the telling question means that other concepts have been thought about and rejected for the time being. The research design, the

selection of specific techniques of data gathering, statistical analysis, the construction of tables and graphs and other ways of presenting the record of the research effort, employ the judgment that these methods are better than others that might be used in this case. Finally, the particular knowledge claims selected as the important ones, the conclusions that are interpreted by the researchers, signify yet another set of (complex) judgments about what is worth reporting and what is seen as having value to other researchers.

The main point to be made here is that the categories of critical appraisal are basically no different from the categories of actual inquiry. The critic should question the judgments made at each stage of the pattern of inquiry. Specifically, he should ask: What other phenomena of interest might be relevant? What other way to pose the problem could be thought of? What different concepts or conceptual systems might have been used? What alternative designs or methods or techniques for data gathering could have been considered? What limits to generalization are found in the particular way the research is reported? What other values might conceivably be found in this research? And critical appraisal, like worthwhile research, depends heavily upon human judgment.

Three Purposes of Criticism

First, to the extent that research is an attempt to establish the fundamental and foundational knowledge claims about education, criticism is the attempt to apply the best human thought to test these foundations. Whether the research effort is directed at aptitude testing, behavior modification, organizational change, instructional material development, nothing of consequence follows if the research is faulty. A science builds upon its foundations, and confidence is a result of a tested faith in those foundations. Further, because research is open ended, criticism can point

to avenues of additional research needed to solidify our foundations of knowledge.

The second aim of criticism concerns policy-making and implementation. Policies are complex judgments, based partly upon facts and knowledge claims and partly upon values and value judgments. Policies are plans for action. To educate, to intervene in the lives of other human beings are serious moral undertakings. If a lack of knowledge is allowed to persist where knowledge could be obtained, the policy made and the action undertaken are grossly negligent of concern for the moral worth of other people. Criticism has a special role in policy analysis because it makes explicit this relation between knowledge and value found in educational policies.

The third aim of criticism concerns educational practice. In spite of rhetorical claims to the contrary, research has had little effect upon educational practice. Because there is always the potentiality that research will be conceived so as to change practice, criticism must obtain here too. The taking of thought to improve practice can lead to finding out facts, to discovering relations, to solving problems, to dispelling the comforting but misleading conventional wisdom. Criticism can be applied directly to the problems of justifying educational practice, but it ties up with research when it suggests the role of research in making practice more efficient, more effective, more humane, more insightful in its complex operation.

Criticism and Literary Criticism

We find it useful to borrow from the field of literary criticism a set of distinctions we think apply to criticism of educational research. Literary critics distinguish aspects of criticism into four elements: the author (or artist), the work, the audience, and the universe.' These distinctions are useful because we find that importantly different criteria

of assessment apply to different elements. For example, when we evaluate research we can begin a criticism by checking the authority of the author and we give the reasons for saying that the author or authors are experts in the area of research. Individuals with a history of high quality research justifiably deserve our attention because they have over the course of years earned the label of expert. Experts are in a sense highly calibrated instruments; we trust their "readings," the points they make. Of course, any person is fallible; experts have their off day, busy people make mistakes and so on. Nevertheless experts continue to deserve the label as they continue to employ high standards for their work.

Many judges of research papers (editors of journals for example) make a practice of not knowing the name of the author. This practice is one way to force attention to the work itself. Criteria of excellence commonly applied to individual works are very familiar: coherence of the reasoning from the problem statement to the conclusion, justification of the significance of the problem in the context in which it is placed, elegance of the design, choice of techniques of measurement, completeness of analysis, originality or novelty or creativity (breaking new ground), generation of new paradigms as well as connection to older paradigms to supply continuity with previous research.

Literary critics also judge the value of a work of art by the effects it has upon an appropriate audience; does it entertain, edify, point a moral, stimulate applause? Research products are also judged for their contribution to individuals who use the research products.

Does the set of knowledge claims of the research report stimulate consideration of educational changes? There exists the balancing of judgment between research that is socially relevant, that solves or contributes to the solution of an immediate social problem--that, versus research

for which no socially relevant consideration is relevant because the research contributes to the furthering of scientific knowledge which at the time does not seem to have any social relevance. This comparison is sometimes referred to as the scientist riding a white horse (to change society) versus the scientist wearing his white coat (to contribute to scientific knowledge).

The process of education is necessarily social. And the conservation and continuity of a social order necessarily requires education. Every adult (indeed every human who acquires a language) is educated in a social context, whether through formal schooling or not. In a common sense way every person knows something about education. This common sense knowledge, or conventional wisdom of the audience, often stands in the way of establishing scientific knowledge.

The fourth element for the focus of criticism is called by literary critics "the universe." The term we have used for this element in these materials is the "phenomena of interest." We have in mind here the "stuff," the subject-matter, the kind of thing the research is about. For example, in one of the studies in this book the authors are concerned with productivity of groups as it relates to the structure of the group. These phenomena are of considerable interest to school principals, industrial managers, and others for the reason that adequately anchored knowledge claims could provide a valuable guide for the administrator. In this regard the research would be judged as potentially significant; we say potentially significant here because the significance of such studies are not achieved by what it is about, but by what it tells us of what it is about. In other words, the phenomena of interest can be very important and the research relatively trivial if it fails to penetrate into the phenomena in any successful way.

As suggestive and fruitful as the model for criticism that the literary critics use, (and we only sketch it here) it has some shortcomings as well. A chief shortcoming is the lack of focus upon methods of work. We do not feel that we are going to mislead an audience of educational researchers, however, because the omnipresent focus of criticism found in contemporary educational research is precisely a concern with methods, with techniques of work, with research design, with statistical analysis.

Old Chestnuts in Dispute

Sometimes it is held that research is creative, that research generates new knowledge about the way the world works. On this ground research is distinguished from scholarship (sometimes called library research) which only puts together or comments on knowledge which others have produced. So, on this ground the products of criticism can be called scholarship. Whatever the label agreed on, the relation between research and scholarship can be very close. Criticism which reveals faults in purported knowledge claims is both creative and valuable. Moreover, as indicated previously, both research and criticism (scholarship) require judgments about the same processes.

We have learned much from the college students who helped by using early drafts of these critical appraisals. One thing which many students reported had inhibited their own appraisal was the lack of knowledge about statistics. We urge students, and other critics, to become knowledgeable about statistics, but we also recommend that one not be too easily blinded by statistics. A kind of mindless reverence for numbers, tests of coefficients, F ratios and the like is to be avoided. One can still use judgment to see whether the data analyzed actually relate in a satisfactory way to the basic question, and how useful the data are in

composing an answer. Any complex statistical analysis can be paraphrased in words, and the relations between variables can be interpreted in terms of the key concepts and the major knowledge claims. Research reports which rely on tests of statistical significance alone to establish educational significance are justifiably criticized.

Many people feel that the best way to criticize research is to compare the work against a checklist of possible faults. Many such checklists have been produced.[1] Checklists can be valuable, for they serve as a reminder of key features of an investigation which should be considered in any appraisal.

Checklists, however, have at least two major shortcomings. First, they do not provide the criteria to judge the criteria. On what basis is the critic to decide if "the instruments are valid" or "the design appropriate"? Such judgments require knowledge of facts, concepts, and research paradigms. The model appraisals in this book are often very lengthy precisely because we have attempted not only to share our judgments but also to provide the basic information needed to reach such judgments.

A second shortcoming of checklists, in our opinion, is their almost total preoccupation with methods of work--i.e., with questions of research design, measurement and analysis. The methods of work are very important, of course, for they can make the difference between securing valid or invalid knowledge claims. But as researchers become more sophisticated about these things and the number of investigators capable of producing reasonably "tight designs" research grows, it becomes increasingly

[1]For a bibliography of such checklists, see Bruce B. Bartos, "A Review of Instruments Developed to be Used in the Evaluation of the Adequacy of Reported Research." Bloomington, Indiana: Phi Delta Kappa, Research Service Center, Occasional Paper #2, 1969.

important to ask, as well, a different set of questions about the research--questions such as its import for education and its implications for policy or practice. In the appraisals in this book, we have attempted not to slight these other dimensions of the appraisal process.

CHAPTER 2

Nine Research Articles and Critiques: Description and Use

Implicit in the development of this book is our assumption that repeated practice is required to learn to appraise educational research critically. Consequently, we have selected several research articles for appraisal. Before beginning this analysis task, the following comments about the articles and the use of these materials are important.

DESCRIPTION

Characteristics used in selecting the nine articles reproduced in this book were problem area, methods of work, value, and difficulty.

Problem Area. A wide variety of educational topics are represented by the articles. Provided in Table 1 is a brief description of the primary problem area associated with each article. Several of the articles have abstracts which indicate more precisely the content of the research report.

Methods. An attempt was made to select articles utilizing a diversity of approaches. In Table 1, brief labels are given for research types, but these tend to mask the differences in methods employed by the several investigators.

Value. All the articles have redeeming features. It is true that we found much to criticize about all the articles, but any article can be criticized negatively. In our opinion, the articles are of reasonable quality from which there is much to be learned.

Table 1

SUMMARY OF THE RESEARCH ARTICLES

Chapter	Senior Author	Problem Area	Type of Research
3	Josephson	Grading and student attitudes	Status
4	Kaplan	Evaluation of a Head Start program	Status
5	Hackman	Prediction of "long-term" success	Prediction
6	Hunkins	Effect of questioning procedures on student achievement	Experimental
7	Durkin	Development of a concept of justice	Status
8	Harris	Reinforcement and behavioral modification	Case study/ experimental
9	Elkind	Factors affecting the validity of creativity assessments	Experimental
10	Bronars	The case against experimenting with live animals in elementary school	Philosophical analysis
11	Bridges	Small group composition and productivity	Experimental

Difficulty. Results of field testing indicate that all articles are understandable to college students. We avoided articles having sophisticated statistical analyses or dealing with topics requiring prior expertise in a specific content area to be understood. Several of the articles are accompanied by special notes in which an occasional technical term or isolated material is defined or explained. Although there are some minor variations among them, all articles are moderately easy to understand.

The level of sophistication of the critiques, however, are not equal. In this respect, the articles are arranged in a crude ordering from simple to hard.

USE

The articles may be read in any order because each illustrates different concepts of research and these are not arranged sequentially.

Nevertheless, it is probably wise to begin with one of the studies listed toward the top of Table 1 and work toward those having a more intensive and sophisticated appraisal. Regardless of the article being analyzed, keep in mind the following points.

1. Any piece of research can be criticized negatively. The perfect study does not exist. Any investigator is operating within a system of constraints and must make compromises. The fact that weaknesses (as well as strengths) are evident in every study should not be interpreted that they are without value. Quite the contrary. We consider each of the investigations in this book worthy of study.

2. Not all articles that one reads deserve the time needed to perform a thorough analysis as provided with the studies reproduced in this book. The professional must place priorities on how he spends his time. There will be occasions, however, when specific studies have particular importance to a researcher or educator, and for these occasions it is most desirable that he can appraise the work critically. Although the articles in this collection will not be particularly important to many readers, it is well, nevertheless, that they practice critically appraising the articles so that this skill can be learned and then applied to works considered by readers to be more important.

3. The reader must be careful not to infer (improperly) that because the problem area of a particular article is "irrelevant" to his specialty that the task of appraising the article is therefore irrelevant or valueless. The primary purpose of this book is to provide the reader with a set of generalizable skills. The specific articles are merely vehicles through which basic concepts can be taught and habits of workmanship practiced. Much is to be learned about the appraisal process regardless of the particular examples used for illustration.

4. The distinction between the research investigator and his work should be kept in mind. The reader should avoid taking sides for or against the investigator; avoid trying to be easy or hard on him. Rather, the task is to identify the strengths and weaknesses of the work itself and what these assessments mean for the educational value of the study and for the interpretations or knowledge claims resulting from the investigation.

5. Frequently not appreciated by readers participating in field tryouts of the materials is that the learner's expectation should not be to duplicate the model critique. Most readers are simply not able to appraise a study to the extent found in the model critiques. The model critiques are more complete and detailed than can be reasonably expected from even experienced researchers. The purpose of the model critique is not to serve as the standard which students are expected to meet. Rather, they are complete and sometimes overblown statements designed in part to teach concepts and principles.

6. It is our intention that the materials be used either for group or individual instruction. In an effort to make the materials self instructional, we have made heavy use of "student responses." Frequently these are representative replies of student readers participating in the field tryouts of the materials. These student responses are likely to be similar to comments that you, the present reader, may have made. By providing our response to these statements, we hope to increase the interactiveness of the materials and their viability for self-instructional use.

7. The reader is expected to read carefully each article and then to appraise the work by responding to one or more questions. Many students who participated in the field testings performed poorly on the appraisals because they failed either to read the article, the questions, or the appraisals carefully. We've heard much about programs designed to increase reading speed. In our opinion, people need to be instructed how to read more thoughtfully. The first principle in research criticism is actively consider what one reads. The world needs more plodders!

8. One can simply read through these materials like a textbook and passively consider the appraisal tasks and model answers. Alternatively, the learner can write a response to each task, thus helping to insure his active involvement. We much prefer the latter. Appraising the work of others is a "doing" task just as performing research is. Neither performing nor criticizing research is easy; attention to detail is required, the work is demanding, the rewards are high.

Appendix II

Charles H. Josephson

Do Grades Stimulate Students to Failure?

Chicago Schools Journal, Dec. 1961, pp. 122-127

Do Grades Stimulate Students to Failure?

Charles H. Josephson

Chicago Schools Journal, Dec. 1961, pp. 122-127

1. Question:

Before you begin, an important point needs to be made. In study after study that we review, all too often the problem which occasioned the research, and which is used to introduce the research report, turns out not to be the problem actually dealt with by the study as conducted. We are reminded of a Peanuts cartoon in which Linus stands at Violet's front door and asks, "Hi, Violet. Can you come out and play?" Violet responds, "You're younger than I am." A puzzled Linus turns to the reader and queries, "Did that answer my question?"

The mismatch between problem statement and answers collected by the investigator are seldom as gross as that confronted by Linus. A good critic must be alert for such incongruity. He may ask: Do the data provide evidence about the stated problem? Given the data actually collected what question could be composed to which the data would be an answer? Has the phenomena of interest shifted as the study progressed? What conclusions and interpretations are the investigator entitled to draw from the findings?

Five possible problem statements about which data could have been collected are listed in this first question. We are asking you to practice an important skill - namely relating data to the question posed. There is a sense in which one does not really know what the problem is until the solution emerges. Another way of stating this point is to say that any question will remain ambiguous until data which count at the answer to the question are specified.

Consider the following statements:

A. "Grades stimulate students to failure."

B. "Students in slum schools find it more rewarding to be considered academic failures than successes."

C. Students "most likely to succeed" feel the strongest pressure to fail.

D. "...in lower-class schools students of low ability will desire high grades, and students of high ability will desire low grades."

E. "There is a discrepancy between aspiration and achievement."

Which one of the above options most accurately reflects the problem statement that the data of this paper deal with? Why? Give reasons for rejecting each of the other items.
Note: We are NOT asking you which statement is true. We are asking you to indicate which statement represents a hypothesis the investigator attempted to test empirically, i.e. the hypothesis about which the investigator collected data.

1. Answer: C

A. Answer A is quoted from the title. Titles of articles are almost always both illuminating and misleading. Except in the most technical of journals, titles are phrased in ordinary language and it is difficult to achieve precise meaning with the looseness and ambiguity of ordinary language.[1] Note ambiguities in a key word of this title, "failure." "To failure" can have three meanings : (a) to fail out of school (as a drop-out, perhaps); (b) to get a failing grade in a single course (as to fail algebra); (c) to fail to achieve at a level commensurate with ability (underachieving).

For the three reasons which follow, option A was not considered the best statement of the question to which the data reported in the study are relevant. First, the word "stimulate" suggests a causal connection and no such relationship between grades and failure was established. Further, the actual grades students receive are not given, and thus we have no data about failure in the sense of a teacher giving a pupil a failing grade. Finally, the data which are gathered pertain to a slum school and the students in several tracks, and these facts are not mentioned in option A.

B. Option B is Davis' position but this investigator does not actually collect data on what is rewarding to students. Nevertheless some support for this position would be a finding that of 106 students interviewed in the slum school, a large number aspire to (i.e., would "select") grade 5, the failing grade. However, not one gave that response and the author paid no attention to this fact. One might go one step further to ask if the data presented by this investigator

1. Answer:

1. For a further discussion hear Robert M.W. Travers, The Limitations of Variables Derived from Common Language. Washington, D.C. American Educational Research Association, Cassette Tape Series 108, 1971.

1. cont'd.
actually could be interpreted as falsifying Davis' position as stated in option B. The answer is yes, if one can establish that what is rewarding to students and what they would "select" are identical.

C. Although the investigator states option C as a beginning hypothesis (paragraph #4), he gathers no data on peer pressure; he thus cannot compare pressure to fail with a measure of likelihood of success.

D. We think this choice is the most accurate one. See page 2, bottom paragraph where the hypothesis is explicitly stated. Note also that the table giving the data closely follows the hypothesis. Recall that in question 1 we asked if the hypothesis was tested in this study and not whether it could be considered true on other grounds.

E. Although we have data on an expectation of achievement, we have no data on achievement itself and therefore cannot compare achievement to aspiration. The three tracks are said to represent ability levels. If they are also viewed as defining an achievement variable, then some gross data on the discrepancy between aspiration and achievement are provided and option E could be considered an acceptable (but probably not the best) answer.

2. Question:
Having thought about the real purpose of this study, cite one very important reason why research on the broad question addressed in this paper is of value.

1. cont'd.

2. Question:
Having thought about the real purpose of this study, cite one very important reason why research on the broad question addressed in this paper is of value.

2. Answer :

The reward system of a slum school is being studied. There are several acceptable reasons you might have given to explain why research on this topic is of value. One that appeals to us is that IF the research should point up the fact that the grading system isn't working as intended, that "the teacher's reward has become the student's punishment", or that the extrinsic rewards (for example, the grades) of the system wield such a powerful influence that the intrinsic rewards of learning are diminished or by-passed, THEN such distortion would provide support for changing present educational policies and practices. The primary aim of an educational system should be its true educational goals and not the external trappings attached to these goals. Florence Nightingale once said of hospitals that at least they should not spread disease; school systems should not discourage true learning.

3. Question :

Refer to the data presented on the bottom of page 2 and to the investigator's descriptive labels for the grade categories on the bottom of page 3. Which one(s) of the following statements is (are) factually correct interpretations of the findings for students in the accelerated class?

A. Only 1/3 prefer superior grades; nearly the same number prefer average or below average grades.

B. About 2/3 prefer grades above average; only 2 students preferred below average grades.

How do statements A and B differ in the impression they give?

2. Answer:

3. Question:

Refer to the data presented on the bottom of page 2 and to the investigator's descriptive labels for the grade categories on the bottom of page 3. Which one(s) of the following statements is (are) factually correct interpretations of the findings for students in the accelerated class?

A. Only 1/3 prefer superior grades; nearly the same number prefer average or below average grades.

B. About 2/3 prefer grades above average; only 2 students preferred below average grades.

How do statements A and B differ in the impression they give?

3. Answer:

Both statements are technically correct given the investigator's interpretation that 1 means superior and 3 means average. They differ in the impression they give the reader. The A statement suggests a failure of the school to keep high the aspirations of good students. The B statement suggests most students in accelerated classes want good grades. The A statement is the way this investigator interprets the findings (last sentence, p.3). We think it acceptable for a researcher to try to find what his reasoning leads him to expect. He should not, however, stop at this point but should examine alternative explanations. We must remember that one can say a cup is half full or half empty and be correct in both instances. A researcher should be able to, and further has an obligation to, say both, realizing the different possible impressions he may give his readers from these different viewpoints.

4. Question:

Note on the top of page 3 that from each of the 3 programs (remedial, regular, accelerated) one class was selected in some unspecified fashion. Alternatively, the investigator could have selected the required number of students randomly from all the students enrolled in each of the programs. We believe this latter selection plan to be far superior? Why?

4. Answer:

The investigator wishes to compare the grade desires of students of different ability levels. Because he selected only one class from each program, he cannot distinguish differences due to program/ ability level from those due to classroom

3. Answer :

4. Question:

Note on the top of page 3, that from each of the 3 programs (remedial, regular, accelerated) one class was selected in some unspecified fashion. Alternatively, the investigator could have selected the required number of students randomly from all the students enrolled in each of the programs. We believe this latter selection plan to be far superior. Why?

4. Answer:

4. Answer cont'd.

influences. We know from other research that on many variables classrooms differ markedly from one another even when the classrooms are composed of students of the same general ability. The particular teacher, classroom peer relations, and other factors can lead to a distinctive kind of response from students in a particular classroom. The responses of pupils from one of the classes might not be typical of those from other classes in the same track. Thus the differences the investigator notes in the data shown on the bottom of page 2, may not be due to program/ability level group differences at all but to other attributes of the three particular classrooms he selected for the study. Had a random sampling procedure been used, students from several classrooms within each program would have been selected and this source of confusion in data interpretation would have been avoided.

5. Question:

Recall that when the students were divided into the three programs (accelerated, regular, remedial) and their desired grades noted (see data on the bottom of page 2), the investigator concludes that the expected, "...inverse relationship between ability and grades desired does not obtain." (p.3) However, when the investigator reclassifies the regular and accelerated students into a single category, "...a significantly different picture emerges." (p.3) Is it wrong for an investigator to manipulate his data in this way in search of confirming evidence? Why?

4. Answer cont'd.

5. Question :

Recall that when the students were divided into the three programs (accelerated, regular, remedial) and their desired grades noted (see data on the bottom of page 2), the investigator concludes that the expected, "...inverse relationship between ability and grades does not obtain." (p.3) However, when the investigator reclassifies the regular and accelerated students into a single category, "...a significantly different picture emerges." (p.3) Is it wrong for an investigator to manipulate his data in this way in search of confirming evidence? Why?

5. Answer:

We don't think so, provided the cautions mentioned in the next paragraph are noted. Such "teasing" of the data in which after-the-fact- hypotheses are tested can provide insights into the subject of the research. Such unplanned analyses, however, are generally more valuable as possible leads for future research than as firm conclusions.

We suggest these cautions. First, the data should be presented in the manner the investigator had expected to present it before the data were collected, (the present investigator does this) or else the departure explained. Second, the investigator should state or imply (as the present investigator does) that the particular analysis presented was suggested to him only after the data were observed. Third, the investigator should also report plausible after-the-fact analyses which do support his expected conclusions. In this regard, it is of interest to note that in this study the largest group differences occur when the extreme groups, the remedial and accelerated classes, are compared to the regular classes. This finding, if replicated by others, would suggest a much different interpretation from that provided by the investigator. Finally, relationships found as a result of such after-the-fact manipulating must not be taken too seriously, especially those: (a) not predicted ahead of time; (b) not amenable to a reasonable interpretation; and (c) emerging from a large number of comparisons. When enough things are examined, some comparisons will seem "significant" by chance alone.

5. Answer:

6. Question:

Both in the case when the data for the three programs (ability level groupings) are kept separate, and in the case when the data for the regular and accelerated classes are combined, the differences among programs in the per cent of students desiring the various grades are not statistically significant according to our calculations. What is the importance of this statement?

6. Answer:

Lack of statistical significance means that the differences among the percentages in the three columns in the table on page 2 could be due, not to differences between program/ability level groups in the grades desired, but simply to errors in sampling. Failure to get statistical significance can be interpreted as a vote of no confidence that the differences which were found will be observed with another sample of students. The investigator should have realized that the program/ability level group differences should not have been taken seriously and refrained from such strong definitive language as, "It seems uncontestable that a distinctively low-aspiration group (i.e. the middle group) emerges from these findings." (p.5).

7. Question:

Although there are serious flaws in this study, there are also some commendable aspects. List four such positive features (not conclusions) of this paper.

6. Question:

Both in the case when the data for the three programs (ability level groupongs) are kept separate, and in the case when the data for the regular and accelerated classes are combined, the differences among programs in the per cent of students desiring the various grades are not statistically significant. What is the importance of this statement?

6. Answer:

7. Question:

Although there are serious flaws in this study, there are also some commendable aspects. List four such positive features (not conclusions) of this paper.

7. Answer

The following list is meant to be suggestive and not necessarily complete.

a) The investigator sees research as having a clear bearing on educational policy and practice, and suggests changes in these practices based on such relevant research.

b) He uses his reasoning powers in the search for an explanation (but not a generalization) of phenomena he thought he observed in the schools.

c) Even though a teacher in the schools at the time of the study, he does a study - collects the data in situ - which makes good use of an educationally relevant context. We think more studies should be done by people who make decisions about practice as a consequence of the studies undertaken.

d) The investigator cites a puzzling observation in the literature (Alison Davis's position) which is an impetus to research.

e) The investigator realizes some of the inadequacies of his study and that more complete and better planned ones need to be made.

f) He publishes locally where the impact of such a controversial study will most likely have an effect.

g) The investigator attempted to obtain valid measures of aspirations. He thought of devices (e.g. anonymous responses and additional questions) as an "honesty check" to the first question. We do not claim he was successful but do commend the attempt.

h) He manipulated his data in more than one way. (See question and answer #5).

7. Answer:

7. Answer cont'd.

7. Answer cont'd.

i) The research was open-ended in the sense that it suggested further investigation.

j) The paper was highly readable and written in an interesting fashion.

Concluding remark: In their classical paper, Campbell and Stanley wrote:[1]

> At present, there seem to be two main types of "experimentation" going on within schools: 1) research "imposed" upon the school by an outsider, who has his own ax to grind and whose goal is not immediate action (change) by the school; and 2) the so-called "action" researcher, who tries to get teachers themselves to be "experimenters", using that word quite loosely. The first researcher gets results that may be rigorous but not applicable. The latter gets results that may be highly applicable but probably not "true" because of extreme lack of rigor in the research. (p. 21)

The present paper clearly falls into the second category.

1. Campbell, Donald T. and Julian C. Stanley, Experimental and Quasi-experimental Designs for Research, Rand McNally, Chicago, 1966.

Appendix III

Eleanor Kaplan

"Head Start" Experience and the Development of Skills and Abilities in Kindergarten Children

Graduate Research in Education and Related Disciplines
Vol. II, No. 1, April 1966

"Head Start" Experience and the Development of Skills and Abilities in Kindergarten Children

Eleanor Kaplan

Graduate Research in Education and Related Disciplines

Vol. II, No. 1, April 1966

SPECIAL NOTES

The present article would be classified as an example of educational evaluation. There is disagreement among experts regarding the distinction between evaluation and research. Some say that the purpose of evaluation is to derive assessments of the worth of particular instances of educational undertakings such as individual textbooks and specific programs; the purpose of research is to produce generalizable conclusions. We see the distinction to be one of degree rather than kind. In both studies we ask whether the activities followed permitted the investigator to accomplish the objectives of the study.

For each set of data the investigator conducted a chi square test of the statistical significance of the difference between the score distribution found for the two groups of children. The investigator is seeking to determine whether the difference in the proportion of students in the two groups who are above a particular score could happen by chance alone. Specifically, the statistical test indicates the probability of getting such a large difference in proportions if only chance (i.e. sampling variability) were operating. When this probability is small (defined in this paper as less than 5%) and thus the chance-alone hypotheses is not very likely, the investigator indicates that the difference was statistically significant and, presumably, the Head Start program had an effect.

On page 17, just before the last paragraph, the parenthetical expression should have been written: $(.05 < p < .10)$. The symbol, $<$, means "less than." Thus, the probability of differences between two groups on enunciation scores as large as those actually found could be expected to occur 5 to 10% of the time, even if chance alone were operating (i.e., program had no effect). This probability wasn't small enough for the investigator to reject with confidence the hypothesis that for a population of children similar to these 70, no differences on this variable would be found.

"Head Start" Experience and the Development of Skills and Abilities in Kindergarten Children

Eleanor Kaplan

Graduate Research in Education and Related Disciplines

Vol. II, No. 1, April 1966

QUESTIONS:

1. "The purpose of this study was to evaluate whether the children who participated in Project Head Start were better prepared for kindergarten than those who did not participate..." To accomplish this purpose, the investigator: 1) reviewed the literature, 2) stated hypotheses, 3) selected subjects, 4) selected and constructed measuring instruments, 5) administered and scored tests, 6) performed analyses, and 7) drew conclusions.

   A. What two importantly different kinds of information are contained in this review of the literature? What, in general, are the main purposes of any review of the literature and how well did the investigator succeed in achieving these purposes?

   B. Write a critical appraisal of each of the other six aspects of the study identified above, being sure to cite strengths as well as weaknesses.

2. The investigator evidently feels that the Head Start programs involved in her study were very effective and worthwhile. Yet there is information needed in addition to that given in the report if one is to reproduce such an effective program elsewhere. What information is lacking in the report which prevents it from serving as a guide to one who must develop and operate a Head Start program? (Assume that the leader has much freedom in how he plans and runs a Head Start program.)

"Head Start" Experience and the Development of Skills and Abilities in Kindergarten Children.

Eleanor Kaplan

Graduate Research in Education and Related Disciplines

Vol. II, No. 1, April 1966

ANSWERS:

1.A.

One kind of information in the literature is the description of the social and political forces which in 1965 were changing drastically the prekindergarten public education of economically and socially disadvantaged children. The Kaplan report indicates by 1965 Project Head Start "benefitted" 560,000 youngsters in 2,5000 communities at an estimated cost of $112,000,000. A second kind of information in the literature review is more commonly found. The investigator cites empirical studies (e.g., Bernstein, 1960, 1962; Deutsch, 1956b) and studies of new educational practices (Graham and Hess, 1965; Hess and Rosen, 1965).

One main purpose of a review of literature section in empirical studies is to describe the educational context in sufficient detail such that the justification of the study is clear. The literature review succeeds fairly well to give us the political, historical and empirical context of the study. These political and social changes to educational practice which the investigator documents serve as an excellent stimulus and justification for educational research.*

A second main purpose of a review is to indicate the sources of concepts and principles used to guide the inquiry. One can find instances in which the evaluation was influenced by the empirical studies and writing quoted in the review. One example of the influence of these sources on the conduct of the inquiry is the

* Among social scientists and educational researchers there often exists a tension between being socially relevant ("on a white horse") and scientifically rigorous ("wearing a white coat"). Whenever social changes take place rapidly and pervasively, the tension can develop into a rift. In our opinion this division is unnecessary and counterproductive. Social changes can be thought of as an excellent stimulus to empirical inquiry, as we indicate about the Kaplan study. More than that, the empirical researcher who can say as a result of inquiry that he knows both the facts and the educational consequences of political and policy decisions can become a valuable influence upon the shaping of future educational policies. Many researchers would prefer to spend money on research before changes are made so that they might be made intelligently in the light of new knowledge. Social urgencies dictate otherwise sometimes. Perhaps the best course is to combine the two: research can change policy and practice, and changes in policy and practice can be a valuable stimulus to further research. For a discussion of some of these issues, see Nevitt Sanford, The American College, John Wiley & Sons, New York, 1962, pp. 1-30.

literature which points to the need for emphasis on language teaching for the disadvantaged. This information justifies the inclusion of language development measures in the study.

A third main purpose is to provide a theoretical context from which the knowledge claims of the inquiry can receive intelligible interpretation.. There is none of this material in the review. Some readers would say that the large differences found after a short summer program are rather remarkable, yet there is no theoretical context, nor even an educational rationale, provided which can help us to account for or make sense out of these findings.

1.B.

1. Hypotheses. The hypotheses on page 10 are a clear statement of the questions to which the investigator is seeking answers. Although it is not always necessary for questions to be in the form of hypotheses in which predicted results are stated, we approve of the investigator's indication in this section of the direction in which she predicts the results will appear. Most experts favor directionally stated scientific hypotheses to those expressed in the less communicative null form.

In assessing the hypotheses, several student readers questioned the investigator's methods of measurement, the failure to consider other variables in the study, and the feasibility of matching students. Valid as these concerns may be, for convenience they will not be considered at this point in our assessment of the study.

2. Subjects. The principal technical flaw in the evaluation is that no control had been exercised over the assignment of children to Head Start or control programs. Further, because such variables as sex, ethnic background, age (only a 10 month range), language spoken in the home, and age of siblings would not be expected to be highly correlated with the measures used in the study, the reader has little assurance that the two groups being compared were initially equal in those skills and abilities the Head Start program most wanted to affect. The investigator also mentioned this problem. (p. 14).

We could assess more accurately the likelihood of this initial equality if we were told in the report the reasons why the control children did not attend Head Start classes. Did they live too far away from the Head Start center, come from more stable homes, or live in better neighborhoods? Did the control children not attend Head Start programs because their parents chose not to send them? If so, then differences in attitudes toward education (as seen by differences in the learning experiences provided in the home - learning experiences such as talking, reading, color identification, etc.) could mean that the Head Start children would have scored higher than the control children even before the Head Start experience was begun, and certainly after an additional year of a better learning situation in the home.

The investigator was wise not to match students on intelligence or other cognitive or attitude variables measured after the Head Start experience. If the Head Start program improved the children's scores on such variables, then matching children on their scores would cancel the very effects to be demonstrated.

Suppose the investigator had been able to administer identical criterion measures (verbal fluency, enunciation, etc.) before the Head Start experience and to match children on the basis of their scores on such measures. Differences between the two groups would still be expected on these measures when the children were tested in kindergarten, even if the Head Start program had no effect in developing the skills and abilities measured by the criterion tests. Such bogus or false differences can be explained by the regression phenomenon. (For an elementary discussion of the regression phenomenon, read: Kenneth D. Hopkins, "Regression and the Matching Fallacy in Quasi-Experimental Research", Journal of Special Education, 3, 1969, 329-336.)

We do not fault the investigator for matching students. We merely wish to point out that such matching was probably largely ineffective in assuring the equality of the two groups prior to training. Matching on variables measured before the Head Start programs were begun and which were more highly related to the criterion variables would have been far more preferable. But even if this were done, the lack of random assignment of children to the Head Start and control conditions still prevents the ruling out of selection bias and regression artifacts.

Frequently expressed reactions of student readers are that 35 children per group is too small a number and the number of Head Start programs being evaluated is not mentioned in the article. More data are always desirable, but an investigator must weigh the increased scope against the increased "costs" associated with having a larger sample size. The differences between the Head Start and control groups were sufficiently great that 35 cases per group were adequate to reject for most of the variables the chance alone null hypothesis. Perhaps more useful than a larger sample size per se would be having as a sample children taken from several Head Start programs. We suspect, but are not certain, that all the children were exposed to the same program and, if this was the case, the generalizability of the results is very uncertain.

3. Measuring Instruments (selection and construction). Given the rather limited goal of assessing the comparative performances of the two groups of children, then ideally the measuring instruments used in the study should represent a diverse collection of reliable and valid devices of measuring the degree to which the intended skills and abilities have been developed and unintended ones are absent.

Many student readers objected to the absence of test reliability and validity data in the report. If a test is unreliable, then it is not measuring any trait or skill consistently; the test score then has a large component of random error. Such inconsistency of measurement and random error are to be avoided since real treatment effects will not be revealed by such unreliable instruments. In the context of this study, Head Start programs can not be judged effective if the measures of effectiveness are largely unreliable. Since the investigator did find group differences, we can assume that the instruments employed had acceptable levels of reliability.

"Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score...One validates, not a test, but an interpretation of data arising from a specific procedure."* The investigator would probebly claim that the test items are representative instances of the skills being described and, thus, her inferences about children's capabilities based on their test performance are valid. Such a claim seems reasonable to us with possibly two exceptions. First, we question whether the Goodenough - Draw a Man Test is as much a measure of motor coordination as it is an indicator of other skills. (Note, the investigator probably meant to say on page 12 that the test's scales rather than norms were used.) Second, we have some qualms about the buttoning-own-clothes measure since the task is not the same for all children. (Some children had harder clothes to button than others.)

Because the specific Head Start programs being evaluated were not described, we do not know for sure the extent to which the abilities and skills measured by the tests used in this study do represent the primary objectives of these programs. Further, we do not know the extent to which the very tasks used in the tests were used in the training programs themselves. This is not to say that it would be wrong to use identical tasks in both teaching and testing. It is just that interpretation of group differences and the value of a program depend upon knowing the relation of tasks tested to the tasks used in training.

We suggest that in an evaluation study of this type three categories of tasks be used in the testing: 1) those tasks directly involved in the training (on which large group differences would be expected); 2) tasks not used in the training but on which it is hoped there will be group differences; and 3) tasks representing unintended outcomes (on which there is expected no group differences).

* Cronbach, Lee J., Test Validation, Chapter 14 in R.L. Thorndike (Ed.), Educational Measurement, American Council on Education, Washington, 1971.

We would like to have seen more of the category two and category three tasks used in this evaluation. As examples of category two tasks, we would like to have seen the differences in performance of the two groups on tasks requiring left-right visual search and production of graphic symbols (e.g., letters). In addition, as a category two or three task, measures of personal-social adjustment to school would have also been of interest.

The investigator engaged in good practice, however, in including several measures of performance rather than relying on just one or two. Where there were no standardized tests to measure the type of performance on which the investigator wished to compare the groups, she devised her own tests for these skills and abilities. This research practice is commendable.

4. Test Administration and Scoring. The importance of administering tests prior to the start of the Head Start program was mentioned earlier.

The investigator indicates that the instruments were administered, "... at the beginning of kindergarten in order to insure that these skills and abilities to be tested were not learned during the kindergarten experience." (p.13) Although there is some merit to this procedure, we feel it would have been desirable if some of the tests had also been administered at the end of kindergarten, or even later. The critical importance of ascertaining the long-term benefits of Head Start programs has been well documented by the investigator herself. The advantage of the Head Start group during the first weeks of the school year may be due primarily to preschool environment and materials which have no carry-over effect on later learning. Although determining if there is an immediate effect is useful, it would be of great value to document that a primary goal of Head Start programs, increased performance in school, was met.

Recall that the measuring was not blinded from the standpoint of the observer, although the investigator claims on page 14 to have made no effort to remember which children were in the Head Start group. This is small comfort to the reader who suspects that the children's membership in either group could have been independently identified and thus could have biased the judgment of the investigator as she administered and scored the tests.

The testing was somewhat subjective, both in administration (e.g., frequency of directions to be given, probing for termination of responses) and scoring.(See especially the cutting, coloring and enunciation tests.) Thus, the results were open to the influence of the evaluator herself. The investigator is not to be faulted for using instruments which were subjective in nature. However, using these instruments in such a manner that the subjective element invalidates the comparison between the two groups is a procedure open to censure.

5. Analysis. The analysis of the data was adequate and not misleading even though more precise statistical techniques could have been employed. The investigator could have utilized the exact scores and not have forced them into two categories (above and below the combined median). Further, the investigator could have made use of the fact that she had matched pairs of children. However, these objections carry little weight since the result of substituting these more refined measures would have been more power (i.e., likelihood of rejecting false "no difference" hypotheses) and almost all of the chance-alone or no difference hypotheses were rejected even without their use.

The investigator is to be commended for not evidencing an unthinking attachment to a particular criterion of statistical significance. (See Special Notes on page for an explanation of the 5% criterion used by the investigator.) Particularly in the case of the cutting-skill variable, the evaluator showed her willingness to accept evidence of a difference even though the obtained test statistic fell somewhat short of the critical value needed to claim statistical significance at the 5% level.

6. Investigator's Conclusions. The investigator is quite correct in stating that,"...kindergarten children who had attended the Head Start program were superior to those who had not in each of the skills and abilities tested." (p.22) This conclusion is merely a factual statement of the results found. Even though a few differences did not reach statistical significance, it is a fact that the Head Start group had superior scores on all the measures.

The investigator is also permitted to say, "The findings support the current view that culturally deprived children benefit from preschool programs." (p.25) "Findings support the current view", is interpreted to mean, findings are consistent with the current view, and does not imply that the results prove that the children benefitted from the programs.

Because of the lack of fundamental controls as specified earlier in our appraisal, we have no assurance that the differences were due to the Head Start programs. Thus, we feel the investigator is not justified in making conclusions that imply the Head Start programs caused the superior performance. We question the validity of such a conclusion as: "The experiences provided in the instructional program made it possible for children in the preschool Head Start project to become more adept..." (p.24)

Finally, before claiming that results will generalize to other Head Start projects, we would want to see such positive results from a larger sample of students and programs.

2. To develop and operate a Head Start program effectively, one would need to have much financial, legal and political information not touched upon in the report. To plan the instructional aspects of the program, that is to decide what to teach and how and when to teach it, a detailed specification of the Head Start programs being evaluated in the present article is needed if the experience reported in the article is to have benefit. Lack of this specification is a major deficiency of this report.

The reader is left completely in the dark as to the components of the programs, their duration, the training and number of staff, the objectives of the programs, the procedures used to achieve these objectives, etc. Without even the most rudimentary description of the programs, the investigator has produced an evaluation report not unlike a research report in which the independent variable was unspecified. As the report now stands, its nearly total neglect of description of the programs makes it of use only to a small number of persons who are intimately connected with the programs being evaluated. No two Head Start programs are alike. Without a description of the programs herein evaluated, we do not know what programs to perpetuate or how the programs should be conducted differently. What good is an evaluation that something works when that "something" is not defined?

Appendix IV

J. Richard Hackman, Nancy Wiggins, Alan R. Bass

Prediction of Long-Term Success in Doctoral Work in Psychology

Educational and Psychological Measurement
1970, 30, 365-374

Prediction of Long-Term Success in Doctoral Work in Psychology

J. Richard Hackman, Nancy Wiggins, Alan R. Bass

Educational and Psychological Measurement, 1970, 30, 365-374

1. Question:

What were the investigators hoping to achieve? That is, what was the purpose(s) of the study?

Answer 1:

We think the investigators had two primary purposes which are well stated in the opening and closing sentences of the initial paragraph of the article: a) to examine, "...the degree to which measures of aptitude and undergraduate preparation obtained before the beginning of doctoral study are predictive of the (short and long-term) 'success' of psychology graduate students."; b) "...to determine the degree to which evaluations made at the end of the first year of doctoral work are congruent with the long-term assessments of success in the program." The relationships mentioned in purposes a) and b) above are shown in Tables 1 and 2 respectively.

Student Responses. Several students inferred that the investigators were trying to make predictions rather than "just to gather information" about relationships between predictors and criteria. They claim that, "...the purpose of the study was to find a kind of cause/effect relationship, so that the Graduate School at the University of Illinois or other graduate schools can make specific recommendations to undergraduate institutions, to future students, and to faculty members about changing or maintaining certain practices."

Our Reply. Worthwhile as such a purpose might be, the investigators did not state it as their aim. If their purpose were to devise a prediction system which could be used by educators, they no doubt would have then followed the recommended practice of cross-validating their results; that is,

1. Question:

What were the investigators hoping to achieve? That is, what was the purpose(s) of the study?

Answer:

trying out the system on a student group different from that used to develop the prediction formula.*

2. Question:

(a) How many specific pre-enrollment predictors (not groups or categories) were used? Your answer should be a specific numerical value.

(b) How many specific criteria were used?

(c) Was it a good idea to employ so many variables in a single study? Why or why not?

Answer 2:

(a) Thirteen predictors were used. These predictors are listed in the left-hand column of Tables 1 and 3 as well as in the body of the article.

(b) Ten criteria were employed; all but one of these are considered short-term criteria. These criteria are listed at the tops of the columns in Table 1, in Table 2, and in the body of the article.

(c) We approve of using multiple predictors and criteria in any study for two reasons. First, we are rarely interested in a dependent variable which can be perfectly measured by a single variable. A good case in point is the present study in which "success" is clearly a complex concept - the more aspects of success we study the better. Second, the more independent, or predictor, variables included in a study, the more information

* For an entertaining account of how failure to cross-validate a prediction system can lead to astounding and unfounded claims, read: E Cureton, "Reliability, Validity and Baloney.", Educational and Psychological Measurement, 1950, 10, 94-96.

2. Question:

(a) How many specific pre-enrollment predictors (not groups or categories) were used? Your answer should be a specific numerical value.

(b) How many specific criteria were used?

(c) Was it a good idea to employ so many variables in a single study? Why or why not?

2. Answer:

Answer 2 cont'd.

we obtain about the relationships we are interested in. Study of a greater network of interrelationships aids in comprehending and explaining the reasons for the relationships.

On the other hand, use of variables poorly measured or lacking rationale for their inclusion should not be encouraged. It should be kept in mind that when a great many relationships are studied, it is probable that some bogus, "significant" ones will appear. Thus, caution is required in interpreting isolated findings. Further, for statistical reasons involving the stability of the prediction equation coefficients, there are too many predictors (for so few students) to construct a prediction system that would be expected to work well (cross-validate) on a different sample. The investigators were wise to focus their analyses on simple, two-variable relationships.

3. Question:

The predictors used are categorized into four groups:

(a) Aptitude and ability
(b) Foreign language facility
(c) Undergraduate grades
(d) Rated quality of undergraduate school.

Evaluate the appropriateness of the specific measures employed in each group of predictors. (In your answer, focus upon whether these measures were reasonable choices and not upon whether, in fact, they seemed to work in this particular study.)

Answer 2 cont'd.

3. Question:

The predictors used are categorized into four groups:

(a) Aptitude and ability
(b) Foreign language facility
(c) Undergraduate grades
(d) Rated quality of undergraduate school

Evaluate the appropriateness of the specific measures employed in each group of predictors. (In your answer, foc upon whether these measures were reasonable choices and not upon whether, in fact, they seemed to work in this particular study.)

3. Answer:

3. Answer:

a) Aptitude and Ability Predictors. At least for short-term success, aptitude measures have been shown to be good predictors. The Graduate Record Examination (GRE) tests are widely employed and have proved useful in the past. The GRE correlations serve as a useful benchmark against which to judge the magnitude of relationships found with other predictors. Both past performance and current practice argue for inclusion of these test scores.

b) Foreign Language Facility. Unfortunately the reader has to wait until the very end of the paper before he is given the rationale for including these predictors. The argument is not terribly convincing. We have no objection to the inclusion of foreign language facility but suspect more interesting and meaningful predictors could have been found. The three specific measures employed in this category leave much to be desired as true indicators of foreign language facility. They may have been used because they were handy. Their inclusion is no crime; it is just that they are not apt to be very enlightening.

Student Responses. In the evaluation of several of these predictors, as well as in the evaluation of some of the criteria (see Question 4), a large number of students were critical of the subjectivity involved in the measures. Many students went so far as to say that some measures were "worthless" or "should not be used" because they were subjective.

Our Reply. "Subjectivity" can have two meanings. In one sense, subjectivity means based on personal experience or a matter of opinion. In another sense, it means unreliable and that judges do not agree. A doctor should not dismiss a patient's complaint of pain because it is based on personal experience or because other judges cannot agree on the amount of pain involved. Likewise, we would caution researchers against an off-hand dismissal of all subjective measurements. The phenomena we may have the greatest difficulty measuring will sometimes be those most worth measuring. One must often ask whether it is better to measure something trivial well or to measure something important poorly.

3. Answer cont'd.

3. Answer cont'd.

c) Undergraduate Academic Performance. It is wise to include these predictors for the same reasons that the aptitude measures should be included. Of course, grades from different institutions are not completely comparable since a C at one institution may show greater achievement than a B at another institution. Nevertheless, even with such a deficiency, grades have been found to be useful predictors in the past and should be included.

Analyzing the grade record by specific course has the advantage of making the grades somewhat comparable, although this comparability is achieved at the loss of reliability. Grade averages based on one or several courses simply are not as reliable as more composite measures for the same reason that tests with one or only a few items are not as reliable as total scores computed on many-item tests. We further wonder why (but are not critical that) grades earned during the first two years of undergraduate study and the number of semester hours of psychology were not included as predictors.

Student Response. "Doesn't the regression phenomena enter in here? One presumes a 2.75 or 2.8 cut-off so you'd be looking mainly at very high grades to start with."

Our Reply. The subjects are a select, extreme group whose performance on other measures is expected to regress toward more average levels. Because this group is not being compared with other groups, however, this regression effect is not a source of bias. The student does suggest a reason why undergraduate grades (and other measures used in student selection) might not be as highly related to the criteria as one would hope. Presumably the 42 subjects in the study all had quite good undergraduate grades (or else they would likely not have been admitted to graduate school). A predictor which does not discriminate among the students (that is, the students' performances are relatively homogeneous) is not likely to correlate highly with a criterion. Had all the students

3. Answer cont'd.

3. Answer cont'd.

who applied to the doctorate program been admitted, regardless of their undergraduate grades or aptitude test scores, the correlations involving these predictors would undoubtedly have been greater.*

d) Quality of Undergraduate Institution. Because grades at different institutions are not comparable, we see the inclusion of this variable as a wise decision both in studying its relationship with the criteria directly, and as a variable to use in adjusting undergraduate grade averages. More information about the number, nature, procedures and criteria used by the committee to arrive at the quality ratings would be helpful to the reader who might wish to use the same variable in a local prediction study. Failure to specify fully how this variable was measured makes it of limited use to others.

Further, we wonder if the judges' ratings of particular institutions could have been biased by knowledge of which students came from which institutions. The judge, for example, might think more highly of institution X because student A, who is given high ratings on the "success" criteria, came from that institution. Conversely, perhaps students coming from institutions thought highly of were expected to do well and a self-fulfilling prophecy was in operation. This latter possible explanation for the positive correlation between the quality of undergraduate institution and most measures of "success" was noted by the investigator. As one student put it: "It looks like a case of circular reasoning. What schools are rated excellent? Those whose students do well at this university. And what students are successful at this university? Those who come from schools that are rated excellent! Shall we go another round?"

* To see why this is the case, consider three persons whose IQs are 110, 111 and 112 (very homogeneous scores). There is no predicting who would do best in college. If their IQs were 50, 100 and 150 (mentally retarded, average and gifted), making correct predictions would be easy.

3. Answer cont'd.

Student Responses. "There should be categories between 1 and 2 and between 2 and 3 which many schools would fit into more fairly and accurately." "The 3-point scale range is too small as to make the differences practically useless."

Our Reply. We disagree. Even 2-point scales (such as above average, below average) have been found to be very useful in predicting criteria. Although we certainly have no objection to a finer scale, the experience has been that added scale values result in rather meager gains in predictability.

4. Question:

The criteria of short term "success" are divided into three groups:

a) Grades earned in first year of graduate school

b) Self-report measures

c) Faculty ratings

Evaluate the appropriateness of the specific measures used in each of these groups of criteria. (In your answer, focus upon whether these measures were reasonable choices and not whether, in fact, they related to the long term "success" criterion.)

4. Answer:

a) Grades earned in first year graduate school. Grades have typically been used as measures of success. It is a good idea to include them both because they are considered important and because they can be used to study their relationship to long-term success. We believe the investigators were wise to separate out the grades earned in core courses, for at least these grades would be comparable among the students. One student's grade in physiological psychology and another's grade in abnormal

4. Question:

The criteria of short term "success" are divided into three groups:

a) Grades earned in first year of graduate school.

b) Self-report measures

c) Faculty ratings

Evaluate the appropriateness of the specific measures used in each of these groups of criteria. (In your answer, focus upon whether these measures were reasonable choices and not whether, in fact, they related to the long term "success" criterion.)

4. Answer:

4. Answer cont'd.

psychology, for example, might not be comparable. The fact that these core courses were included in the end-of-year average means that this composite measure will have a built-in dependency (correlation) with the other criteria in this category.

b) Self-report measures. These two measures seem to be reasonable indicators of speed and persistence toward achieving the Ph.D.

Student Responses. "Why was the student rating expressed as slow or fast progress to a Ph.D. rather than feeling with satisfaction with progress whatever the speed?" "A student can be deeply engrossed in his study for learning's sake and be highly successful and motivated and yet be totally unconcerned with his speed toward his Ph.D. It is unfortunate that large universities often place the degree above the actual learning taking place."

Our Reply. Of course, other student self-report measures could have been used. We suspect that administrators and professors associated with the degree program were more concerned about the students' perceptions of their actual progress than these students' feelings of satisfaction about their progress. Since the investigators did include grades earned in first year graduate school among the criteria, "actual learning" was not ignored in this study.

Student Response. "Self-report measures should not be obtained after the grades were issued, but before."

Our Reply. We found this to be an interesting reaction to which we could both agree and disagree. By requiring the student to report his progress before he receives the formal grades, we can obtain a measure of how he truly thought he was progressing and such an evaluation might be less contaminated by faculty opinion. On the other hand, by permitting the student access to the formal grades as information to use in making a considered judgment of his progress, a more realistic estimate of his true progress might result.

4. Answer cont;d.

4. Answer cont'd.

c) Faculty ratings. Such categories as, "excellent progress, assured of financial aid", and, "dropped from graduate program", simply do not seem to be points on a common scale. These ratings appear to encompass a whole gamut of possibilities, including good performance, persistence and voluntary withdrawl, and we would like to have seen all of the scale values for these ratings. Further, the shorthand labels of these variables in Tables 1 and 2 do not seem especially appropriate.

Student Response. Many students felt that, "there could be bias in a faculty member's opinion", and that faculty ratings, "...are an unfair criterion." Further, many students wanted, "...to know how divergent the various faculty members were in the rating of the same student."

Our Reply. We agree that faculty ratings might have bias and not be a fair rating of a student's real progress. In addition, as one student pointed out, "...faculty ratings can be influenced by predictor ratings." (We discussed this a bit toward the end of our answer to Question 3d.) Since each student's faculty ratings were averages of several ratings, the influence of a single professor's bias or susceptability to contamination by a predictor was lessened. Although we recognize faculty ratings will have at least some shortcomings, we nevertheless support the use of faculty ratings as a criterion of short-term success. The fact that the hiring of recent Ph.Ds depends heavily on the recommendations of the students' professors serves to remind us that colleges and universities consider such faculty ratings to be a suitable criterion.

5. Question:

A key concept in this study is "long-term success."

a) Give two or more reasons for rejecting the definition given this term in the paragraph starting at the bottom of page 367.

b) How might this definition be defended?

4. Answer cont'd.

5. Question:

A key concept in this study is "long-term success."

a) Give two or more reasons for rejecting the definition given this term in the paragraph starting at the bottom of page 367.

b) How might this definition be defended?

5. Answer:

a) Before listing many objections to the definition of long-term success, it is necessary to point out a confusion on the part of several students. The professors of the students did NOT make the long-term success ratings. Rather, these professors were asked only to define what they thought "constitutes 'success' for a psychology doctoral student." Based on these answers, the investigators, in some unexplained way, constructed the 9-point long-term success scale. Two judges (probably two of the investigators or their assistants) then made the rating using the information (available at the time the student left the university) of where the student was going and the "circumstances " of his leaving. Thus, for practically all students, the long-term success index was determined from data available well before six years had elapsed.

Probably the most serious criticism of the long-term success measure is that it fails to consider many factors commonly thought of as indicators of success. Not included, presumably, (presumably because we do not know the intermediate scale values), are such indicators as quality of teaching, service to the profession, grants awarded, number and quality of publications, etc. "Acceptance to a highly prestigious institution," is not usually thought of as the only or even the most valid indicator of long-term success. The incompleteness and irrelevancy of the measure of long-term success is clearly the most serious flaw in the study

Typical student comments which we included under this first objection are the following: "There seems to be little concern for the performance on the job in this research." "Those who drop out are automatically excluded from being judged successful." "It would be possible for a student to withdraw from the program and later continue the study of psychology and be successful."

5. Answer:

5. Answer cont'd.

Second, we believe that the success measure would be strengthened if it took into account at least the program from which each person has graduated. For example, a long-term success measure for a graduate of the clinical program might be number of patients or fee charged per client.

Third, we agree with one student who points out that the investigators' long-term measure is, "...not measuring long term success; it is merely rating a student as to the circumstances under which he left the University. To measure long-term success his career has to be followed up after he left." Other students worded this objection as follows: "Success cannot be measured immediately after graduation. Determination of long-term success must be made after a period of time has elapsed." "Notation was made of where a student intended to go but no follow-up on the students was made." "The 'long-term success is not long enough. A person may have accepted a prestigious position, but may not have been able to retain it." "The use of the word long-term is unfortunate. The long term aspect of the question would deal with careers."

Fourth, it should be noted that the 9-point scale is not fully identified. We can only speculate as to what description (if any) is given to the intermediate points. As we indicated earlier in our discussion of ratings of short-term success (see our answer to Question 4 c), the points on the long-term success scale do not seem to be tapping a common dimension. It is difficult to know where to place a person who, for example, drops out of a program but yet demonstrates "success" in other ways.

Fifth, as one student pointed out, "The definition may be rejected on the basis of the narrow sampling of experts used in determining what is and what is not success. They are professors at the same institution in the same department who are probably prone to similar thoughts on an issue such as this." The restricted

5. Answer cont'd.

5. Answer cont'd.

nature of the sample (number not given) of faculty whose opinions were used in developing the long-term success scale increases the likelihood that the criterion will not seem appropriate to other faculty groups.

Finally, note a built in dependency between short and long-term criteria. If a person drops out of a graduate school he must necessarily receive a low rating on both the short-term and long-term assessment. Thus, the relationship between long-term and short-term criteria is almost predetermined even though the study of this relationship is presented as a primary objective of this research.

Student Responses. "With only two judges, what does a reliability of .95 mean?" "I have difficulty with the 'inter-judge reliability' which was .95 when there were only two judges."

Our Reply. The records of the 42 students were rated on the 9-point scale twice, once by each judge. The correlation coefficient computed on these 42 pairs of ratings was .95. These two judges agreed almost perfectly on the relative ratings assigned to the students. A reasonable inference is that the high agreement resulted from a clear definition of the scale points.

b) The senior author of the article, in personal communication defended the definition of long-term success by writing: "I believe it is important to be able to predict things like which graduate students are most likely to flunk out vs. withdraw vs. get a Ph.D. and take a job at Podunk University vs. get a Ph.D. and accept a job at a prestigious university such as Cornell. Certainly the faculty of graduate schools feel that such 'long-term' criteria are important."

5. Answer cont'd.

6. Question:

Are investigators permitted to define key terms (such as "long-term success") any way they wish? Explain why you answered as you did.

6. Answer:

Our first three objections to the investigators' definition of "long-term success" (see our answer to Question 5 a) suggest what we think the term to mean. In commenting on our critique, the senior author wrote us: "Long-term, which to us meant 'after the end of a student's graduate education' apparently implied some kind of career-long perspective to Messrs. Millman and Gowin. They can mean whatever they want to, but in discussing our study I'd suggest they talk about our operational measure (which indeed has some problems) rather than focus entirely on the name we put on our measure."

We would agree that investigators should be permitted to define their terms as they like. On the other hand, they do have an obligation to foster accurate communication about their work and this goal is not sufficiently achieved when labels are used which convey meanings markedly different from those intended. In such situations, it is often possible to be misled into thinking that accounts of a study are more generalizable and significant than they actually are because persuasive labels (such as long-term success) are given specific meanings.

7. Question:

A student reviewer of this study stated that this investigation merely demonstrates what was already common knowledge. Do you agree? Support your answer.

6. Question:

Are investigators permitted to define key terms (such as"long-term success") any way they wish? Explain why you answered as you did.

6. Answer:

7. Question:

A student reviewer of this study stated that this investigation merely demonstrates what was already common knowledge. Do you agree? Support your answer.

7. Answer:

We disagree. We were surprised, for example, that the investigators found negative correlations between undergraduate grades and their "global assessment of success" rating. In spite of our misgivings about this "long-term success" index we would have anticipated at least small positive correlations. Further, we would not have expected rated quality of institution to be such a good predictor of long-term success. Many readers would not have predicted these findings and they could not be considered common knowledge.

Student Responses. "I thought that facility in a foreign language would be a great asset towards long term success in doctoral studies." "I was surprised that GRE-Quantitative and GPA mathematics correlated .00." "I didn't expect there to be so many negative correlations."

Our Reply. There is probably very little that is common knowledge. Like beauty, surprise is in the eyes of the beholder.

8. Question

Question 8 a) through 8 d) are based upon the following sentence quoted from the first paragraph on page 371:

> For example, the quality of the undergraduate school, which predicted first year grades negligibly, was found to be significantly related to student and faculty global assessments of progress toward the Ph.D. at the end of the year, and correlated more substantially with the long-term criterion of success.

Look again at the tables in the report of the study.

7. Answer:

8. Question:

Question 8 a) through 8 d) are based upon the following sentence quoted from the first paragraph on page 371:

For example, the quality of the undergraduate school, which predicted first year grades negligibly, was found to be significantly related to student and faculty global assessments of progress toward the Ph.D. at the end of the year, and correlated more substantially than any other predictor with the long-term criterion of success.

Look again at the tables in the report of the study.

8. Question cont'd.

a) Find those numbers which indicate the degree of relationship between quality of the undergraduate school and the other variables mentioned in the quotation. What numerical values of correlations did the investigators find to lead them to make their statement which is quoted previously?

a) Answer:

From the last line of Table 1, the "negligible" correlations between rated quality of undergraduate school and graduate school grades are: .00, -.13, .21, .16, and .15. The significant relations to students and faculty ratings of progress toward the degree are .30 and .31, also found in the last line of Table 1. The same line also shows a .43 correlation with the long-term measure of success.

8b) Question:

Can a negligible correlation be statistically significant from zero?

8 b) Answer:

Yes. Some people would say that for prediction purposes the significant correlations referred to in the quotation of .30 and .31 are negligible. The "negligible" correlation of .21 would be statistically significant if 88* instead of 42 students were involved in the study. Correlations of .30 and .31 would not be significant at the 4]/2% * level of significance. We wish to make two points. 1) In this context, negligible is an adjective describing the magnitude of a correlation; significance describes a different attribute - the likelihood of correlations of a given value occurring in a random sample of a population in which the actual correlation is zero.

* Arrived at by a t test of the significance of a correlation from zero.

8. Question cont'd.

a) Find those numbers which indicate the degree of relationship between quality of the undergraduate school and the other variables mentioned in the quotation. What numerical values of correlations did the investigators find to lead them to make their statement which is quoted previously ?

8. Answer:

8 b) Question:

Can a negligible correlation be statistically significant from zero?

8 b) Answer:

8 b) Answer cont'd.

All four combinations of these two descriptive adjectives are possible: negligible, significant; negligible, not significant, not negligible, significant; not negligible, not significant. 2) One must be careful not to have an unthinking attachment to correlations (or other statistical indices) which are barely statistically significant at some level of confidence, and distain for correlations which do not quite make the cut-off between significant and not significant. (Note our answer to Question 8 d).

8 c) Question:

For these 42 students, which pre-enrollment predictor was able to predict the long-term global assessment criterion most accurately? On what evidence do you base your answer?

8 c) Answer:

Undergraduate GPA in the physical sciences is the best predictor. The correlations between the pre-enrollment predictors and the long-term success criterion are given in the last column of Table 1. GPA in the physical sciences has the highest correlation (-.50) and, thus, would predict the criterion best. Because the correlation is negative, students having a low GPA in physical science would be predicted to have the highest long-term success rating and vice versa. (The predictor having the highest positive relation with the long-term success criterion is, as the investigators state in the sentence quoted, quality of the undergraduate school.)

8 d) Question:

If the appropriate statistical test were run, guess whether the quality of the undergraduate school would correlate with the long-term success criterion significantly more (in the statistical sense) than, say, the Quantitative score on the GRE?

8 b) Answer cont'd.

8 c) Question:

For these 42 students, which pre-enrollment predictor was able to predict the long-term global assessment criterion most accurately? On what evidence do you base yous answer?

8 c) Answer:

8. d) Question:

If the appropriate statistical test were run, guess whether the quality of the undergraduate school would correlate with the long-term success criterion significantly more (in the statistical sense) than, say, the Quantitative score on the GRE?

8 d) Answer:

This difference is not statistically significant. It is true that both correlations are significantly different from zero at the 5% level. This fact is indicated by the asterisks affixed to the correlations of .32 and .43 in the last column of Table 1. The point being illustrated is that correlations which are significantly different from zero need not be, and indeed frequently are not, significantly different from each other. A clear cut example might be two correlations of .80 and .79 being each significantly different from zero but having a difference (.01) that is not significant. The investigators did not test the difference between correlations in any of the tables and statements comparing the relative sizes of them should be made cautiously.

9. Question:

One of the findings of the study, as pointed out in Question 8 earlier, is that rated quality of undergraduate institution has a fairly high correlation with long-term success. Does this mean that if you were an admission officer in the Psychology Department at the University of Illinois and primarily interested in this measurement of success you should give preference to students coming from highly rated undergraduate institutions? Why or why not?

9. Answer:

If, and this is a big if, you were interested in predicting this long-term success measure, then yes, you should give preference to students coming from undergraduate institutions rated highly by the same (vaguely described) procedures used in this study. (Quality of undergraduate school should not be the only factor considered, of course.) It is true

9. Question

One of the findings of the study, as pointed out in Question 8 earlier, is that rated quality of undergraduate institution has a fairly high correlation with long-term success. Does this mean that if you were an admission officer in the Psychology Department at the University of Illinois and primarily interested in this measurement of success, you should give preference to students coming from highly rated undergraduate institutions? Why or why not?

9. Answer:

9. Answer cont'd.

that this high correlation may not show up in another sample, but chances are better that the variable will be positively related than that the relationship will disappear. The existence of a high correlation between quality of undergraduate institution and long-term success does not mean that the quality of undergraduate institution caused the students to have long-term success - indeed, the same forces which are responsible for a student's selecting (or being selected by) a highly rated institution might be operating at the time he selects (or is being selected by) his employer upon his graduation.

Student Responses. "I would not give any preference to those students coming from a highly rated institution. What a person puts into an institution is what he will get out of it." "Absolutely not. To me the entire record should be evaluated and equal weight given to all variables, insuring fairness and a chance for the student to achieve this goal if he really has the desire to try." "As an admission officer in the Psychology Department at the University of Illinois, I wouldn't show a preference to students coming from highly rated undergraduate institutions. I would, however, carefully consider all information in the folders of all applicants." "The GREs would be important to me as a comparison of the individual students so this would definitely affect my decision." "I would not give any preference to students coming from highly rated institutions. The rating scale was too narrow and biased."

Our Reply. The above comments of student readers appear to be a denial of the facts in the case; namely that the quality of undergraduate institution was predictive of the long-term success measure. The rating scale may indeed be narrow and biased, but it worked. Most of the other information in the folders, particularly undergraduate grades and some GRE scores, have either low or negative correlations with the success measure or unknown predictive validity for this criterion. That is, there is not sufficient evidence to believe that these

9. Answer cont'd.

9. Answer cont'd.

other variables will be effective in predicting the student who will be rated high on the long-term success measure. One can find good reasons for objecting to the appropriateness of this criterion, but that is not the issue under consideration. (Reread Question 9.)

Student Responses. "It is more likely that strong individuals are selected by high quality institutions. Therefore, the judgment should be made on the basis of the individual and not the institution." "The long-term 'success' of students coming from highly rated undergraduate institutions may be due in part to the self-fulfilling prophecy. Students coming from a highly rated undergraduate institution may get hired by prestigious universities because such universities may expect him to do well merely on the basis of this undergraduate school."

Our Reply. The student who made the first response overlooks the fact that it was the institution and not the "individual" predictors that worked. Both students, and many other readers whose responses we did not quote, quite properly attempted to explain the reason for the success of the undergraduate institution quality rating as a predictor.

Student Response. "Students must be selected for more justifiable reasons than what the names of their undergraduate schools were. A better measure must be found."

Our Reply. In this student reader comment the frustration of many of us is given expression. One student discussed the dilemma in these words:

> Being intelligent, perceptive, and liberal, I, of course, would not discriminate against a student from a low-rated school. However, if I had to bet on which student would succeed, I would go with a student from a high-rated school. These statistics indicate I would have a better chance of winning.

9. Answer cont'd.

9. Answer cont'd.

9. Answer cont'd.

An example of another area of concern may help to clarify the situation. Suppose you own a company that produces screws and nuts (metal variety)! The more screws and nuts turned out by an employee, the more money you, as a company owner, will earn. Your personnel office reports a company study in which, let us assume, race correlates .43 with production - white employees producing more per man hour than black employees. Now, you know that skin color per se is not the cause of this differential production, and that there is some more basic reason. But, while you ponder the underlying causes, a vacancy occurs and two applicants, one black and the other white, apply for the job. The applicants are equal on all other factors you usually consider. Whom do you chose? If money is the only criterion you would "bet" of the white man. If, as company owner, you are willing to consider more unselfish motives relating to, say, society's needs, you might give the black applicant a chance.

None of the predictors in the present graduate school study tells the whole story and they may well discriminate against the poor, the late bloomer, and the special student. An admissions officer may try to be fair to such people by deviating from total reliance on the best predictors available to him by choosing individuals with a lower probability of success. When he does so, it is because he feels that criteria other than his success measures are important. Of all institutions, educational ones are perhaps best able to afford using multiple indicators of success.

It would certainly be nice if each person could, "...be given the chance to fail or to succeed on his own without a survey telling him he can or cannot do it." When demand greatly exceeds supply (e.g., only some of the applicants to graduate school can be admitted), not everyone can be given that chance, and choices have to be made.

Appendix I

Francis P. Hunkins

The Influence of Analysis and Evaluation Questions on Achievement in Sixth Grade Social Studies

Educational Leadership Research Supplement
January 1968, p. 326-332

The Influence of Analysis and Evaluation Questions on Achievement in Sixth Grade Social Studies

Francis P. Hunkins

Educational Leadership Research Supplement

January 1968, p. 326-332.

SPECIAL NOTES

Page 326. The Taxonomy of Educational Objectives is a book by Benjamin Bloom and others in which are described types of cognitive abilities organized into the following categories: knowledge, comprehension, application, analysis, synthesis and evaluation. Each of these categories is further subdivided into more specific skills and abilities. The authors of this volume hypothesize that these six major categories are hierarchically arranged with knowledge at the bottom of the scale and evaluation at the top, and with each step of the hypothesize scale dependent upon mastery of previous categories. Thus, for example, they/ that an individual cannot properly evaluate (category 6) a statement about, say, atoms without first learning certain facts about atoms (#1), comprehending certain ideas about them (#2), being able to analyze these facts and ideas (#4), and so on.

Page 330, Table 1. Recall from the design that there were two treatments (Condition A in which questions requiring analysis and evaluation were stressed, and Condition B in which questions requiring only knowledge were in the majority), four reading levels, and the two sexes. Each student was considered to fall into one of these 16 possible categories (i.e. 2 x 4 x 2 = 16). One category, for example, would be Condition A, reading level 2, girl.

Table 1 shows the results of a statistical analysis designed to test if there were significant differences in achievement scores among students in certain combinations of the categories. In each of the first seven rows of the table are reported the results of an analysis involving a different such comparison. The Source of Variation column identifies the comparisons involved. By Treatment is meant the comparison between Condition A and Condition B, or more precisely, between the scores of students in the eight categories involving ConditionA with the scores of students in the eight categories involving Condition B. Similarly, the reading level comparison involves a test of the

significance of the differences among the scores of students in the four reading level conditions. If boys score significantly higher than girls (or vice versa), it will be reflected in the results of the sex comparison shown in row three. By statistically significant is meant that the differences are sufficiently large that it is unlikely that they could occur by random sampling, or by chance alone.

The next four rows involve interaction comparisons. Since these interaction effects were neither significant nor of much concern in this study, they will not be discussed further here. The concept of interaction is discussed in regard to other articles in this series.

The last effect represents differences in scores among the students within each of the 16 categories. These differences are not tested for significance; rather they serve as a base from which to evaluate the other differences associated with the first seven effects.

The investigator performed two kinds of analyses - one involving the actual achievement scores and the other involving scores that were "adjusted" for pretest score differences. In both cases, the d.f. column represents degrees of freedom which relate to (but do not exactly equal) the number of groups being compared. The S.S. columns and the M.S. column stand for the sum of squares and mean square respectively, and are intermediate calculations in the analysis of variance.

The numbers in the F column are used to indicate if results are statistically significant. For a given number of degrees of freedom (d.f.), the higher the F number the more unlikely that chance alone could account for the differences, and the more statistically significant the results. You'll note that the difference in test scores (when adjusted for preachievement score differences) of students in Conditions A and B, and of students in the four reading conditions, were statistically significant.

Page 330, column 1, lines 14 and 15. The symbol, $>$, means "greater than" and the Q represents quarter. $Q_4 > Q_3 > Q_2 > Q_1$ means that the mean achievement scores of students in the fourth (i.e. top) quarter in reading was significantly greater than the mean score of students in the third quarter in reading, and so on. Just above the Results section toward the bottom of page 329, the investigator incorrectly uses the word, "quartile", to mean quarter. The first quartile is 31.5, the point below which 25% of the scores lie. The first quarter of scores covers the range 0 through 31.

"The Influence of Analysis and Evaluation Questions on Achievement in Sixth Grade Social Studies."

Francis P. Hunkins
Educational Leadership Research Supplement
January 1968, pp. 326-332

1. Do you think the title a good one? Why?

A good title for a research report will describe the contents of that report as accurately and as completely and consisely as possible. This particular study is an investigation of the effects of several kinds of questions asked upon subsequent achievement. The written materials used and questions asked dealt with social studies, the grade level was sixth; relationships with reading level and sex were investigated as well as teacher differences and pretest scores for students. Not all of these elements can be easily mentioned in a title and therefore the investigator must choose those he considers most essential for inclusion.

Not clear from the title was that the principal independent variable was the kind of question asked and answer provided. A better (but not the best) title would have been, "A Comparison between Knowledge and Higher-level Questions and Answers on Achievement in Sixth Grade Social Studies." But other than this concern that a description of the manipulated variable be given priority in the title, we think the title a fairly good one.

Some students objected to the use of the word "influence" in the title and felt that the study was inconclusive and influence was not demonstrated. We do not share this concern because a title need not convey the specific finding, only the intended problem. Thus, a study with a title that begins, "The Relationship Between," might have as its finding that there is no relationship between the variables investigated. Although titles such as "A Study of the Influence of..." and "An Investigation of the Relationship Between ..." would be less ambiguous, it is accepted practice to use the abbreviated version.

1. Do you think the title a good one? Why?

Answer:

2. Reread the Introduction. Does this research provide a test of the heirarchical hypothesis implicit in the Bloom et al. taxonomy? Give reasons for your answer. (See Special Notes for a discussion of this hypothesis.)

The study does not prove the hierarchical hypothesis is true; nor does the investigator claim that it does. Since only three levels are involved, this research can not provide a complete test of the hierarchical nature of all 6 levels. Further, just because the research is "concerned" with Bloom's taxonomy does not mean it provides a test of it. Whether the study even gives some support for it is a question about which experts disagree. Some say NO and argue that the hierarchical hypothesis is assumed to be correct and merely used as a starting point about which the research is organized. Others say YES and argue that the results are consistent with the hierarchical hypothesis and thus give some support for its validity.

A lesson to be learned from this specific question and answer is that a criterion for a hypothesis to be tested is the presence of data which can count as evidence in support of or against the hypothesis. Those who answered YES should be able to point out such evidence. The personal views expressed by many students that the hypothesis is most reasonable and that the distinctions among the cognitive levels are very important do not, in themselves, justify the conclusion that the hypothesis was being tested by the research.

3. Reread the Objectives section. Do you think the overall hypothesis is a clear and accurate statement of the hypothesis the investigator wishes to test?

The statement is probably quite accurate, although awkwardly phrased. We would have preferred deletion of the ending: "... in relationship to..." A separate sentence could have been added to describe these secondary concerns.

2. Reread the Introduction. Does this research provide a test of the hierarchical hypothesis implicit in the Bloom et al. taxonomy? Give reasons for your answer. (See Special Notes for a discussion of this hypothesis.)

Answer:

3. Reread the Objectives section. Do you think the overall hypothesis is a clear and accurate statement of the hypothesis the investigator wishes to test?

Answer:

3. cont'd

We might note that the standard procedure is to express the hypotheses in the direction the investigator really expects them to be true, rather than in the "null" form used. The use of the null form is not incorrect, but it does represent upsophisticated reporting. It is the substantative question (hypothesis) the reader wants to know about.

4. Reread the section, General Plan of the Study.

a) Did the investigator construct the materials about Africa and Oceania? Give reasons for your answer.

b) Were the questions used in the instruction of the multiple-choice type? Give reasons for your answer..

c) The investigator attempted to reduce the influence of the teacher on the experimental situation by avoiding active teacher participation. Was this wise? Give reasons for your answer.

a) It is true that, "...two sets of text-type materials...," were constructed by the investigator. However, these sets stressed questions and must have been widely supplemented by other materials, primarily the textbook. Note that: "Pupils in both treatment conditions were directed to read designated sections of their textbooks..." We believe that the special instructional materials constructed by the investigator consisted only of questions and answers.

b) No, at least not all the questions used in instruction were of this type. Note that the pupils had, "...to respond in writing to the questions on their worksheets." This suggests that students had to construct their responses rather than simply select their responses as it is the case with multiple-choice questions. (Do not confuse the criterion test of achievement [which did consist of multiple-choice questions], with the questions asked as part of the instruction.)

4. Reread the section, General Plan of the Study.

a) Did the investigator construct the materials about Africa and Oceania? Give reasons for your answer.

b) Were the questions used in the instruction of the multiple-choice type? Give reasons for your answer.

c) The investigator attempted to reduce the influence of the teacher on the experimental situation by avoiding active teacher participation. Was this wise? Give reasons for your answer.

Answer:

4. cont'd

c) Yes and no. By reducing teacher participation, the investigator can be more certain that the differences in scores of students using the two sets of materials are actually due to the experimental variable, type of question asked. We say that the study is more likely to have internal validity. However, the price paid for this internal validity is a lessening of the external validity because the study results may be reliably applied to limited classroom practices. By minimizing the role of the teacher we cannot determine what the effects might be if teachers asked the different types of questions rather than presenting them in written form alone. The investigator has gained control at the expense of conducting the investigation under fairly narrow and less typical conditions. Many research experts argue, as this investigator evidently does, that it is more important to guarantee that the comparisons made are valid even though this validity necessitates confining the research to a study of less typical practices. But compromises must be made and we certainly do not fault the investigator for restricting the role of the teacher.

5. Reread the section, Subjects.

a) Note that the proportion of boys to girls (67:60) in Condition A does not equal the proportion (55:78) in Condition B. Does this fact mean that the comparison on the criterion achievement test between students in Conditions A and B is misleading? Why:

b) "Background data were collected and analyzed for both pupils and teachers." What data were collected and was it important that the investigator analyze them?

c) Primarily for students having had a course in statistics: On page 328, column 1, the investigator indicates that his criterion for determining whether a background variable should be used as, "...a possible covariant on subsequent analyses

5. Reread the section, Subjects.

a) Note that the proportion of boys to girls (67:60) in Condition A does not equal the proportion (55:78) in Condition B. Does this fact mean that the comparison on the criterion achievement test between students in Conditions A and B is misleading? Why?

b) "Background data were collected and analyzed for both pupils and teachers." What data were collected and was it important that the investigator analyze them?

5. cont'd

of the criterion data...," is whether pupils in the two conditions differ significantly (in a statistical sense) on that variable. Is this a good criterion to use? Why?

a) Unless controlled for in the analysis, a comparison between all the students in the two conditions would be misleading if boys and girls do not perform equally on the dependent variable. Since performance on the criterion test is related to reading ability, and since the girls in this study were reported to be better readers than the boys, it is not unreasonable, therefore, to expect that on this basis boys and girls will score differently on the criterion test. Thus, condition B with a higher proportion of girls, could have an unfair advantage. As it turned out, such bias is a little concern since Condition A was still judged to be superior to Condition B in spite of the possible advantage given to the latter treatment.

One way to control for these differences is to report criterion test scores separately for boys and girls. Another way is to weigh equally each of the 16 subcategories. (See Special Notes for a description of these categories.) The investigator did not indicate the procedure he followed to handle the disproportionate frequency-in-categories problem. If an acceptable procedure for controlling the disproportionate number of boys and girls in the two conditions were used, then the comparison between conditions is <u>not</u> misleading.

b) Pupil I. Q. and reading test scores were mentioned as well as (in the Results section) some kind of pre-test. Information about age, teaching experience and college degree was obtained from the eleven teachers. Other information concerning both pupils and teachers may have been collected but was not reported.

Yes, it was important that such data were collected and analyzed, especially in the case of the pupils. Since only eleven classes were involved in the study, and thus gross inequalities between groups possible, it is important

c) Primarily for students having had a course in statistics: On page 328, column 1, the investigator indicates that his criterion for determining whether a background variable should be used as, "...a possible covariant on subsequent analyses of the criterion data...," is whether pupils in the two conditions differ significantly (in a statistical sense) on that variable. Is this a good criterion to use? Why?

Answer:

5. cont'd

to know how these background variables differ for conditions A and B. Student data are also of value for the pupose of understanding the limits of permissible generalizations. Since the teacher influence was minimal, teacher differences are not so important as pupil differences.

c) This is a technical question, the answer to which you are not necessarily expected to know. Our answer is no. The criterion which was quoted assumes, incorrectly, that failure to reject the null hypothesis is equivalent to establishing its truth. Merely because the differences in reading scores and I. Q. scores for students in Condition A and Condition B are not statistically significant does not mean the two groups are identical in these regards. There is sufficient difference between the groups which could go a long way toward explaining the difference on the criterion variable. Further, the investigator fails to recognize another important reason for including a covariate--namely, to increase the precision (power) of the statistical test. Although beyond the score of these notes, suffice it to say that even if the two groups were equal on these background variables, it would still be a good idea to employ them as covariates in order to increase the likelihood of a true difference on the criterion variable being detected.

6. The construction of the criterion test of achievement is described in the section, Collection of Data. Reread this section.

a) Do you agree that, "...only the total achievement score was of concern in this phase of the investigation."? Give reasons for your answer.

b) From a pool of 59 items, 42 items were selected and 17 were eliminated. On what basis was the decision made to accept an item? On what basis were the 17 items eliminated?

6. The construction of the criterion test of achievement is described in the section, Collection of Data. Reread this section.

a) Do you agree that, "...only the total achievement score was of concern in this phase of the investigation."? Give reasons for your answer.

b) From a pool of 59 items, 42 items were selected and 17 were eliminated. On what basis was

6. cont'd

c) Do you feel that each item measured the level of cognitive ability that was intended? Why?

d) How important is it that this classification task be done accurately? Explain.

e) Publishers of tests used to making decisions about individuals often consider reliability indices of .90 or more as high (i.e., good) and indices of less than .70 as poor. The investigator seems to be unhappy with a reliability index of .68. Should he be? Why?

a) Absolutely not! Although the investigator may expect achievement to be better on all questions for pupils in Condition A, surely he must expect the most dramatic differences to occur on questions tapping higher level skills. When possible, as in this case, it is important to provide data which relate to predictions growing out of one's conceptualization of what is going on. Failure to provide mean and variability measures for both groups on all subtests is a serious weakness of this study. (In a later study the investigator provides such data.)

b) We are told that there was almost unanimous agreement on the classification of the 42 items actually included in the criterion test. We are not told, however, the reasons for excluding 17 or the original 59 items and we can only assume that at least some of them were eliminated because the judges could not agree on the appropriate level.

c) We remain skeptical especially that the higher level abilities of synthesis and evaluation were actually measured since it is most difficult to devise multiple-choice questions which truly measure these skills. Further, since the instructional materials were different for the two conditions, it is possible that a single question could be measuring at different cognitive levels for

the decision made to accept an item? On what basis were the 17 items eliminated?

c) Do you feel that each item measured the level of cognitive ability that was intended? Why?

d) How important is it that this classification task be done accurately? Explain.

e) Publishers of tests used to making decisions about individuals often consider reliability indices of .90 or more as high (i.e., good) and indices of less than .70 as poor. The investigator seems to be unhappy with a reliability index of .68. Should he be? Why?

Answer:

6. cont'd

different students because of this differential prior instruction. For example, suppose that a question and answer used in the instruction under Condition A concerned the evaluation of a particular content area. A question in the criterion test asking for an evaluation of a similar area would not be as novel a task for students in Condition A, and thus for those students would not be measuring at this "highest" level (evaluation) but rather it would be measuring lower level skills. Just because an item contains the word "evaluate" does not mean that it will necessarily measure a student's evaluative ability. It is indeed unfortunate that no examples of questions from the instructional materials and the criterion test were shown as evidence that the investigators were able to overcome these difficulties.

It is, of course, important that competent judges be used to classify the items. One point made above is that proper classification requires more than competent judges. The items cannot be classified accurately into the categories employed in this study without knowledge of the students' prior instruction.

d) Had subtest scores been reported, as we suggested they should be in our answer to 6a, then the correct assignment of item to taxonomy category would have been very important indeed. Since only the total score was reported and the same criterion test was given to both groups, it probably wasn't important that the six categories be equally represented and some classification mistakes certainly could be tolerated.

e) This is a technical question, the answer to which you are not necessarily expected to know. Our answer is NO. A high reliability coefficient is not required for a criterion measure in a research study comparing groups. Here's why. High reliability assures us that differences in test scores are not due to measurement error. Unless a test has high reliability, then differences in an individual's test scores (used to measure

6. cont'd

"gain" or to determine his relative strong areas) may be due to measurement error. In a research study comparing groups, what are being compared are not the differences between two individual scores but rather the differences in means, each based on the scores of many individuals. Although any one score may have measurement error, the "too high" and "too low" errors will balance out over many people, leaving us quite confident that this mean score is fairly free of measurement error.** That is why we can tolerate a lower reliability in the measures we use in research studies. The value of .68 reported by the investigator is quite acceptable.

7. Reread the section on Experimental Material and Procedure on pages 328 and 329.

a) At the top of page 329, the investigator indicated that it was important that the unit to be studied be one about which the subjects did not have "...abundant prior knowledge." Do you agree? Give reasons for your answer.

b) Note that 47.53% of the questions used in Condition A were in the analysis and evaluation categories; in Condition B 87.38% of the questions were in the know ledge category. Alternatively, the

**Reliability means consistency of measurement. If a test consistently gives systematic errors, i.e., errors which are consistently too high or consistently too low, then we say the test is invalid, but still it can be reliable. Unreliability occurs because of random measurement errors, which, if averaged over enough people (or over many test items) will balance out. Means of many scores (or very long tests) are usually very reliable.

7. Reread the section on Experimental Material and Procedure on pages 328 and 329.

a) At the top of page 329, the investigator indicated that it was important that the unit to be studied be one about which the subjects did not have "... abundant prior knowledge." Do you agree? Give reasons for your answer.

b) Note that 47.53% of the questions used in Condition A were in the analysis and evaluation categories; in Condition B 87.38% of the questions were in the

7. cont'd

investigator could have had all the questions in Condition A in the analysis and evaluation categories and all the questions in Condition B in the knowledge category. Would this have been an improvement? Why?

c) Was the readability analysis a wise thing to do? Explain.

d) Is it possible to compare the content of the questions and answers used in the instruction to the criterion test questions? If not, is this inability a serious shortcoming? Give reasons.

a) Yes, we feel it was wise for the investigator to select a topic about which the students did not have abundant knowledge. Our reason is that using such a topic insured that the question and answering procedures would have a chance to make a difference because there were still many things the students could learn. In other words, if students already knew a great deal about a topic leaving little to learn before the study began, then one procedure of instruction could not be expected to result in more learning than the other procedure. A second reason is that when students have different levels of prior knowledge about a subject, it is difficult to construct items which will measure at the same cognitive level for all students. (Recall our answer to question 6c.)

Many students answered YES for a reason different from the ones we gave above. They felt that the students in the study with prior knowledge would have an unfair advantage and another extraneous factor would be introduced. This would certainly be true, but it should not be a cause for concern unless it is suspected that the students in one of the two groups had, on the average, more prior knowledge than the other group.

knowledge category. Alternatively, the investigator could have had all the questions in Condition A in the analysis and evaluation categoeies and all of the questions in Condition B in the knowledge category. Would this have been an improvement? Why?

c) Was the readability analysis a wise thing to do? Explain.

d) Is it possible to compare the content of the questions and answers used in the instruction to the criterion test questions? If not, is this inability a serious shortcoming? Give reasons.

Answer:

7. cont'd

Other students answered question 7a NO and remarked that knowledge of some facts is important for without such knowledge the students in the study could not be expected to analyze and evaluate. This is true, but the issue is not whether the students in the study should have this information (they should) but whether they should be given this information before they are exposed to the instructional materials.

b) Before giving our answer to Question 7b, note that the kinds-of-questions-asked represents, in this study, the variable which is under the control of the investigator--the variable being manipulated. In appraising the work of others, pay particular attention to the levels or conditions which are being used. The results depend upon it.

If it is true that asking higher level questions really makes a difference, then the alternative distribution proposed involving 100% or 0% in a given category, would give the investigator the best chance to discover differences between conditions. As one student put it, to do otherwise would, "...water down the effectiveness...," of the experimental treatment.

On the other hand, the ratios of the different kinds of questions the investigator chose to compare are more typical of what one would expect to find in existing materials (or in the questioning patterns of teachers) and what one would hope to find in materials that emphasized higher level questions. We personally approve of the investigator's decesion to make the balance of question types more closely resemble present and sound practices rather than to use conditions as different as possible. Clearly, use of either distribution is justified.

c) Yes, to conduct a readability analysis was a wise decision, although reporting readability data separately for the two groups would have been preferable. The readability analysis would have been important had the

7. cont'd

investigator failed to find differences in favor of Condition A. If that had occurred, one explanation for finding no differences, namely that the reading materials were too hard, could be ruled out by the fact that the mean reading level was within the range of fifth and sixth grade pupils. As one person answering question 7c put it, the readability analysis, "...knocked out the possibility of massive inability to comprehend the questions."

Further, if Condition A students had not done better than the other students, we might have wondered if the higher level questions and answers were more difficult to read. The plausibility of this explanation could be assessed by having readability figures shown separately for materials used in Conditions A and B.

Notice that a reading difficulty index was computed for the answers as well as for the questions. This alerts us to the fact that the answers are, in all probability, more than cryptic responses and that by providing answers, additional instruction must have been given. The implication of this will be evident in the answer to the next question.

d) No, we are not told how similar the questions asked in the criterion test were to the questions and answers given in the instruction. This omission is probably the most serious shortcoming of the study. We know only that during instruction different questions and answers were given to the two different groups. It is extremely difficult, if not impossible, to choose criterion test items upon which these differences in the instructional materials had no bearing. The fact the investigator does not mention this problem and report in detail how it was circumvented is a serious weakness. It suggests that the differences between the two conditions could be accounted for entirely by the different content of the instructional materials (and especially in the answers provided

7. cont'd

which were admittedly more complicated [p. 330, column 1] in the case of the evaluation and analysis questions than the knowledge ones) rather than by the practice of answering analysis and evaluation questions alone.

8. Reread the section, Analysis of Data, page 329. Do you approve of using sex reading achievement as additional variables in the analysis? Why?

Yes, inclusion of these variables helps to determine the generalizability of the findings. Because there was presumably little interaction between these variables and the treatment variable, it means that the differences between the scores under the two conditions seemed to be about the same for both sexes and across the reading groups. If, for example, the higher level questions and answers were relatively less effective with poor readers this would show up as an interaction between treatment and reading. By including reading and sex as variables in the analysis the investigator could identify the limits to the generalizability of the results in these respects.

It was also important to include these two variables because they were explicity mentioned in the statement of the hypothesis of the study (p. 327) and, thus, a complete test of this hypothesis requires their inclusion.

Some students made a good case in support of inclusion of one of the two variables. A separate analysis by sex was deemed necessary because of the difference in boy/girl ratios in the two treatment conditions. Others argued that the reading variable was very important to include because of the suspected relationship between reading and performance on the criterion task.

8. Reread the section, Analysis of Data, page 329. Do you approve of using sex and reading achievement as additional variables in the analysis? Why?

Answer:

9. Reread the results section on pages 329 and 330, including Table 1. (You may wish to review the Special Notes regarding the interpretation of Table 1).

a) Note that at the very bottom of page 329 the investigator refers to pre-achievement scores. Did he ever report how these scores were obtained? Regardless whether or not he did so, do you think it was a good idea to obtain such scores and once obtained, should they have been used? Why?

b) Find the number 9.85 in the F column of Table 1. Is the difference in mean scores for students in the two treatment groups statistically significant? Here's a question you may not know the answer to. Does the number 9.85, by itself, indicate which treatment group performed better?

c) Does the investigator ever indicate the numerical value of the differences in mean scores for students in the two treatment conditions? If so, what is it? If not, should he have done so?

a) The nature of the pre-achievement scores was not specified. They could have been previous achievement grades in social studies. They could have been scores on the criterion test administered before the textbook and special materials were used. If the latter is the case, there is a slight danger that seeing the criterion test ahead of time would be of greater help to students in one condition than in the other. The investigator did mention on page 328, however, that the reading and I. Q. scores were not used as the covariates; that is, they were not used as the pretest.

Once the pre-achievement scores were obtained, it was a good idea to adjust the criterion scores on the basis of differences on the pre-achievement scores not only to equate the groups, but (as mentioned in answer to Question 5c) to give greater power

9. Reread the results section on pages 329 and 330, including Table 1. (You may wish to review the Special Notes regarding the interpretation of Table 1.)

a) Note that at the very bottom of page 329 the investigator refers to pre-achievement scores. Did he ever report how these scores were obtained? Regardless whether or not he did so, do you think it was a good idea to obtain such scores and once obtained, should they have been used? Why?

b) Find the number 9.85 in the F column of Table 1. Is the difference in mean scores for students in the two treatment groups statistically significant? Here's a technical question you may not know the answer to. Does the number 9.85, by itself, indicate which treatment group performed better?

c) Does the investigator ever indicate the numberical value of the differences in mean scores for students in the two treatment conditions? If so, what is it? If not, should he have done so?

Answer:

9. cont'd

to the analysis. We find it strange that the investigator did not tell us how the two groups differed on these pre-achievement variables or describe them more clearly.

b) This F number, as indicated by the **footnote, signifies that the means for the two treatment groups are significantly different. The so-called F test, however, does not indicate which group scored higher but only that the differences could not reasonably be accounted for by chance alone. We have to look at the mean values to find out which group did better. In this report, we must rely on the statement in the text that Condition A pupils performed better.

c) A surprising deficiency is the failure to report the criterion means for the two groups. We don't know if the difference in means is large or small. To find statistical significance in mean differences is only the initial step in a proper interpretation of a research study. If the difference were as much as 1/2 a standard deviation (a variability measure like the standard deviation should also have been reported) the difference would have important practical implications; if the difference were only 1/100th of a standard deviation, even though the difference was statistically reliable, it would lack much practical significance. The magnitude of the differences should definitely have been given.

**Note that the number 10.05 in Table 1 is not the difference in group means. Rather, it is the result of an intermediate calculation in the analysis of covariance.

10. Reread the discussion section. column 1, page 331. The investigator indicates that this study suggests the following: that questions requiring analysis and evaluation, "...stimulated individuals to utilize general viewpoints regarding the information embedded in the task."; forced "mental juggling" of the materials; led

10. Reread the discussion section, column 1, page 331. The investigator indicates that this study suggests the following: that questions requiring analysis and evaluation, "...stimulated individuals to utilize general viewpoints regarding the

10. cont'd

to greater "...interaction with the materials presented," and have the potential, "...to make pupils uneasy." What evidence supports these suggestions?

None that we know of. It is true that students given the greater proportion of "analysis" and "evaluation" questions and answers performed better on the criterion test. But the study was not designed to determine how this superior performance came about. The statements of the investigator quoted by us in Question 10 represent admitted guesses on his part of changes occurring inside the student rather than assertions based on reported evidence. It is quite acceptable for an investigator to report his speculations as long as they are clearly labeled so that the reader can recognize them as unsupported views.

information embedded in the task."; forced "mental juggling" of the materials; led to greater, "...interaction with the materials presented," and have the potential, "...to make pupils uneasy." What evidence supports these suggestions?

Answer:

11. Assume that the research were redone so as to overcome the criticisms mentioned earlier and that similar findings in favor of Condition A resulted. What limitations would still remain to this single study which would prevent one from generalizing with confidence that questions of higher cognitive levels generally stimulate higher achievement?

The study investigates one topic, in one subject area, for students in one grade, from one surburban school system. Further, it is limited to written self-instructional materials and we don't know if the findings would hold up for the situation in which teachers ask the same questions. Further, only achievement immediately after study was measured. Of more importance is the long-term impact as measured by a delayed post-test. A single study cannot have universal applicability. This study did look at both sexes and various reading levels. We do not fault the investigator for not including more topics and delayed post-testing, etc. We only mention these "extensions" to alert you to those situations to which the results might not apply.

11. Assume that the research were redone so as to overcome the criticisms mentioned earlier and that similar findings in favor of Condition A resulted. What limitations would still remain to this single study which would prevent one from generalizing with confidence that questions of higher cognitive levels generally stimulate higher achievement.

Note: If you are interested in reading a review of the research on the effect of questions on learning, see the December 1970 issue of the Review of Educational Research.

Answer:

Appendix VI

Dolores Durkin

Children's Concepts of Justice: A Comparison with the Piaget Data

Child Development, 1959, 30, 59-67

CHILDREN'S CONCEPTS OF JUSTICE: A COMPARISON WITH THE PIAGET DATA

Dolores Durkin

Child Development, 1959, 30, 59-67

QUESTIONS

1. Appraise the educational significance of the study. In this context, by educational significance we mean the import for those responsible for the education of children and what they might do as a consequence of the assertions established by the study.

2. Give your critical appraisal pro and con concerning how well the investigator has accomplished the first named principal purpose for the study (as indicated in the Special Notes). Evaluate the adequacy of the design (subjects, interview procedure), appropriateness of the analyses (categorization scheme, statistical tests), and the validity of the interpretations and conclusions (of both the Rampert and the present studies).

3. Note the second purpose for the study (as indicated in the Special Notes). Briefly evaluate how well this purpose has been accomplished. Pay particular attention to the investigator's notion of intelligence.

CHILDREN's CONCEPTS OF JUSTICE: A COMPARISON WITH THE PIAGET DATA

Dolores Durkin

Child Development, 1959, 30, 59-67

Special Notes

Introductory Section:

The investigator has indicated two principal purposes for the present study. The primary purpose is, using American children, to test Piaget's empirical claims that children up to about 8 years of age typically appeal to adults to redress wrong and to provide appropriate punishment; that from 8 to 11 they shift to an equalitarian notion of justice characterized by reciprocity (an eye for an eye); and that from 11 or 12 onward they associate reciprocity with equity (retribution takes account of circumstances).

A second purpose of the study is to investigate whether intelligence, rather than chronological age, is the significant factor in the development of a child's concepts of justice.

Description of Piaget's Study:

A thorough critical appraisal of a research report requires familiarity with the context out of which the study comes. This is especially true of the present study which has as its focus a comparison with the research results from another investigation. The experiment being replicated was actually conducted by Mlle Rambert, and reported by Piaget in his 1932 book, The Moral Judgment of the Child. The results of that study which most relate to the Durkin paper may be found on page 302 of the Piaget book and are shown below.

PERCENT OF GIRLS AND BOYS RESPONDING IN VARIOUS CATEGORIES TO THE QUESTION, "IF ANYONE PUNCHES YOU, WHAT DO YOU DO?"
N = 167

	"It is naughty"		Give back the same		Give back more		Give back less	
Age	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys
6	82	50	18	37.5	--	12.5 --	--	--
7	45	27	45	27	10	46	--	--
8	25	45	42	22	8	33	25	
9	14	29	29	57	--	14	57	
10	--	8	20	54	--	31	80	7
11	--	--	33	31	--	31	67	38
12	--	--	22	67	--	10	78	23

The table can be read as follows: 82% of the responses of girls, age 6, were categorized as "it is naughty": the remaining 18% were placed in the category, give back the same. The children did not say "it is naughty" as a direct response to the question, "what do you do?" It is only when asked additional questions such as, "do you hit back?", that the child might respond, "it is naughty." Some children did say they would tell someone in authority as their first response to the "what do you do?" question. In such cases the determination of which of the four categories to use depended upon the children's replies to further questions.

On the basis of the above data and the transcriptions of the complete interviews, Piaget concludes that:

> ...the children who do not hit back (most of them are from the younger ones), are primarily submissive children who rely upon the adult to protect them and who are more anxious to respect or make others respect the orders that have been received than to establish justice and equality by methods appropriate to child society. As for the children who hit back, they are far more concerned with justice and equality than with revenge properly so called...Among those who give back more blows than they receive there is, of course, a combative attitude, which goes beyond mere equality: but it is precisely this attitude which diminishes with age. (p. 305)

Statistical Analysis of Durkin's Study

The author makes extensive use of chi square statistical procedures to test several hypotheses. In each one, the investigator is comparing the observed frequencies of responses to those expected on the basis of a chance distribution of responses or those expected if the variables of interest were not related. When the discrepancies between observed and expected frequencies are large (as evidenced by a large numerical value for the chi square statistic) the hypothesis of chance distribution or no relation is rejected and support is evidenced for a nonchance, or statistically significant relationship. The end product of a chi square test is a probability (p) of obtaining discrepancies between observed and expected frequencies as large or larger than those found in the study. A small probability of getting such large discrepancies (in this study, small is 5% or less) is the criterion for rejecting the chance relationship hypothesis and for claiming statistical significance.

Page 62, last paragraph. This paragraph describes the results of three chi square tests. The set-up for the test for grade two is shown below.

	Authority	Agression	Other	
Observed	15	8	5	28
Expected	(9 1/3)	(9 1/3)	9 1/3)	28

the numbers 15, 8, and 5 are taken from Table 1: the numbers in parentheses (9 1/3) are the expected frequencies. Under the chance distribution hypothesis, the 28 second graders would be expected to give responses in the three categories equally, or 28/3 or 9 1/3, responses in each. The end product for this 2nd grade significance test is reported at the end of the paragraph as ".05 < p < .10". The symbol, < means "less than". Thus, the discrepancies between the actual frequencies and 9 1/3 could be expected to happen by chance alone between 5 and 10% of the time. By the 5% criterion, these discrepancies were not statistically significant.

Page 63, first paragraph. Here is described the second use of the chi square test. The set-up is shown below. The expected frequencies under the hypothesis of no relationship between age and kind of response are shown in parentheses.

	Grade 2	5	8	
Tell Authority	15 (15.2)	13 (20.7)	27 (19.1)	55
Return Agression	8 (7.5)	15 (10.2)	4 (9.4)	27
Other	5 (5.3)	10 (7.1)	4 (6.6)	19
	28	38	35	101

no relationship means that the ratio of 2nd, 5th and 8th graders giving each of the three kinds of reasons will be the same. The discrepancy between observed and expected frequencies is very unlikely if there were no relationship in some hypothetical larger population as evidenced by the probability figure between 1/2 of one percent (.005) and 1% (.010). The no relation hypothesis is thus rejected and the results are statistically significant.

Page 65, footnote 3. The probabilities computed from a chi square test are only approximate. When the expected frequencies are not too small (say, all more than 5) the approximation is extremely good. In goodnote 3 the investigator is saying that when the responses were spread over more categories, the theoretical (i.e., expected) frequencies for these categories were too small to permit accurate estimates of the desired probabilities.

CRITIQUE

DURKIN, DOLORES. "Children's Concepts of Justice: A Comparison with the Piaget Data," Child Development, 1959, 30, 59-67.

Educational Significance

Question

1 The educational significance of a study dealing with children's concepts of justice should be clearly evident. Elementary school years are usually seen as a time when children learn to cope with agression by learning standards of fairness to apply to interpersonal conflicts. Elementary school teachers are expected to be able to understand these conflicts and to aid pupils in developing appropriate standards of conduct. A commonly-held assumption in teacher education as well as developmental psychology is that studies of child development help determine teachers' expectations of children of different ages and abilities. It could therefore be easily assumed that studies such as Durkin's would have educational significance.

2 One student reader asked, "How can we focus on ways to teach children until we first focus on children?" This question assumes that descriptive studies of what is the case sets limits on what teachers ought to do. Thus, if descriptive studies show that, say, eight-year-olds have not yet developed certain moral concepts of, say, autonomy, then teachers should not try to teach them these new ideas. In one sense, of course, a child must crawl before he can walk. Some things do come before other things in the development of a child. The assumption of the stages of development underlies much of Piaget's work. But information about what children "naturally" learn in the course of their development does not overlap completely with what they might learn under conditions of schooling. Teachers intervene in "natural" development. Thus, we see that the significance of the Durkin study (and others like it) for teachers is considerably less than what we might at first suppose. The Durkin study does not give evidence or advice to teachers about what they might (positively) do with children in teaching them about proper responses to actual or threatened physical agression.

3 After reading paragraph two, one student reader complained that it is wrong to criticize the author for failing to study the teaching of moral concepts. After all, "Isn't it unfair to criticize a work for something the author did not intend?" It is unfair, or rather, inappropriate, to confuse a criticism of the author's intentions with a criticism of other points. Clearly any work has an audience beyond the audience for which the author intended the work. That the author may not have intended the work for teachers does not mean that the work cannot be criticized from the point of view of teachers and teaching.

4 Another student reader defended the significance of the study another way. She wrote: "Piaget has proposed a theory in child development which is quite well known. This study has educational significance because Piaget has earned the attention of educators. Any attempt to expand upon or reconfirm his findings is important."

5 We agree in part. We commend the attempt to rework ideas in one culture (American) which have earned recognition in another culture. Because of contextual considerations, studies done in one place need to be redone in another place if they are to be utilized there with confidence. Nevertheless, just because a study related to Piaget's work does not automatically confer significance on it. Everything depends upon what the study asserts about the significant phenomena of interest.

6 A reader who thought the study lacked educational significance wrote as follows: "This study reports the obvious, namely that older children grasp conceptual complexities younger ones don't. Educators do not need this point demonstrated." There are some not-so-obvious things to say about the obvious.

7 The obvious is usually the conventional, when it comes to educational matters, and the conventional usually has aspects of both right and wrong.

8 Secondly, what seems obvious at the end of an inquiry might not have been so obvious at the beginning. Presumably any inquiry is an attempt to find something out that is at least somewhat in doubt. If we really knew for certain in the beginning of an inquiry what we wanted to know by the inquiry, then it is not likely that we would undertake the study.

9 Consider this table:

BEGINNING OF INQUIRY*	END OF INQUIRY	RANK IN SIGNIFICANCE
Conventional Wisdom	Conventional Wisdom Reaffirmed	4th
Conventional Wisdom	Surprising (new) Results	1st
Puzzling Phenomena	Conventional Wisdom Reaffirmed	3rd
Puzzling Phenomena	Surprising (new) Results	2nd

*Research does not necessarily have to begin at either of these two starting points. It can begin in theory, for example. This table reflects only one way to look at the question about research into the obvious.

10 Our significance rankings are arguable, of course. Nevertheless, to take the first case, simply to reaffirm what everyone already knows is perhaps of only mild interest to either practitioners or researchers. To obtain the highest ranking of the four possibilities one must begin with the conventional and hope to find out something which is surprising to both practitioners and researchers, The third case ranks fair, in our judgment, because it begins with something puzzling and reaffirms a portion of the ambiguous conventional wisdom; we now know which horse to back in the ordinary races of the day. The fourth case ranks high because we find out something we did not know about something which had been puzzling us.

11 Thus, although we agree with the reader mentioned in paragraph six that the findings are what one would expect and therefore the study has limited educational significance, at the same time we support occasional "high risk" studies because from such investigations high levels of educational significance can result.

12. The educational significance of the research problem is discussed in paragraphs 1-5 and, in paragraphs 6-11. Consideration is given to the significance of the research findings. One can also assess the significance of research as conducted and ask whether the actual study contributes significantly to the solution of the research problem, as defined by the investigator.

13 Metaphorically we may say that any inquiry is only a beam of light on a vast, clouded and perhaps dark area of interest. No beam of light will illuminate the whole area; thus, any single study has to be less than comprehensive. Granted the necessity to limit any study, special care must be taken to see that the actual study is not too small (only a pinhole rather than a beam of light). For reasons discussed above and subsequently, we think the Durkin article is more like a pinhole than a beam of light in its educational significance.

Question 2: Purpose 1

Adequacy of the Design.

14 Subjects. Since all the subject children came from the same community and school, the investigator may safely claim that the differences in responses observed are not the result of broad environmental differences. Since the environmental factors have been equated, the observed differences most likely reflect age differences although some underlying factor (such as intelligence) cannot be ruled out as at least partially responsible.

15 Whatever its advantages, the use of such a small and homogeneous sample makes generalization to other U. S. school children hazardous. Age differences in responses for this single school may not be similar to those one would find in other schools or other communities in the United States. For example, we do not know the "rules" of the school and teachers regarding fighting and other forms of agression, rules which might have a disproportionate influence on the children's responses.

16. The investigator wishes also to test for cultural differences by comparing the responses of her sample of children with those made by the children used in the study reported by Piaget. The latter were described as coming "from the poorer parts of Geneva (Switzerland)." At least five characteristics of the sample as described prevent an adequate comparison for these cultural differences: a) the two samples (from Piaget, from present study) were selected roughly 30 years apart and thus time as well as cultural determinants are involved; b) the sample used in the present study is a homogeneous one and may not be representative of American culture: c) the sample used in the present study is from a rural community whereas the earlier study employed urban (Geneva) children--thus the differences may not be strictly cultural; d) the oldest children used in the sample reported by Piaget were 12, whereas in the present study the oldest group had a mean age of almost 14--thus differences in response could reflect age differences. e) although residents are placed as "poor," "average," or "rich," we do not know how, for example, the resources of the family of an "average" child of the present study compared with the resources of the family of a child used in the study reported by Piaget.

17. The fact that the differences in sample may be multi-dimensional and not purely cultural, is probably not as serious a weakness as the above discussion would lead you to believe. One aspect of Piaget's moral theory is that the changing responses reflect a basically genetic development. Thus, if the investigator notes substantial discrepancies in the responses of the two samples of children, the genetic dev elopment position is weakened--a finding of some theoretical importance.

18. Interview Procedures. Differences between the two samples are to be expected because a purpose of the study was to replicate the findings of Piaget with a different type of child. But differences between the studies in the interview procedures clearly make any valid comparison between the two studies unlikely and represent a serious weakness of the study.

19. One modification the present investigator made in the interview procedure is in the initial question asked. The investigator asked the child what should Vann do. In the Rambert study reported by Piaget, the child himself was asked what he does do. Not only is the question more detached than the other because it involves two fictional children, but should do is substituted for does do. These modifications might make a considerable difference in the responses. (Once the investigator decided upon fictional names, the use of rarely used names like Vann and Bennett was a good strategy since it reduced the likelihood that the responses would systematically be affected by the characteristics of a real Vann or Bennett known to the children. We are assuming the names Vann and Bennett are rare in this context.)

20. A second modification in the procedure is the avoidance of the "clinical method." Recall in the special notes that, in the earlier study, much inquiry took place after the child's initial response. Indeed, it was the responses to these later, more probing questions which determined

the response category for the child. The present investigator restricted herself to a single question, except in one case when an eye-for-an-eye response did prompt her to provide a follow-up question.

21. Anyone who tries to replicate a Piagetian experiment faces a real problem because Piaget's experimenters are trained in the "clinical method" and routinely conduct a short inquiry into the meaning of a child's answers to any standardized question. In many cases not even the initial question is standardized. A good case in point is the study being replicated by the investigator.

22. As we noted, the investigator has modified the procedures, no doubt in the interest of objective reporting and to avoid the "clinical method." This modification, however, was made at the price of losing any valid basis for comparison with the earlier study. It is unfortunate that the investigator does not mention any of these problems, but instead conveys the impression that she merely replicated the earlier experiment using a different cultural group.

23. Somewhat parenthetically, we might add that had the investigator not wanted merely to replicate the earlier work, but rather had wanted to study the development of the concept of justice in the best way possible, then other procedures of gathering data should obviously have been considered. For example, what a child says he would do when speaking to an adult may be quite different from his actual behavior. Some check for such a discrepancy might be included. Further, the single question focuses on too narrow a range of the factors involved in making a moral judgment and concerns itself with only a single aspect of justice. Other factors might be investigated. However, because the purpose of the present study was replication, we do not fault the investigator for not including such extensions in her data gathering procedures.

Appropriateness of the Analyses.

24. Categorization of the Responses. Even a cursory glance at the special notes accompanying the article will reveal that the investigator used a categorization scheme different from that employed in the earlier study. This permits at best only a rough comparison of the data from the two studies.

25. Once the decision was made to drop Rambert's classification system (presumably for a system believed to be more objective), it seems to us that a finer and more productive set of categories could have been established. For example, we believe that the following categories would have been preferable: a) tell authority, b) retaliate, c) conflict, with resolution in direction of telling authority, and d) conflict, with resolution in direction of retaliation.

26. Finally, we note that the interviews were tape-recorded. We wonder why individuals ignorant of the respondent's age were not used to categorize the responses. Such a procedure would guard against at least one form of experimenter bias.

27. Statistical Tests. Although admittedly a minor point, we fail to see the purpose of the statistical test mentioned at the bottom of page 62 and interpreted in some detail in the special notes. The question being tested is whether, for a given grade, the responses tend to pile up more in one or two categories than would be expected if the probability of assignment to the three categories were equal. Failure to find significance (as was the case for 2 of the 3 grades) could be interpreted to mean that there is no model response for a grade level and that it would be misleading to say that such and such a grade level child is characterized as being in one or more categories. But such an interpretation was not given by the investigator.

28. A more serious problem with the statistical analysis relates to the chi square analysis. The chi square test measures only the discrepancy between actual and expected frequencies and does not take into account the order or pattern in which the discrepancies occur. Thus, for each hypothetical layout shown below, the chi square test will give identical results even though the direction of the effects is different. As before, expected frequencies are in parentheses.

Layout (a)

2	5	8
4 (6)	6 (6)	8 (6)

Layout (b)

2	5	8
8 (6)	6 (6)	4 (6)

Layout (c)

2	5	8
6 (6)	4 (6)	8 (6)

Layout (d)

2	5	8
8 (6)	4 (6)	6 (6)

Since the investigator is hypothesizing a specific direction or trend (linear with age) and, after seeing the data a curvilinear trend, statistical tests which would be more powerful in detecting such trends should have been used. In other words, the statistical analyses employed should match the research question.

29. Finally, we wonder why the investigator: a) did not analyze her results by sex in view of the sex differences evident in Rambert's data: and b) went to the trouble to place each child into one of three categories of economic status when no analysis was conducted by level of economic status. The analyses performed by the investigator are not incorrect; they are merely incomplete.

Validity of the Interpretations and Conclusions.

30. Interpretation of Rambert's Data. In speaking of the earlier study, the investigator writes: "They generally proposed two quite different solutions. Younger subjects favored reporting to an authority person; older subjects, a return of the aggression." (p. 59). Piaget never reported a specific

27. Statistical Tests. Although admittedly a minor point, we fail to see the purpose of the statistical test mentioned at the bottom of page 62 and interpreted in some detail in the special notes. The question being tested is whether, for a given grade, the responses tend to pile up more in one or two categories than would be expected if the probability of assignment to the three categories were equal. Failure to find significance (as was the case for 2 of the 3 grades) could be interpreted to mean that there is no model response for a grade level and that it would be misleading to say that such and such a grade level child is characterized as being in one or more categories. But such an interpretation was not given by the investigator.

28. A more serious problem with the statistical analysis relates to the chi square analysis. The chi square test measures only the discrepancy between actual and expected frequencies and does not take into account the order or pattern in which the discrepancies occur. Thus, for each hypothetical layout shown below, the chi square test will give identical results even though the direction of the effects is different. As before, expected frequencies are in parentheses.

Layout (a)

2	5	8
4 (6)	6 (6)	8 (6)

Layout (b)

2	5	8
8 (6)	6 (6)	4 (6)

Layout (c)

2	5	8
6 (6)	4 (6)	8 (6)

Layout (d)

2	5	8
8 (6)	4 (6)	6 (6)

Since the investigator is hypothesizing a specific direction or trend (linear with age) and, after seeing the data a curvilinear trend, statistical tests which would be more powerful in detecting such trends should have been used. In other words, the statistical analyses employed should match the research question.

29. Finally, we wonder why the investigator: a) did not analyze her results by sex in view of the sex differences evident in Rambert's data; and b) went to the trouble to place each child into one of three categories of economic status when no analysis was conducted by level of economic status. The analyses performed by the investigator are not incorrect; they are merely incomplete.

Validity of the Interpretations and Conclusions.

30. Interpretation of Rambert's Data. In speaking of the earlier study, the investigator writes: "They generally proposed two quite different solutions. Younger subjects favored reporting to an authority person; older subjects, a return of the aggression." (p. 59). Piaget never reported a specific

percentage of subjects who replied that they would tell an authority, although in other contexts he does describe the early period of the child's moral development as being marked by a submissive attitude to authority. So, although it is not unreasonable for the investigator to suggest the trend above, it is not an accurate rendering of Piaget's report of Rambert's experiment itself.

31. On page 64 the investigator quotes Piaget as saying that, "children maintain with a conviction that grows with their years that it is strictly fair to give back the blows one has received." This statement, which is repeated in a different form in conclusion 1 (p. 66), is not an accurate summary of the Rambert data and the investigator should have been more critical of Piaget's interpretation. For girls, this option of strictly equal retaliation declines with age in favor of under-retaliation (see the table in the special notes section). For boys, there is an increase in equal retaliation responses but the trend is not at all clear.

32. Interpretation of the Data from the Present Study. One objection to the data analyses in the present study has already been mentioned: the chi square test is inadequate to test the significance of the linear hypothesis stemming from Rambert's data and the curvilinear hypothesis suggested by the findings of the present study.

33. One interesting finding reported by the investigator is that although both 2nd graders and 8th graders favor telling an authority, their responses are not identical--8th graders think more about it. It is too bad that the investigator describes only a single interview to illustrate this difference. A different category system, such as the one we proposed earlier, would have made possible a more penetrating and precise set of conclusions. Except as noted above, conclusions 1 through 3, dealing with the first purpose indicated for the study, seem to us to follow from the data reported.

Question 3: Purpose 2

34. The investigator has given no reason for expecting intelligence to be related to the development of a concept of justice. Since justice is a social concept, we suspect that a case could be made for expecting many other correlates of moral judgment development more worthy of investigation.

35. Once the decision was made to relate intelligence to moral judgment development, a mistake was made, we believe, in using I. Q. rather than mental age as the measure of intelligence. I. Q. is a measure of rate at which the child is able to grow in intelligence--about half the second graders have a higher I. Q. score than the average eighth grader. But in terms of sheer amount of intelligence, approximated by the measure of mental age, very few, if any, of the second graders would surpass the average eighth grader. It isn't the "brightness" per se that is believed to be related to degree of development, but rather the amount that this bright child has learned.

36. When analyses are conducted within a single grade (thus using children of approximately the same age), then I. Q. and mental ability will be very highly related and the choice of variable will make little difference. We suspect, however, that when using several age groups simultaneously to test the relationship of intelligence to the development of a justice concept, had mental age been used, a different result would have been found and the "conflicting" results referred to in conclusion 4 would be less likely. (Technical note: because of the markedly different standard deviations in I. Q. scores among the three grades--see paragraph 6, p. 62--the calculation of mental age scores directly as the product of I. Q. times chronological age divided by 100 would not be recommended. Preferred are standard scores computed at each grade level.)

Appendix VII

Florence R. Harris, Margaret K. Johnston,
C. Susan Kelley, and Montrose M. Wolf

Effects of Positive Social Reinforcement on Regressed Crawling of a Nursery School Child

Journal of Educational Psychology, February 1964.

Effects of Positive Social Reinforcement on
Regressed Crawling of A Nursery School Child

Florence R. Harris, Margaret K. Johnston,
C.Susan Kelley, and Montrose M.Wolf.

Journal of Educational Psychology, February 1964.

SPECIAL NOTES

The chronology of this study can be conveniently divided into several periods. During the first two weeks of her nursery school experience, Dee showed strong withdrawal behavior and was off her feet most of the time. During the third and fourth weeks, the teacher reinforced on-feet behavior with the result that, "Dee's behavior was indistinguishable from that of the rest of the children." Next came a crucial 2-day period in which Dee was given special attention during her off-feet behavior (the reversed reinforcement contingencies). The results of this change in reinforcement pattern are shown in Curves 1 and 2 in Figure 1. Thereafter, the regular reinforcement of on-feet behavior was resumed. The results for the first two days after the start of this second reversal of procedures are shown in Curves 3 and 4 on page 119.

Figure 1 is a bit difficult to interpret. The length of the line in the horizontal direction indicates the length of time Dee was being systematically observed. Thus, the longest observation period was during the second day in which attention procedures were reversed; the shortest for the two days immediately following. The steepness of the curve to the horizontal axis indicates the degree to which Dee was off her feet. Thus, Dee was on her feet the greatest length of time for the last day shown because Curve 4 is not very steep. On the critical first day in which reverse procedures were followed (Curve 1) Dee was off her feet most of the time except toward the end of the observation period when, as shown by the bend in the curve to the horizontal position, she was on her feet. Do not be misled by the fact that Curve 4 is "in the air." The positioning of the curves was arbitrary. We suspect that Curve 4 was placed along side Curve 3 in order to save space, or to remind the reader that the last two curves refer to consecutive days under the same reinforcement condition.

Effects of Positive Social Reinforcement on
Regressed Crawling of A Nursery School Child

Florence R. Harris, Margaret K.Johnston,
C.Susan Kelley, and Montrose M. Wolf.

Journal of Educational Psychology, February 1964.

Question 1.

Puzzling behaviors occurred during the reversed reinforcement period where off-feet behavior is reinforced. Describe these unexpected phenomena.

Answer 1.

Dee became more socially adjusted during the period when attention was given to off-feet behavior. It was not expected that Dee's return to her off-feet behavior would be accompanied by greater social adjustment. She began, "...for the first time to accept, even seek, attention from the other teacher." * She also exchanged a few words with the other children, something entirely new for Dee. "The positive effects of reversing reinforcement contingencies seemed to outweigh by far the momentary negative results." (p. 121)

We were also puzzled by another event which was not commented upon by the authors. We would not have expected Dee to return to her predominantly off-feet behavior as quickly as she did on the first day that the reverse reinforcement procedures were instituted. As indicated by the steepness of Curve 1 at its lower left portion, on that first day Dee appears to have been off her feet from the moment she entered nursery school.

Question 2.

The authors conclude that the increased ratio of on-feet to off-feet behavior in Dee

* P. 120.

Question 1.

Puzzling behaviors occurred during the reversed reinforcement period where off-feet behavior is reinforced. Describe these unexpected phenomena.

Answer 1.

Question 2.

The authors conclude that the increased ration of on-feet to off-feet behavior in Dee was caused by the teacher's

Question 2 cont'd.

was caused by the teacher's positive social reinforcement of the on-feet behavior. Other explanations are possible. Dee's increased on-feet behavior might be explained by at least some of the following: a) the reinforcement of walking itself; b) increased familiarity with the nursery setting; c) the expanded range of rewarding objects (toys and people) made possible by walking; d) possible physical factors (such as illness, fatigue, physiological maturation). Decide which explanation, if any, you think is correct. Give reasons for your choice.

Answer 2:

At least some of the factors mentioned in Question 2 would be reasonable explanations for the increased on-feet behavior were it not for the fact that the investigators could change the on-feet to off-feet ratios merely by changing the focus of the teacher's reinforcement. These four factors were present during the off-feet reversal time. We are thus led to conclude that the return to high off-feet behavior is most likely due to one factor that was correspondingly changed - the teacher reinforcement procedure. If the teacher's social reinforcement were not a causal factor, removing this reinforcement would not change Dee's on-feet to off-feet behavior ratio.

If you answered that one of the four factors could account for Dee's increased on-feet behavior, you are in a predicament. If any of these factors were responsible for the increased on-feet behavior then Dee should have continued her improvement during the reversal

Question 2 cont'd.

positive social reinforcement of the on-feet behavior. Other explanations are possible. Dee's increased on-feet behavior might be explained by at least some of the following: a) the reinforcement of walking itself; b) increased familiarity with the nursery school setting; c) the expanded range of rewarding objects (toys and people) made possible by walking; d) possible physical factors (such as illness, fatigue, physiological maturation). Decide which explanation, if any, you think is correct. Give reasons for your choice.

Answer 2:

Answer 2 cont'd.

procedure because these factors were all present at that time. The fact that Dee reverted to her off-feet behavior implies that any effects of these factors were overshadowed by the teacher's positive social reinforcement.

Whether researcher or critic, be alert in any research for other explanations and assess their plausibility. The investigator's use of a manipulated variable design was effective in dealing with what otherwise would have been reasonable alternative explanations.

Question 3.

This study is a cause-and-effect study: attention to on-feet behavior (X) causes a child to change her behavior (Y). It is commonly thought that phenomena are explained when causes can be correctly identified. How can we best explain Dee's behavior? Three forms of explanation are as follows:

A. The covering law form. A single instance is explained when it is subsumed under a general law which "covers" the particular case. For example, the specific instance in which a spherical object can pass through an iron ring only when the ring is heated is explained by the general law that heat causes a metal object to expand.

B. The manipulated variable form. Event X is said to be a cause of Y because when the experimenter permits X to be present, he gets Y, and when he removes X, he fails to get Y.

C. The coherent pattern form. Event X is said to relate to event Y when the many descriptive elements in these events are shown to "fit" together to form a pattern of relations. In such a case, multiple causes, some occurring together and some

Question 3.

This study is a cause-and-effect study: attention to on-feet behavior (X) causes a child to change her behavior (Y). It is commonly thought that phenomena are explained when causes can be correctly identified. How can we best explain Dee's behavior? Three forms of explanation are as follows:

A. The covering law form. A single instance is explained when it is subsumed under a general law which "covers" the particular case. For example, the specific instance in which a spherical object can pass through an iron ring only when the ring is heated is explained by the general law that heat causes a metal object to expand.

B. The manipulated variable form. Event X is said to be a cause of Y because when the experimenter permits X to be present, he gets Y, and when he removes X, he fails to get Y.

occurring as a sequence of events, are described. Thus, to explain the causal relations between X and Y it is necessary to give a full account of the elements involved. This is the historian's, or case study, form of explanation.

One reason we found this study to be particularly interesting is that the investigators provide a rich assortment of evidence to support the claim that attention to on-feet behavior caused Dee to change her behavior. It is possible to (and we would like you to) explain Dee's behavior using each of the three forms of explanation described above. Specifically, in regard to each of these three forms of explanation:

1. cite material from the article itself which could be used to explain the change in Dee's behavior; and

2. give reasons for being critical of each of these explanations.

Thus, for example, your answer to Question 3A 1 would need to identify a general law and show how one could claim it "covers" this particular case. In 3A 2 your response will be a criticism of the explanation presented in 3A 1.

Note: A complete answer to this question will have six sections: 3A 1, 3A 2, 3B 1, 3B 2, 3C 1, 3C 2. Further, we are not asking you to pick one form of explanation as "correct." Critically discuss how each applies in this study.

Answer 3.

A. Covering law form.

1) Evidence:

One way to express the covering general law is: behavior is strengthened when it is followed by a reward (reinforcement);

C. The coherent pattern form. Event X is said to relate to event Y when the many descriptive elements in these events are shown to "fit" together to form a pattern of relations. In such a case, multiple causes, some occurring together and some occurring as a sequence of events, are described. Thus, to explain the causal relations between X and Y it is necessary to give a full account of the elements involved. This is the historian's, or case study, form of explanation.

One reason we found this study to be particularly interesting is that the investigators provide a rich assortment of evidence to support the claim that attention to on-feet behavior caused Dee to change her behavior. It is possible to (and we would like you to) explain Dee's behavior using each of the three forms of explanation described above. Specifically, in regard to each of these three forms of explanation:

1. cite material from the article itself which could be used to explain the change in Dee's behavior; and

2. give reasons for being critical of each of these explanations.

Thus, for example, your answer to Question 3A 1 would need to identify a general law and show how one could claim it "covers" this particular case. In 3A 2 your response will be a criticism of the explanation presented in 3A 1.

Note: A complete answer to this question will have six sections: 3A 1, 3A 2, 3B 1, 3B 2, 3C 1, 3C 2. Further, we are not asking you to pick one form of explanation as "correct." Critically discuss how each applies in this study.

Answer 3 cont'd.

Answer 3:

conversely, behavior is weakened or eliminated when it is not rewarded. The on-feet behavior was strengthened because it was followed by a reinforcement (adult attention). Animal trainers, teachers, and parents have used something like behavior modification for centuries by providing food, gold stars, or treats when their charges performed desired behaviors. We generally conclude that the cause of behavior change is reward. Dee's change in behavior (from off-feet to on-feet) is the special case subsumed under the law of reinforcement.

2) Criticism:

The specific instance (the Dee case) in this study does not fit the general law completely. Recall in connection with Question 1 that some of the positive behavior (e.g. greater social adjustment, playing near other children, etc.) was NOT weakened when Dee's on-feet behavior was no longer rewarded. It is perfectly acceptable to claim, as the investigators do, "...that there were sources of social reinforcement not in coordination with those controlled in the experiment." (p.121) It is important, however, to be able to identify these reinforcers independently of whether or not they have an effect. If only those actions which change behavior are called reinforcers, then the law that reinforcement changes behavior must be true by definition; it is untestable.

Because it is convenient to do so and not because it is a direct answer to question 3A 2, we make the following two observations about laws and the covering law form of explanation. First, more than one law can account for the same observation. The covering law form thus permits multiple explanations of the same observations. Second, when causal explanations take this covering law form we generally do not ask for further explanation since these events are common and familiar in the experience of most people. But we may find a law-like relation between events X and Y

Answer 3 cont'd.

and still feel that the relation has not been adequately explained. For example, we may see that heat causes a metal object to expand, but we may still feel that we do not have a completely satisfactory explanation of expansion of metals. Similarly, we may feel that "reinforcement" is not a satisfactory explanation.

B. The manipulated variable form.

1) Evidence:

Clearly, the investigators were manipulating an event and studying the resulting effects. The investigators purposefully increased teacher attention to Dee's on-feet behavior and gave no attention to off-feet behavior (X present), then purposefully reversed attention procedures (X withdrawn), and finally reinstated the original attention procedure (X again present). Increased on-feet behavior (Y) was evidenced when attention was directed toward it (X present) and when such attention was reversed, (X withdrawn), the investigators failed to get (Y).

2) Criticism:

The manipulated variable form of explanation is usually attacked on grounds that other factors vary as X is manipulated, or that event X is too broadly stated and that the real cause of Y is only some component of event X

Some students have correctly observed that reinforcement was not withdrawn but rather the particular behavior reinforced was varied. Since Dee received teacher attention all the time, it is not too surprising that some of Dee's changes (e.g. greater social adjustment) did not deteriorate during the 2-day reverse reinforcement period. Nevertheless, the conclusion that the specific focus of the reinforcement caused the change in Dee's on-feet to off-feet ratio, is not weakened by the reasoning above and this conclusion would seem inescapable if it were not for some added reservations spelled out below.

Answer 3 cont'd.

Answer 3 cont'd.

We are struck by the puzzling fact that Dee resumed her old off-feet habits immediately upon beginning the first day of the 2-day period in which attention procedures were reversed. If these attention procedures were such powerful conditioners, then why, on the day immediately following the final reversal of procedures was there virtually no lessening of off-feet behavior (see Curve 3)?

We are left wondering how much off-feet behavior would have occurred during the critical 2-day period had no change in procedure been instituted. The data in Curves 1 and 2 would have been more convincing had similar data been shown for the other children. We are told, for example, that Dee's playmates in the doll corner were also off their feet. The proportion of time a child will spend on his feet depends to some extent upon the type of activity in which he is engaged. We speculate that during these particular two days Dee perhaps chose to spend more time indoors involved in activities that naturally lent themselves to off-feet behavior.

Answer 3 cont'd.

C. The coherent pattern form.

1) Evidence:

A case study admits complex events taking place over a significant period of time with many variables. The investigators report many of these descriptive details: the family background; Dee's entry behavior; the puzzling fact that she regressed to crawling (strong withdrawal behavior to usual friendly, warm teacher approaches); mother's reports; the development of the study through the various reinforcement procedures; social adjustment; and post checks made at irregular intervals for a year subsequent to the study. (Teachers agreed that Dee's improved behavior was stable.) The investigators attempt to show how all these facts fit together in a sensible way.

Answer 3 cont'd.

2) Criticism:

The main criticism is that not enough of Dee's life prior to entry into nursery school is given; nor do we know enough about Dee's life outside nursery school (21 hours of each day). Specifically, we have no information about why Dee might have started to regress to crawling behavior. We could speculate that her younger brothers (aged 8 months and 18 months) were both still crawling and that while they were rewarded (attention given) Dee was not. Finally, we need more description of what other adult and child behavior might have been reinforcing to Dee.

We believe that events are explained when a sufficiently rich description of these events leaves us without further significant questions to ask. The incompleteness of this report leads us to describe it as a demonstration study rather than a case study. It is a demonstration of the application of the principle of reinforcement rather than an explanation of how and why these principles work.

Question 4.

One person wrote that, "...the study would have been better if: a) reliability checks had been made on all the recordings; b) recordings had been available for more time; and c) there had been better documentation of what was happening at the various times the child was on-feet and off-feet." Do you agree? If not, why not? If yes, do you think such improved record keeping might have changed the authors' conclusions? In what way(s)?

Answer 3 cont'd.

Question 4:

One person wrote that, "...the study would have been better if: a) reliability checks had been made on all the recordings; b) recordings had been available for more time; and c) there had been better documentation of what was happening at the various times the child was on-feet and off-feet." Do you agree? If not, why not? If yes, do you think such improved record keeping might have changed the authors' conclusions? In what way(s)?

Answer 4.

We agree. The procedures used in this study were much more casual than those we would generally expect to find in educational and psychological research. Dee was observed systematically only a portion of the time. The change in percent of time off-feet from the 2-day reverse reinforcement period to the second reversal period might be accounted for by normal, expected variation. Without a longer, more detailed accounting of percent off-feet statistics, we have little basis for assessing the fluctuation which did occur.

We have no indication of how reliably the observers were able to record Dee's behavior. Particularly welcome would be information on inter-rater reliability; that is the extent to which the ratings of several judges observing the same occurrences agree. Further, the investigators rely on teachers' judgments and impressions which may be subject to bias due to their own expectancies.

If we had more precise and complete data, the speculations we made in our answer to Question 3B 2 would not have been necessary. Although we doubt that such data would change the main conclusion, there nevertheless remains the possibility that such alternative explanations (chance fluctuation, nature of the activities Dee engaged in, and so forth) would be supported.

Question 5.

Cite four strengths (i.e. desirable features) of this investigation. Attempt to identify distinct types of strengths.

Answer 5.

There are many positive things to be said about this investigation. A few strengths are listed below, but this list should not be considered complete.

Answer 4:

Question 5.

Cite four strengths (i.e. desirable features) of this investigation. Attempt fo identify distinct types of strengths.

Answer 5:

Answer 5 cont'd.

1. An experimental situation was manipulated by the investigators. When the independent variable is manipulated by the experimenter, we have a very strong technique for investigating the existence of causal relations. The fact that the investigators provided, then removed, and then restored attention to the on-feet behavior of Dee provides strong evidence that this attention was a cause of whatever effects varied systematically with changes in attention. The evidence is much stronger than, for example, if the investigators merely increased attention to the on-feet behavior and observed the results.

2. The investigators displayed a concern for the possible negative consequences on Dee of their study. The investigators were prepared to terminate the reversal condition if Dee showed, "...any evidence of detrimental effects, such as loss of speech, crying, or other emotional behavior." Researchers do not have unlimited rights to manipulate their subjects. The rationale that society will benefit from such findings is not sufficient to harm, psychologically or physically, the particular subjects used in search for greater knowledge. The researcher has an obligation to protect the individual.

3. We commend the investigators for seizing an interesting opportunity (the discovery of Dee) for special study. Significant research is often conducted when the research is triggered by a puzzling observation or fortuitous event. Had the investigators first planned a careful system of observation, for example, and then sought to find a Dee, a more smoothly executed study might have been the result, but only if a Dee could then be found. It is better to do what you can with an interesting situation that presents itself than to let it pass unstudied.

Answer 5 cont'd.

Answer 5 cont'd.

4. The study has direct relevance for educational practices. The major variable is one that can be manipulated by teachers in classrooms or other situations. That is, it is in the power of teachers to reinforce desired behaviors by such social rewards, although we admit that, in some senses, the situation described in this investigation was well suited for this purpose. The fact the study was conducted in a schooling situation using techniques easily learned facilitates its adoption by others.

5. The study did provide evidence about the effect of positive reinforcement. Thus, in the conduct of the study, supportable knowledge claims were made.

6. One person commented that a strength of the study was:

> To make nursery school teachers and student observers more sensitive about the effectiveness of their own behavior in shaping children's responses. One child changed; many teachers were. One can imagine the hours of meeting time that were devoted to planning and discussing this study and its implication; no doubt this was excellent in-service education of the teachers and their apprentices.

We note the values of research are not limited to the supportable knowledge claims. Inquiry is a form of learning and is often as valuable as a process itself as for the direct results it supports.

7. The investigators considered several dependent variables (e.g. social adjustment) and not just the single variable of on-feet behavior. They were concerned both with the long range results of the experiment (as evidenced by their follow-up checks) and with the unintended as well as the anticipated outcomes.

Answer 5 cont'd.

8. One child was benefited directly.

9. The problem was well stated, the article was logically organized and the nature of the reinforcement was explicitly defined.

Answer 5 cont'd.

Appendix VIII

David Elkind, Joann Deblinger, and David Adler

Motivation and Creativity: The Context Effect

American Educational Research Journal, Vol. 7, No. 3, May 1970

Motivation and Creativity: The Context Effect

David Elkind, Joann Deblinger, and David Adler

American Educational Research Journal, Vol. 7, No. 3, May 1970

SPECIAL NOTES

p. 352, p. 356: Putative creativity measures means generally considered to be creativity measures.

p. 354: Replace the first sentence under the subheading "Design" with the following:

> The experiment lent itself to an analysis of variance design with motivating-condition as the within subjects variable and order-of-motivating condition and students within order-of-motivating condition as the between subjects variables.

The above sentence, which replaces the inaccurate statement in the published report, indicates that the statistical analysis of variance involved three variable 1) motivating condition (interesting task interrupted, uninteresting task interrupted), 2) order-of-motivating-condition (interesting task interrupted first, uninteresting task interrupted first), and 3) students within order-of-motivating condition (16 students who were interrupted first from an interesting task and 16 students who were interrupted first from an uninteresting task). Motivating condition is considered as a "within subjects variable" because the two conditions being compared (interesting task interrupted, uninteresting task interrupted) involve the same 32 students. Variables 2) and 3) are "between subjects variables because the comparison of the two orders and the comparison among the students each involves different subjects.

p. 355: Omit the last two lines above the heading, DISCUSSION. By "Groups Under Order of Motivating Condition" the investigators must mean students within order-of-motivating condition, and a significance test of this variable is not possible given the design used in the study.

p. 356: The creativity-intelligence dichotomy is the separation of creativity and intelligence into distinct traits so that being highly intelligent does not necessarily mean being creative and vice versa.

ORIENTING QUESTION OF APPRAISAL

Any study will probably contain key weaknesses and strengths as well as several more minor ones. By key, we mean those aspects of the study upon which th value of the work rests most heavily, and without which the study would be reduced markedly in worth. In an empirical study such as this one, key areas include: A) quality of reasoning from problem statement data to conclusions and implication B) methods of work (including instrumentation, design and analysis); and C) defens of the problem's significance. Provide a critique of the key aspects of the study which emphasized its key flaws and is organized into the three areas just identifi Do not concern yourself at this stage with key strengths of the study.

CRITIQUE

A. Reasoning from Problem to Data to Conclusions:

Background Considerations.[1]

1. "Motivation and Creativity: The Context Effect," is similar to the vast majority of empirical investigations in education in that the purpose of the research is to discover and to explain relationships between variables. How these variables are defined and described is therefore crucial to the value of any such investigation.

2. The variables of this study are described at varying levels of abstraction. At a level close to the events, the variables are referred to as the kind of task interrupted (i.e., crossing out n's and 6's, or activities indicated by th teacher as interesting) and as the total number of responses and total number of unique responses to several specific questions. At the highest level of abstraction, context and creativity are the variables being related, and motivation is seen as the construct (or abstract "mechanism") which explains the relationship.

3. The import of a scientific study increases greatly when an investigation is concerned with variables at higher levels of abstraction. There are two relate reasons for this point. First, predictions which cover a wider range of observables are possible. Thus, for example, if a relationship is described in terms of creativity, then we can predict the relation to hold for other valid measures of creativity. On the other hand, if a relationship is examined in terms of number of uses of a newspaper, then we have a poorer basis for predicting performance with other kinds of measures. The more specific the terms of the examinatic the more specific and therefore limited will be the valid applications. Secondly the use of constructs helps us to explain the reasons for the relationship. If we want to understand the reasons for a relationship, if we want to know the extent of this relationship, and if we want to know how to allow for this relatio ship in the practice of education, it is important to set the observed relationship into an explanatory system or theory. Some of these ideas are illustrated by Figure 1.

4. Constructs:

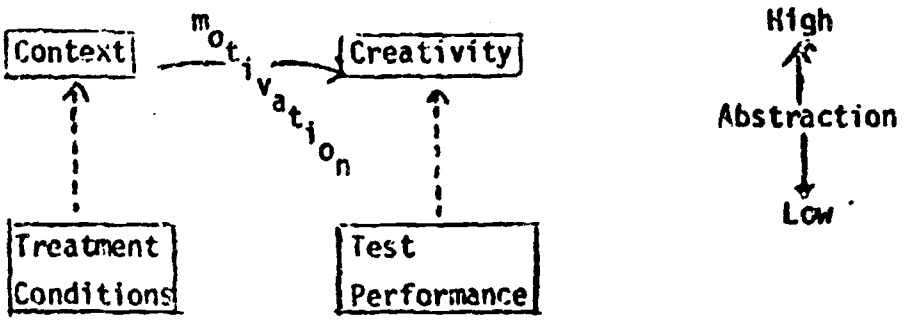

Figure 1. Variables considered in the present study and their professed interrelationship.

---

[1]Not expected to have been stated in a model critique but offered here for pedagogical purposes.

5. As just indicated above, the importance of a study is greatly enhanced when it rises beyond providing the relation between observed variables and yields an inferred relation among constructs. But the validity of this inferred relation depends in turn upon the validity of the observed variables used to measure the constructs and upon the adequacy of the intervening constructs. Consequently, following is our assessment of the context, creativity and motivation constructs as they are involved in the study.

Context:

6. In the broadest sense used in this study, context is seen as, "the ongoing activities interrupted by the test procedures." Other statements in the study lead one to believe that a defining property of the independent variable is the knowledge of the child that he will be returned to the task designated as "interesting" or "uninteresting". But it isn't clear, when we interpret the results, whether differences in scores are to be attributed to some perceived contrast between past activity and present test-taking tasks, or to anticipation of some future experience. This difficulty is illustrated by Figure 2.

7.

Condition	Before Testing		After Testing
Interesting	Interesting Activity	T E S T	Resume pretesting activity- i.e., return to interesting task
Uninteresting	Cross out n's and 6's	T	Resume pretesting activity- i.e., return to crossing out n's and 6's

Fig. 2 Study Design

8. The difference in test performance under the two conditions might be due to the nature of the interrupted before testing task (as the investigators suggest) or to the differential pull of the anticipated post-testing activities. For example, students in the uninteresting condition might have performed better on the tests not because they were happy to get out of an unpleasant task but because they persisted on the test to delay their return to the unpleasant task. The study design does not permit one to assess which context (the pre-testing or the post-testing) is the more important.

Motivation:

9. The investigators appear to have been too quick to infer that motivation is the appropriate construct to be used to explain the differential test performance. Other concepts which could explain the results include: need for novelty, desire to return to a pleasant state, drive for optimal stimulation, etc. Still other concepts are suggested by the vast literature dealing with work conditions and production. He did not include the Ziegarnik effect (the ability to recall unfinished tasks more than completed ones) because there is no reason to believe that the "interesting" and "uninteresting" tasks would have a differential "pull" since both were interrupted tasks.

10. It is possible to argue that all the additional concepts being suggested are really what is meant by motivation--that is, that motivation is any drive, desire or need. Such a concept of motivation is so broad and pervasive that it can be used to "explain" just about everything and, consequently, explains nothing. There is no way to distinguish motivating from non-motivating contexts that is independent of test performance.

Creativity:

11. There is no attempt made to show that the tests employed are adequate tests of creativity. In fact, this task is disavowed:

> It is not our intention here to deal with the issue of whether these tests measure 'creativity' or something else. In this regard we tend to side with Cronbach who argues that 'creativity' is too value laden and that names for particular tests should be used to designate the measure in question.

This move on the part of the investigators succeeds in insulating the argument from the objection that creativity is not really measured by the tests, but the price of this success is the triviality of the conclusion. We can only conclude that there is a relation between context and some test scores rather than between context and creativity.

However, in spite of the quotation cited above, the investigators believe themselves to be dealing with creativity. In numerous places in the text, as well as in the title and the abstract, the investigators refer to the

dependent variable as creativity measures. Further, the most plausible reason for lumping together the scores of the three tests when drawing conclusions is that the three tests are measures of the same thing, presumably creativity.

13. The investigators are trying to have it both ways. They want to eat their cake (protect themselves from the objection that creativity is not really measured by the tests) and have it too (use "creativity measures" in the discussion and conclusions).

B. Methods of Work

14. At this point in the review, we shall consider as the variables of interest the kind of task interrupted and the score on selected tests. Even at this low level of abstraction, three aspects of the research procedure hinder our interpretation of the relationship found: namely, the atypical character of the interrupted tasks, especially the uninteresting one; the inadequate description of the testing situation; and the use of a special school (MOIS) as the locus for the research.

Atypical Character of the Interrupted Tasks:

15. An immediate, practical result of the discovery of a context effect would be to alert the educational practitioner and researcher to the need to be concerned with the activity a child is engaged in prior to any testing situation. By using tasks seen as unrepresentative of the school situation (such as crossing out letters and numbers) the investigators reduce the likelihood that their findings will have relevance to other situations.

16. In defense of the investigators, it is sometimes wise to attempt to obtain the relationship desired using extreme conditions which, in this case, would be very boring and very interesting tasks. If the effects are not evident given extreme conditions, the investigator can feel quite safe in concluding the independent variable is not an important determiner of test performance in a more usual situation. If the effects are evident under extreme conditions then future research can be directed toward the assessment of the effect in a variety of more realistic situations.

Inadequate Description of the Testing Situation:

17. There is a gross lack of information given about the conduct of the creativity testing. The reader needs assurance that the testing conditions were identical under the two motivating conditions. Were the tests administered in a group situation (with the two motivating conditions separate) so that if some child were brave enough to get up and leave others might follow? What subtle cues about how long the children could work on the task may have been present? What were the children told about how long they could work on the tests, and were these instructions consistent with instructions usually given for such tests? Did the children think that if they "finished" early they could (or had to) return early to the task which had been interrupted? In short, there are far too many unanswered questions about the administration of the

tests. The entire difference in test scores under the two conditions might be explained by the factors just identified.

Use of a Special School:

18. The importance of the nature of the interrupted task might not have been so marked had the study been conducted in a more typical school setting. The child in a special school may view the interesting and uninteresting activities as quite different in kind, and thus capable of producing marked differences in test performance. In a regular school, the variety of tasks which could be interrupted by testing would likely differ by degree rather than by kind - in fact the testing situation might be seen as an enjoyable diversion regardless of the task interrupted. We do not claim that such speculation on our part is correct; we only remind you that studies conducted in a very special situation may not generalize beyond it.

Note About Analysis:

19. We do take some exception to the way the data were analyzed and reported. We doubt, however, that our interpretation of results would be much changed had a more adequate analysis been performed. A great many readers cited the use of only 32 pupils as a weakness of the study. We were not bothered by this sample size for two reasons. First, since each pupil was tested under two conditions, the effective sample size was greater than 32. Second, large samples are desired to help insure that real differences in conditions will be detected and not attributed to change. However, a larger sample was not needed because the results of the present study were statistically significant even with the "small" sample used. (The specific questions and answers section of these appraisal materials deals, in part, with data analysis and interpretation.)

C. Significance of the Problem

20. One student, in appraising the research problem, argued:

> This study has very little educational significance. The primary reason is that they sought to demonstrate something which is already well accepted by psychologists and educators. It should surprise no one to find that level of motivation has an effect on the performance on a test which is at least partially scored on the basis of number of responses emitted.

The investigators themselves admit that others have shown the effect of motivation on test performance. The present study, however, deals with a motivational context of special interest to educators - namely, what the child was doing before taking the tests. It seems to us very important to know whether scores on tests such as those given in this study can be influenced to such a large extent by something as seemingly innocuous as the nature of the activity preceding the test administration.

21. Thus, we view the research problem to be significant. Because of the many concerns discussed above, however, we feel that the chief value of the study as conducted is only to remind and caution us that some attention to context is required when testing for creativity. The findings, nevertheless, are provocative enough to warrant additional research on the question.

SPECIFIC QUESTIONS*

1. The "...study was suggested by some unexpected findings that we encountered in our evaluation of the innovative educational program...". Is it legitimate to develop research from unexpected findings? Explain why you answered as you did.

1. Yes, it is legitimate to develop research from an unexpected finding. A puzzling observation, an anomaly, or an unusual situation has often been the precursor of significant research. What is unexpected is, of course, a function of what is expected. In this case the educational expectation of increased creativity for children in a school setting stimulating inquiry and free choice was not upheld. The educators expected these children to be more, not less, creative than children in traditional schools.

   Every situation has its idiosyncratic aspects. If one looks at enough features in a given situation, one or more of the observables is apt to look unusual by chance alone. Thus, although a puzzlement in one situation might well stimulate further study in an effort to seek replication or explanantion of the phenomena, the first occurrence of a puzzlement should not be taken too seriously in and of itself. All perplexities are not worthy of serious further investigation.

Student Response

Unexpected findings, "suggest that some important variables had not been considered or that there is some flaw in the experimental design."

Our Reply

A good point. Before rushing out to seek replication of an unexpected finding, the researcher should re-examine all the procedures of the study which resulted in such phenomena in search of "flaws" which may alone account for the puzzling observation.

1. The "...study was suggested by some unexpected findings that we encountered in our evaluation of the innovative educational program...". Is it legitimate to develop research from unexpected findings? Explain why you answered as you did.

*NOTE: The order of these questions is the same as the paragraph sequence of the published paper.

2. The unexpected findings which motivated the present study resulted from a comparison of World of Inquiry School children with children selected "from names on the waiting list for acceptance into WOIS". In this earlier evaluation of the WOIS and its effect on creativity do you approve of matching the experimental group children with children from the waiting list, or do you think the investigators should have selected the control children randomly from the public schools regardless of whether they were on the waiting list? Why?

2. The unexpected findings which motivated the present study resulted from a comparison of World of Inquiry School children with children selected, "from names on the waiting list for acceptance into WOIS". In this earlier evaluation of the WOIS and its effect on creativity do you approve of matching the experimental group children with children from the waiting list, or do you think the investigators should have selected the control children randomly from the public schools regardless of whether they were on the waiting list? Why?

2. When comparisons are to be made, validity can be maximized if the comparison groups are as identical as possible except for the variables to be examined. If the two groups of school children were different before the one group attended the WOIS, then it is difficult to separate these initial differences from the effects the WOIS was responsible for. Of the two choices given in Question 2, taking control children from the waiting lists appears to be more valid. We can infer that such children are more likely to come from a home environment more similar to the actual WOIS children than children whose parents chose to send them to public schools.

Of course, the answer to Question 2 would be different if there were only a few children on the waiting list (and thus obtaining a good match with the WOIS children would be impossible) or if there is some systematic bias in the way in which the children who were made to wait were different from those who were accepted immediately. Thus, before giving a firm answer to Question 2, it would be helpful to know why some children were accepted and some children were made to wait. In the absence of a differential selection policy, we support the investigators' tactic of choosing controls from the waiting list group.

Student Response

The investigators should have chosen the control group randomly from the general population to, "give more credence to the generalizability of the study." Further, "those on the waiting list are a special population, perhaps more 'creative' (or motivated) than the average public school student."

Our Reply

Our question refers to an earlier study which had produced the unexpected findings regarding "creativity". The purpose of that earlier study was to evaluate the WOIS. It was therefore necessary to use children who were as identical to the WOIS children as possible so that the differences between the two groups of students could be attributed to

Our Reply (continued)

the WOIS experiences rather than other factors. To be sure, the students in the WOIS may be atypical and the ratings of effectiveness of the WOIS may not generalize to a more typical student population. But for the rather specific purpose of determining if the WOIS itself had any impact at all, it is necessary that the comparison group be as atypical (in the same ways) as the students in the school. A good available source for a control group was the WOIS waiting list, the source actually used.

Student Response

"It seems to me that the choice of control group depends on the question the researchers wanted to be able to answer. If the question was, Are children in the WOIS more 'creative' than children in public schools?, then the control group should be chosen randomly. However, such a question would not tell us anything about the effect of the WOIS curriculum on encouraging 'creativity' in its students. Children on the waiting list, however, have already been admitted to the school, and therefore ought to be more 'similar', by whatever criteria the WOIS uses for admission, to the WOIS school population than a random sample of public school pupils. Thus, if WOIS pupils performed better on the battery than potential WOIS pupils, one would be in a position to infer that, for children likely to meet WOIS standards, the WOIS curriculum does promote 'creativity' to a greater degree than does the public school curriculum."

Our Reply

We agree.

3. The investigators state (p. 353, paragraph 1): "Inasmuch as each child who participated in the study served as his own control, we made no attempt to control for or to equate individual differences in ability." (a) What does it mean for a child to serve as his own control? (b) Were the researchers justified in not equating individual differences?

3. The investigators state (p. 353, paragraph 1): "Inasmuch as each child who participate in the study served a his own control, we made no attempt to control for or to equate individual differences in ability." (a) What

3. (a) To serve as their own control means that observations to be compared involve the same objects (usually people). In this study, the students served as their own controls since the test scores to be compared were produced by the same children - once after they were interrupted from an interesting task and once after they were interrupted from an uninteresting task.

does it mean for a child to serve as his own control? (b) Were the researchers justified in not equating individual differences?

Student Response

When a child serves as his own control, you have "a repeated measures design."

Our Reply

That is correct. "Repeated measures" occurs most frequently in the statistical literature concerned with the analysis of experimental data obtained when subjects serve as their own control.

Student Response

To serve as his own control means "the behavior and responses of each child were reflected from his own personal experience", or that "children were stratified by age," or thet "children were matched."

Our Reply

These answers are incorrect.

(b) Yes, the investigators were justified in not equating individual differences in ability. They did not wish to match students and restrict the population of children any more than was already the case by virtue of the fact that only WOIS students were involved in the study. Further, it was not necessary to pick carefully the children because of primary interest in the study was the comparison between each pupil's score under motivating condition 1 and his own score under motivating condition 2.

4. Should different (but matched) children have been used in the two motivating conditions rather than exposing the same children to both conditions? Why?

4. Should different (bu matched) children have been used in th two motivating conditions rather than exposing the same children to both conditions? Why?

4. This is an extremely difficult question to answer. Any research design is a compromise. Using children as their own control in a repeated measures design, as in this study, has both advantages and disadvantages over the design in which matched groups of children are employed. It is a trade-off.

In the case of performance measures, individual differences account for most of the variability and treatments (such as the two motivating conditions) often make relatively little difference. Since this is the case, it is important that differences among the individuals in the two treatment groups be as small as possible so that the relatively small treatment (context) effect will not be masked. The great advantage of the design actually used (the subjects as their own control design) is that the differences between the individuals in the two treatment groups has been minimized. Indeed, they are the same people. A proper evaluation of the motivation conditions effects would involve a comparison of the magnitude of the difference in test performance under the two conditions, with the magnitude of the "unaccounted" for differences. When the subjects are their own control, the unaccounted for differences are reduced tremendously, making us more confident of the accuracy of the treatment differences observed.

The drawback to using subjects as their own control is that such a design does not protect against what is called a differential "carry over" effect. To illustrate this effect, assume that instead of two motivating conditions, two drugs, A and B, were used. Further assume that drug A affects performance while drug B does not, and the effect of drug A is carried over to the time that drug B is tested. Any measure of the performance of the group that received drug B second would not be a true indication of the effect of drug B alone, since drug A would still be in effect, and any conclusions would thus be in error. Since differential carry-over effects are unlikely in the study as actually conducted, we would support the repeated measures design actually employed by the investigators.

Student Responses

"Using different children would have produced a tighter control on any testing or practice effect."

"There was a definite possibility of contamination in the design. Being exposed to the first form of the test might well have influenced the nature of the responses to the second form."

"Having to take two equivalent forms of the same it might involve a carry-over so that the child ld remember and become more proficient the second time the test is taken."

"Some type of learning took place during the first testing and perhaps some modifications occur between the first and second testing."

Our Reply

The practice or testing effect is controlled since the order in which the two treatment conditions are given is counterbalanced -- half the children are first removed from an interesting task, half the children are first removed from the uninteresting task. Thus, the above student responses are not completely accurate; they need to specify that any "contamination" or "carry-over" effect would be differential in nature as explained in the second paragraph of our initial answer to this question. Why would this practice effect or learning be greater (or less) going from motivating condition 1 to motivating condition 2 then going from condition 2 to 1?

Student Responses

Use of the same children is preferable to employing matched groups because: (i) "there was no pretest to help make a good match," (ii) "selection bias would take place," (iii) "it is hard to equate groups," and (iv) there are "too many variables to match children."

Our Reply

For matching to be maximally effective, two conditions need to be met. First, using matching variables highly related to the criterion measures (creativity scores in this study). If, as implied in student response (i), such matching variables are not available, then the effectiveness of the matching strategy would be reduced. Second, random assignment to the two treatment conditions be made after matching has taken place. This procedure protects against the selection bias referred to in student response (ii). As long as the above two conditions are met, matching can be highly effective and free from bias. Contrary to that implied in responses (iii) and (iv), the groups need not be equated on numerous variables.

5. The investigators state (p. 353, paragraph 1) that having children from several grades and of different ages allows for greater generality of conclusions than if a more homogeneous group were used. Do you agree? Explain.

5. The investigators state (p. 353, paragraph 1) that having children fro several grades and different ages allo for greater general of conclusions tha if a more homogenec group were used. I you agree? Explai

5. Yes, an investigator can make broader conclusions when he has employed a variety of subject types or has done his research in a variety of research settings. This statement presupposes that the investigator has analyzed his data by these subgroups. In this study, the investigators have performed such analysis for several age groups. We are NOT saying that the same conclusions will necessarily be valid for each of the various groups and research settings, but only that given proper analysis, one will be able to make a more general set of conclusions using a heterogeneous mixture of subjects than if a homogeneous subject pool is used.

Student Responses

"One must have adequate numbers to generalize with confidence."

"Since the sample was so small, it may be difficult to generalize for several grades and different ages."

"WOIS kids are not a normal group."

Our Reply

This point is well taken. Because of the few children in each subcategory, the likelihood is small that the investigator will be able to make statements about the differences by such subgroups with confidence. Further, the special school setting limits generalizability to such schools. Thus, while we definitely agree with the investigators' statement as provided in question 5, at the same time we recognize that the generalizability actually achieved in the study is limited.

6. The researchers should have used more than one Puerto Rican student because it is too likely that an unusual student was in some way chosen. Comment.

6. If the investigators wished to generalize their results to Puerto Rican students, then clearly more Puerto Rican students are needed for this purpose. The one student may not be typical. If the researchers wish to generalize to the WOIS population (as they clearly state they do), then one Puerto Rican, the investigators assure us, makes about the right proportion. In fact, to use many more than one such student would make the sample unrepresentative of the WOIS population and hinder attempts to make accurate generalizations about the school population.

6. The researchers should have used more than one Puerto Rican student because it is too likely that an unusual student was in some way chosen. Comment.

7. Do you think it was important that the investigators show the two forms of the creativity tests to be equivalent? Why?

7. Although a sensible thing to do, having strict equivalence between the two forms of the tests was not essential. Recall that the order in which the test forms were administered was counterbalanced: that is, half the time one form was given first and half the time the other form was given first. Although not explicitly stated, we believe it reasonable to assume that half the time one form was given after the uninteresting task was interrupted and half the time the other form was given after the uninteresting task. Thus, we can expect any differences in the forms (such as degree of difficulty) to balance out since neither of the motivating conditions or order-of-motivating conditions is associated with one form of the test more than the other form. In this study, equating the forms of the tests is a reasonable, but not essential, procedure to follow.

Student Response

It was important to show that the two forms of the tests were equivalent because "if the difficulty of the tests were different, no conclusions could be reached." Further, "the differences found in the results could be due to the tests rather than to the treatment."

Our Reply

We disagree, for reasons given in our initial answer to this question. Had the investigators not used a counterbalanced design, then we too would have wanted the tests equivalent.

Student Responses

"If the tests had not been equivalent, it would not have been possible to accurately measure score changes."

"It would be very difficult, if not impossible, to get accurate, valid measurement of a child's difference in scores if the difficulty of the tests were different."

Our Reply

It is certainly true that it is difficult to interpret an individual child's difference in the scores of two tests if the tests don't have equivalent units. But in this investigation an individual's difference score was not even computed. Each mean that was computed (see Table 1 in the research report) involved an equal number of scores on the two forms of the test.

7. Do you think it was important that the investigators show the two forms of the creativity tests to be equivalent? Why?

8. "The 'interesting' condition was determined by the child's own interests as indicated by the teacher." Do you approve of this procedure? Explain.

8. We approve of this procedure and other approaches too. Of course, the investigators and readers want some assurance that the task engaged in was interesting to the child. One way to do that is to get such assurance from the child himself. Another reasonable way, it seems to us, is to trust the teachers' judgments, the procedure actually followed. Both of these approaches, asking the child himself and asking the teacher, may be subject to a bias produced when activities are reported or judged to be more interesting than they really are. (The effect of such a bias is to reduce the difference between "interesting" and uninteresting activities and, consequently, to make more difficult finding significant differences between the two treatments.)

Another possible procedure would have been to parallel what was done for the uninteresting task and put all the students in a situation the investigators believe to be interesting to the vast majority of the children. The problem with this procedure is that it is difficult to devise a task one can be sure will be of high interest to a substantial number of children. The advantage of this procedure is that it makes it possible to specify exactly what the interesting task is and to control when the child will be ready to begin testing.

As said before, research design involves compromises and trade-offs. Using teachers' judgments seems to be a reasonable choice, although we would defend as well the other two procedures we mentioned.

8. "The 'interesting' condition was determined by the child's own interests as indicated by the teacher." Do you approve of this procedure? Explain.

Student Responses

"I'm not sure that the teacher can accurately determine those conditions which are interesting to a child."

"Teachers may sometimes be deceived as to a particular child's interest."

"It would have been better to get the child's interest from himself."

"I would tend to trust the involved person's judgment more."

"Let the child speak for himself."

Our Reply

These are all reasonable responses. As indicated in our initial answer to this question, "One way ... is to get such assurance from the child himself." We admit that when possible, measuring something in the most direct available way is often the best procedure. In this case, such a procedure would involved going straight to the child and asking pointedly how interested he is in a particular activity.

Student Responses

"I could only accept the teacher's determination of each child's interest if I knew exactly how she determined it. There is evidence of a lack of control here."

"I don't feel the researchers described this procedure well enough."

Our Reply

Perfectly appropriate reactions.

Student Response

I do not like the procedure of using a teacher's judgment because "this is not an objective method of assessing the interesting condition."

"No -- these are subjective observations."

Our Reply

By "subjective" we assume the students mean that not all observers would agree with the teacher that the child was interested in a particular task. It is true that there is a subjective element in this method of assessment, and it is also true that when inter-judge agreement is absent, the ratings of any one person are very likely to be invalid. Nevertheless, we would caution against an off-hand dismissal of all subjective measurements. The phenomena we may have the greatest difficulty measuring may sometimes be those most worth measuring. One must often ask whether it is better to measure something trivial well or to measure something important poorly.

9. One critic of this study stated that the research assumes for its validity that all children were equally interested in the "interesting" activity. Do you agree with this statement? Why?

9. We disagree. The assumption being made is that any given child will be substantially more interested in the "interesting" task than in the "uninteresting" activity. There is no reason why all children must be equally interested in their "interesting" activity, nor is it reasonable to assume they will be. The comparisons of interest are between individual performances under different conditions and not among children in the same condition.

Student Responses

"There is no way to say that all children were equally interested."

"It is very difficult to measure equal interest."

"The general category, is interested, is too gross."

"The researchers did not take into account the degree of interest in the interesting activity."

"The term interesting can change from day to day with this age group."

Our Reply

The above statements seem to us to be irrelevant to the question asked. The implication in the above student responses is that it would be virtually impossible to demonstrate whether or not the children were equally interested in the "interesting" activity. Although this claim may be true, Question 9 merely asks if there must be equal interest for the study to be valid.

10. The researchers state that each child doing the uninteresting task "was given the same instructions about leaving to 'play games' and about returning to the ongoing activity," as the children in the interesting task condition had received. Do you approve of using the same instructions in both situations? Tell why you answered as you did.

9. One critic of this study stated that the research assumes for its validity that all children were equally interested in the "interesting" activity Do you agree with this statement? Why?

10. The researchers stat that each child doing the uninteresting tas "was given the same instructions about leaving to 'play games' and about returning to the on-going activity," as the children in the interesting task condition had receiv Do you approve of us

10. We do approve of using the same instructions in both situations. We want the two motivating conditions to be as identical as possible in all respects except for those variables the investigators explicitly wish to study. Such similarity makes interpretation of results less ambiguous.

the same instructions in both situations? Tell why you answered as you did.

Student Responses

"The use of the words 'play games' could have influenced the attitude of the child."

"It does not make good common sense to instruct a child to leave an interesting activity to go play games."

"The 'return to ongoing activity' phrase seems to provide a key to the results, i.e. child interrupted from the uninteresting task took more time and gave more responses before returning."

Our Reply

One can take exception to the wording of the instructions, as did the students whose responses are quoted above, and still believe, as we do, that the instructions should be the same for both treatment groups. (Of course "ongoing activity" will mean different things depending on which kind of task was interrupted. But this difference was precisely the difference the investigators wanted to study.)

11. In order to support the claim that the interesting and uninteresting tasks indeed held those qualities for the children tested, the investigators reported their qualitative impressions of the students' feelings about being interrupted; e.g., "That the ongoing activity was indeed interesting to the child was evidenced by the groans, grimaces and footdragging that accompanied the examiner's request" and "The children complained while doing the (uninteresting) task, some called it 'stupid,' and...were uniformly delighted when their participation in the games was requested." Do you approve of such impressionistic reporting in research studies of this type? Why?

11. In order to support the claim that the interesting and uninteresting tasks indeed held those qualities for the children tested, the investigators reporte their qualitative impressions of the students' feelings about being interrupted; e.g., "That the ongoing activity was indeed interestir to the child was evidenced by the groans, grimaces and

11. We approve of such impressionistic reporting. The researcher should be alert to make observations of all phenomena associated with the research investigation. Such observations help us to interpret the more objective data which are available. They provide a fuller picture of the research context and, in this study, lend support to the judgments about task interestedness. Of course, the investigator must be alert to the possibility of experimenter bias and to evidence which is contrary to his position as well as to that which supports his position, and to report both kinds of observations.

footdragging that accompanied the examiner's request" and "The children complained while doing the (uninteresting) task, some called it 'stupid,' and...were uniformly delighted when their participation in the games was requested." Do you approve of such impressionistic reporting in research studies of this type? Why?

Student Responses

We do not approve of such reporting because "children do not mean what they say."

"Emotions could have been made for other reasons."

"Children very often imitate the expressions of their peers without actually feeling the same way."

"The investigators appear to have jumped to conclusions in the matter of children's behavior."

Our Reply

The above responses clearly suggest that the children's behavior should not be taken at face value and that caution should be exerted in interpreting these impressions of the children's feelings. Because the impressions may be difficult to interpret, however, does not lead us to abandon them altogether.

Student Responses

We do not approve of such reporting because it "calls for subjective judgment."

"Findings should include only quantitative measures."

"Impressions are not an empirical measurement."

"Reports are not objective -- but interesting!"

Our Reply

See our reply to the last set of student reactions to Question 8.

12. Note the design as indicated in the first full paragraph of page 354 above the subheading "Design." The investigators want to claim that the kind of activity engaged in (interrupted) before taking the "creativity" tests affects test performance. How many of the following variables have been controlled; that is, which variables are ruled out by the design and alternative explanations for the differential test results found: (a) order in which the two kinds of pretest activities were interrupted; (b) form of the creativity tests; (c) sex of the child; (d) age of the child; (e) "real creativity" of the child?

12. All five variables were controlled. We cannot attribute the observed differences between scores on the "creativity" tests which were taken after an interesting task was interrupted and the scores of the tests after the uninteresting task, to differences in the order in which the two kinds of pretest activities were interrupted; half the children were interrupted from the uninteresting task and the other half were interrupted from the interesting task first. Further, scores from the two forms of the test are equally represented in the two sets of scores being compared. (See Table 1.) Finally, since each child was his own control - that is, was being compared against himself - the sex, age and other characteristics such as "real creativity" were also being controlled. The utilization of a design which rules out so many rival explanations to account for the observed differences in test scores under two different motivating conditions is one of the strengths of the present study.

12. Note the design as indicated in the first full paragraph of page 354 above the subheading "Design." The investigators want to claim that the kind of activity engaged in (interrupted) before taking the "creativity tests affects test performance. How many of the following variables have been controlled; that is, which variables are ruled out by the design as alternative explanations for the differential test results found: (a) order in which the two kinds of pretest activities were interrupted; (b) form of the creativity tests; (c) sex of the child; (d) age of the child; (e) "real creativity" of the child?

Student Responses

"No attempt was made to measure 'real creativity'."

"No one dared to define 'real creativity'."

"Creativity is only a function of the test used."

Our Reply

Variables, such as size of little toe and "real creativity" can be controlled in an experiment even if measurements of these variables are not made or are not possible. One way to do this is by random assignment to treatment groups. Another way to control ability and personality characteristics is to administer the different treatments to the same person -- the technique actually used by the investigators. Since the same people are involved in the two conditions, one cannot claim that the reason for differences in test scores between treatments is because the subjects one condition were older, had more "real creativity," had longer little toes.

Student Responses

"I don't know what you mean by 'real creativity'."

"The concept 'real creativity' confuses me."

"What the hell does 'real creativity' mean?"

Our Reply

We too don't know what "real creativity" means. We used this vague term to emphasize the point that it doesn't matter what such terms mean (for purposes of the issues discussed in this question) since each child is being compared to himself/herself. It is in this sense that we say that "real creativity" has been controlled; it has been ruled out as an explanation for the finding of treatment difference.

13. Competent critiques of experiments require the reviewer to comprehend fully the research design. Especially when several variables are used, many readers find it useful to construct schematic diagrams to serve as a visual reminder of the experimental set-up. For example, if three students (S_1, S_2 and S_3) received treatment M_1 and three other students received treatment M_2, this arrangement might be pictured as shown in equivalent figures 13-1 and 13-2.

M_1	M_2
S_1 S_2 S_3	S_4 S_5 S_6

Figure 13-1

M_1	S_1
	S_2
	S_3
M_2	S_4
	S_5
	S_6

Figure 13-2

For the present study, consider how the design used for the analysis of variance calculations might be illustrated. First, review the special note about p. 354. Second, pick which one(s) of the four schematic diagrams below correctly display(s) the design used.

You do not need to know anything about analysis of variance to answer this question. You do need to know that M_1 was used to represent the interesting motivation condition and M_2 the uninteresting

13. Competent critiques of experiments require the reviewer to comprehend fully the research design. Especially when several variables are used, many readers find it useful to construct schematic diagrams to serve as a visual reminder of the experimental set-up. For example, if three students (S_1, S_2 and S_3) received treatment M_1 and three other students receiv treatment M_2, this arrangement might be pictured as shown in equivalent figures 13-1 and 13-2.

M_1	M_2
S_1 S_2 S_3	S_4 S_5 S_6

Figure 13-1

motivation condition. Further, O_1 and O_2 represent the two orders in which the motivating conditions were present, S_1 to S_{32} the 32 students, and the symbol X a score on the dependent variable. Note: symbols to differentiate the two sexes, the two test forms and the four age groups were not needed as these three variables were not included in the analysis of variance calculations.

(a)

		M_1	M_2
O_1	S_1	X	X
	⋮	⋮	⋮
	S_{16}	X	X
O_2	S_{17}	X	X
	⋮	⋮	⋮
	S_{32}	X	X

(b)

O_1		O_2	
M_1	M_2	M_1	M_2
$S_1 \ldots S_8$	$S_9 \ldots S_{16}$	$S_{17} \ldots S_{24}$	$S_{25} \ldots S_{32}$
X ... X	X... X	X... X	X... X

(c)

	O_1			O_2		
	S_1	...	S_{16}	S_{17}	...	S_{32}
M_1	X	...	X	X	...	X
M_2	X	...	X	X	...	X

(d)

	M_1		M_2	
	O_1	O_2	O_1	O_2
S_1	X	X	X	X
S_2	X	X	X	X
⋮	⋮	⋮	⋮	⋮
S_{31}	X	X	X	X
S_{32}	X	X	X	X

M_1	S_1
	S_2
	S_3
M_2	S_4
	S_5
	S_6

Figure 13-2

For the present study, consider how the design used for the analysis of variance calculations might be illustrated. First, review the special note about p. 354. Second, pick which one(s) of the four schematic diagrams below correctly display(s) the design used.

You do not need to know anything about analysis of variance to answer this question. You do need to know that M_1 was used to represent the interesting motivation condition and M_2 the uninteresting motivation condition. Further, O_1 and O_2 represent the two orders in which the motivating conditions were presented, S_1 S_{32} the 32 students, and the symbol X a score on the dependent variable. Note: symbols to differentiate the two sexes, the two test forms and the four age groups were not needed as these three variables were not included in the analysis of variance calculations.
(See left side of the page for the four diagrams.)

Note: For ease of representation, the dots are used to signify the omission of some of the students and their scores.

13. Diagrams 13a and 13c are equivalent and correct ways to illustrate the analysis of variance design which was used. In both diagrams note that a different set of 16 students belongs to each order-of-motivating condition. (In technical jargon, the variable, student, is nested within the variable, order-of-motivating condition.) Further, note that each person provides scores under both motivating conditions. (In technical jargon, the variable, student, is crossed with motivating condition.)

Diagram 13b is not a correct representation because each student is shown receiving only one of the two motivating conditions and as contributing but one (rather than two) scores on each dependent variable. Design 13d has each student contributing four scores on each dependent variable and has him receiving the motivating conditions under both orders. It too is incorrect.

14. On p. 354, under the subheading "Design," the researchers mention both an age effect and an "age by motivational condition interaction effect." If there were an age effect in this study, it would mean that the average creativity test scores for the several age groups differed -- that children of different ages, as a group, did not do equally well on the creativity tests.

One of the key concepts of empirical research is the interaction between two variables. What would have to be true about the creativity test scores of the children if there was an "age by motivational condition interaction effect?" (The purpose of this question and the discussion to follow is to help you be clear about the meaning of the term, interaction, rather than to ask you about your opinion whether it is reasonable to expect such an interaction.)

Note: For ease of representation, the dots are used to signify the omission of some of the students and their scores.

14. On p. 354, under the subheading "Design," the researchers mention both an age effect and an "age by motivational condition interaction effect." If there were an age effect in this study, it would mean that the average creativity test scores for the several age groups differed -- that children of different ages, as a group, did not do equally well on the creativity tests.

14. The presence of an age by motivational condition interaction effect would mean that the differences in creativity test performance under the two motivating conditions would vary among the several age groups. In other words, such an interaction effect would mean that the differences in the effect (on test performance) of the kind of activity interrupted depends upon the age of the child involved.

One of the key concepts of empirical research is the interaction between two variables. What would have to be true about the creativity test scores of the children if there were an "age by motivational condition interaction effect?" (The purpose of this question and the discussion to follow is to help you be clear about the meaning of the term, interaction, rather than to ask you about your opinion whether it is reasonable to expect such an interaction.)

Student Responses

Interaction between motivational condition and age occurs when "motivational condition affects each age differently."

"It would tell that the affect that motivational condition had was not the same for all age levels."

"The older the child the greater the difference between the two test scores."

"Each age group's amount of score change (under the two conditions) is different from that of each of the other group's."

Our Reply

These responses are essentially correct.

Student Responses

"If there was an age by motivational condition affect, the scores would differ depending on the age of the child."

"As one grows older, his creativity scores increase (go higher) or vice versa."

"The older the child, the higher the test scores would be."

"The scores would vary from age to age."

Our Reply

These responses are incorrect. They describe what would be true if there were an age effect, but they do not describe an interaction effect between age and motivational condition on test score.

15. (a) What dependent variables were used in the study? (b) Is it a good idea to use more than one dependent variable in a study? Why?

15. (a) The dependent variables are those which are affected by the values of the other variables and whose values "depend" upon the conditions under which an investigation is conducted. In this study, the dependent variables are measured by the creativity tests for the effect on such tests is of interest. More specifically, three separate creativity tests were used and two scores - number of responses and number of unique responses - were computed for each test. However, for use in the analysis of variance, the researchers added the number of responses from all three tests to form a new composite variable. They also computed a total uniqueness score by adding the unique response scores for the three tests. These latter scores are reported in the paragraph below Table I on p. 355.

15. (a) What dependent variables were used in the study? (b) Is it a good idea to use more than one dependent variable in a study? Why?

Student Responses

"The dependent variables were the interesting tasks and the uninteresting tasks."

"Conditions before testing, interesting or uninteresting."

"Motivation."

Our Reply

These responses are not correct. The nature of the interrupted task (that is, the motivating condition) was the primary INDEPENDENT variable of the study whose effect on the dependent variables was being studied.

Student Responses

"Score change was the dependent variable."

"The changed scores between the two tests."

Our Reply

It is reasonable to think of the dependent variables as change scores on the several indices of creativity. For example, the statistical test of the difference between creativity scores under the two motivating conditions is equivalent to the statistical test of whether the mean change score is zero.

15. (b) Although there are some inconveniences and possibilities for contamination, on balance we approve strongly of multiple dependent variables in a study since it is possible that the effects being sought will show up for some dependent variables and not for others. A study of the pattern of these results can provide a more complete insight into the phenomena under consideration.

Student Responses

"Using more than one dependent variable is a good idea because it gives a better check on treatment effects."

"Yes, you have a stronger case for generalizability when you use more than one test."

"Yes, it is well to use more dependent variables in order to get more information."

"Yes, especially with a concept like creativity where a definite universal instrument is not available."

"Using several dependent measures is an efficient way of collecting a lot of data at once. Also, if the variable measured is not well defined, as is the case here, using more than one measure provides a way of converging on the concept under consideration."

Our Reply

We concur with these reasons.

16. On p. 354 under the main heading RESULTS, is written: "Motivating condition. The F for this variable was 51.56 and was significant beyond the .01 level." (a) What was significant? (b) What does it mean to be "significant beyond the .01 level?" (If you have not studied statistics, you probably will not be able to answer these questions. Nevertheless, you should study our discussion for it is intended to help you understand frequently used statements like the one quoted above.)

16. (a) Strictly speaking it is the value of F which is significant. (F refers to a statistic computed as part of the analysis of variance.) Also, the 25 point difference in mean number of responses produced under the uninteresting (57.09) and interesting (32.09) conditions was "significant" in the statistical sense of the word.

(b) If the null hypothesis of no difference in the means of the test scores obtained under the two motivating conditions were true, then the probability of obtaining the size differences reported in Table 1 (or differences even more extreme) is less than one chance in a hundred. In this case, significance means rejecting the notion of equal group means in the population. "Beyond the .01 level" means that the probability is less than (i.e., "beyond") .01 that sample results as extreme as those found would occur if the no difference hypothesis were true.

Student Responses

To be significant beyond the .01 level means that "only 1% of the time will such (extreme) results occur because of chance or sampling error."

"The probability is less than 1% that the observed data (or those more extreme) could have occurred only by chance."

16. On p. 354 under the main heading RESULTS, is written: "Motivating condition. The F for this variable was 51.56 and was significant beyond the .01 level." (a) What was significant? (b) What does it mean to be "significant beyond the .01 level?" (If you have not studied statistics, you probably will not be able to answer these questions. Nevertheless, you should study our discussion for it is intended to help you understand frequently used statements like the one quoted above.)

Our Reply

These interpretations are correct. Note that they discuss the probability of the observed data occurring if something (chance alone operating) were really true. A statement of a student that "the (expression) indicates such results as these would happen less than 1% of the time" is correct as far as it goes -- but it needs the qualifying phrase, if chance alone were operating, to be completely correct.

Student Responses

"Significant beyond the .01 level means that the probability of results having been influenced by chance is less than 1%."

"The probability for the chance could be happened less than 1%." (sic)

"Less than 1% possibility that results were obtained by chance."

"It means that less than .01 of the time, chance will be the only causative factor."

Our Reply

The above responses and their variants are the most frequently made, and they are not correct. Equally incorrect are statements that you are 99% confident that chance alone was operating -- e.g. "the chance that differences in creativity scores are caused by manipulation of motivating conditions, and not by chance, is at least 99/100."

The difficulty with these responses is that they state the probability that something is really true beyond the sample results. (In this classical use of probability, either chance alone was operating or it wasn't -- the probability is either 1 or 0.) You should carefully compare our initial answer to this question and the first set of student responses which we said were correct to the set of student responses directly above which we labeled as incorrect. The former give the probability of sample results given a correct chance-alone hypothesis (the correct interpretation); the latter give the probability of the chance-alone hypothesis being correct given the sample results which were found (the incorrect interpretation).

17. Although admitting that norms are not required to test the hypothesis of the study, one student suggested that if national norms for the creativity tests had been reported by the investigators, we could see whether the uninteresting task was responsible for better-than-expected performance or, alternatively, whether the interesting task was responsible for poorer-than-expected performance. Do you agree? Explain.

17. We do not agree. Because the WOIS students in the study may not be typical of the test standardization group, we cannot determine how the WOIS students might have scored, without any unusual pretest conditions, compared with a norm group. Their mean score might have been either lower or, more likely, higher than the norm group mean. Since we cannot establish that the norm group mean and the WOIS group mean under normal testing conditions would be the same, any comparison of the test results with norm group scores would be a meaningless endeavor. For example, if the uninteresting task group scores above the norm group mean and the interesting task group scores at the norm group mean, it could be that: (a) the uninteresting task spurred the students on to better-than-expected performance or that, (b) the interesting task lowered the performance below the level expected of WOIS students.

Somewhat aside, it might have been useful to have a third matched group from the same WOIS population take the tests under standardized administration conditions. Such a third group (a) could help determine if the uninteresting task had a positive effect, the interesting task a negative effect, or both, and (b) could provide data on the typicalness of the WOIS children on the creativity measures. However, an investigator cannot study all the questions he/she might like to, or, in a single study, cannot gather all the data of some benefit. Priorities must be made. We do not criticize the researchers of this study for failure to include such a control group.

17. Although admitting that norms are not required to test the hypothesis of the study, one student suggested that if national norms for the creativity tests had been reported by the investigators, we could see whether the uninteresting task was responsible for better-than-expected performance or, alternatively, whether the interesting task was responsible for poorer-than-expected performance. Do you agree? Explain

Student Responses

"What conditions were the norms obtained under? Conditions of motivation were probably not considered for the norms; therefore, they are not relevant to this experiment."

"No, norms would not be useful because they were not derived under the same experimental conditions as the study."

Our Reply

These students seem to miss the point of the question. It is recognized that the context of testing was different between that found in the study and that present when the tests were normed. The question asked whether, therefore, the difference between the WOIS and norm results could tell us anything about how the context of testing effects test performance -- specifically whether it tended to raise the results (of one of the motivating groups) or lower the results of another. For the reasons given in our initial answer, we concluded that the norm information would not be of much value.

18. About the middle of p. 355 the researchers speculate why, contrary to all the other children, two children gave more responses when taken away from an interesting task than when taken away from an uninteresting task. Should they have made this kind of speculation in a paper of this type? Comment.

18. Definitely. The purpose of research is to explain phenomena. It is quite proper, in fact laudatory, that the investigators share their insights with the reader even though they cannot prove their claims. It is considered good research form to separate speculation and after the fact opinionating from the line of theorizing to which the study was specifically directed. The investigators have clearly made this division.

18. About the middle of p. 355 the researchers speculate why, contrary to all the other children, two children gave more responses when taken away from an interesting task than when taken away from an uninteresting task. Should they have made this kind of speculation in a paper of this type? Comment.

Student Responses

"Yes, it seems reasonable to suggest a possible reason for a result that doesn't 'fit'. It might lead to further investigation, just as their original speculation lead to this study."

Our Reply

We agree.

Student Responses

"This speculation was not necessary, especially since the investigators did not state any attempt toward age-group analysis."

"...suggests that age should have been one of the variables."

"Although the age variable was not incorporated into the design..."

"If they want to replicate, they may want to consider the age factor."

Our Reply

Although age was not included in the analysis of variance calculation, the creativity test data associated with the two motivation conditions reported in Table 1 were further subdivided by age groups. Contrary to the student responses quoted above, the age factor was a variable in the design and was considered.

21. A bit further down on p. 355 the investigators write: "...nor were there significant sex differences with respect to the motivational factor". To what kind of significance (statistical or practical) are the investigators referring here? Or can't we tell?

21. The type of significance is not clear. Not enough information is given to answer the question with certainty. The researchers could be referring to statistical significance although they do not report conducting any significance test of such an hypothesis.

21. A bit further down on p. 355 the investigators write: "...nor were there significant sex differences with respect to the motivational factor". To what kind of significance (statistical or practical) are the investigators referring here? Or can't we tell?

On the other hand, the investigators could merely have noted the differences in mean scores under the two motivating conditions for each sex and concluded, without conducting a statistical test, that the differences of these differences were not of practical significance (that is, were not very important).

Student Responses

(a) "I think the significance they were referring to is the fact that girls tend to do better on tests requiring verbal responses."

(b) "No significance differences is an interesting observation to make because I would guess girls to be more creative than boys."

(c) "I assumed the researchers were saying that there was no difference in the motivation of boys and girls."

Our Reply

These responses illustrate a confusion about the differences being discussed. Regarding response (c), differences in motivation are not involved for motivational condition (type of task interrupted) is an independent -- not a dependent -- variable which is assigned to the children. By no significant sex differences with respect to the motivational factor the investigators are referring to a lack of interaction between sex and motivational condition. That is, they are not claiming that boys and girls scored about the same on the creativity tests (as implied in response b), but rather that the differences under the two motivating conditions for boys and the corresponding differences for girls were themselves not significantly different.

20. In the abstract and at the botton of p. 355, the researchers state that one group was almost twice as creative as the other. What assumption about test scores is necessary to justify this remark?

20. In the abstract and at the botton of p. 355, the researchers state that one group was almost twice as creative as the other. What assumption about test scores is necessary to justify this remark?

20. The assumption is that stating twice as many uses for objects (the results reported by the investigators) means having twice the creativity. (In more technical language, the assumption is that the test scores measure creativity at the measurement level called a ratio scale.) The sentences in question would have been more accurate and less misleading if they had been worded either in terms of "twice as many responses" or "significantly more 'creative'."

Student Responses

Many students said that "the tests must be valid measures of creativity."

Our Reply

This is correct, as far as it goes. More than validity is required, however, before we can make the twice-as-creative interpretation. Our thermometer is a valid measure of temperature, but we would not say a reading of an outside temperature of 6° indicates twice the heat of a reading of 3°. (0° Does not indicate absolute lack of heat just as zero number of responses on the creativity test does not indicate absolute lack of creativity.)

Student Responses

"The standardized norms of a test have to be known before one can make the assumption that one group was twice as creative as the other."

Our Reply

We disagree. Although having norms would permit us to make a comparison with the performances of such a standardized sample, they would not, by some mysterious process, give to the scores this ratio scale property about which we spoke.

21. In the top paragraph of p. 356, the investigators imply relevance of their study to the creativity-intelligence dichotomy. Is their study relevant in this regard? (See special note about p. 356.)

21. In the top paragraph of p. 356, the investigators imply relevance of their study to the creativity-intelligence dichotomy. Is their study relevant in this regard? (See special note about p. 356.)

21. We know of no relevance of this work to the creativity-intelligence dichotomy and are at a loss to explain why any reference to it was made.

Student Responses

"While this study does not disprove the creativity-intelligence dichotomy, it implies that motivating factors may be as important or more important in generating creative responses."

Perhaps the study has relevance to the creativity-intelligence dichotomy "in the sense that a relationship between creativity and intelligence should take motivation into account".

"Tangentially relevant to the larger problem of defining 'creative ability' distinct from 'intelligent behavior'."

Our Reply

Although the above student responses have merit, at best they only make a case for a most indirect kind of relevance that the study might have to the question of whether intelligence and creativity are distinct traits.

22. What would you say is the main conclusion of the study?

22. What would you say is the main conclusion of the study?

22. The investigators would probably claim that their main conclusion is that their results: "...highlight the importance of considering motivational context effects whenever we evaluate psychological or educational test performance." (This conclusion can be worded many ways and still retain its essence -- that test performance depends upon motivation or, in less abstract terms, that the type of task engaged in prior to testing can markedly effect a child's measured creativity.) Regardless which wording you prefer, because of many weaknesses (especially those discussed in the general critique) we cannot assess this study as a rigorous examination of motivational context effects.

Student Response

"A valid conclusion cannot be drawn from an invalid study."

Our Reply

We disagree. For example, fortune tellers are frequently right, especially when predictions are made which agree with one's expectations. One student, frustrated by our answer to question 22, said: "But I feel that we can and should consider the motivational factor in psychological testing." We feel that way too, but our conviction was but slightly strengthened by this particular investigation.

Jason Millman and D. Bob Gowin
Cornell University

Appendix IX

Joanne Reynolds Bronars

Tampering with Nature in Elementary School Science

The Educational Forum
November 1968

"Tampering with Nature in Elementary School Science."

Joanne Reynolds Bronars
The Educational Forum
November 1968

1. Bronars is responding to the need to undertake, "A careful examination of the assumptions underlying experimentation with living things in the elementary school science program." What common name or classification do we give to this kind of critical analysis?

Usually we think of such studies as philosophical research. Somewhat aside, we would like to point out that, one of the traditional tasks taken on by philosophers of education has been the examination of educational theories and practices so that the basic assumptions and values inherent in them may be uncovered and clearly displayed for all to see. The Bronars article follows this traditional form of philosophical research, as it presents an aspect of the elementary school science curriculum and probes beneath the surface of a set of particular activities to ask normative questions about what values we may inadvertently teach by engaging students in such activities.

The obvious audience to which this and a host of similar philosophical articles is addressed is the educational practitioner, forcing him to be reflective about his practice, not in terms of its efficiency or technical propriety, but more fundamentally in terms of its broadly human and ethical dimensions. Since the time of Socrates, philosophers have served as such "gad-flies" to force the public and personal reflection upon our basic values, beliefs, and attitudes, and to thereby bring us to lead the "examined life." Especially in so basic a human activity as education, such an examination is essential to allow us to consider wisely what we are about in terms of its deepest dimensions and far-reaching ramifications for the nurturing of human beings in the ways of civilized life.

1. Bronars is responding to the need to undertake, "A careful examination of the assumptions underlying experimentation with living things in the elementary school science program." What common name or classification do we give to this kind of critical analysis?

Answer:

2. Is the Bronars article an educational research paper?

Yes. It is a study of educational practices, but it is not primarily an empirical study (i.e., it is theory-based rather than experiment based). The predominance of empirical research in education, and the consequent stress on the methodologies of such research, seems to lead many people to believe that only empirical studies conforming to certain methodological norms are properly called "research." Typically, philosophical research continues the oldest tradition of research--that based on careful observation of the world and reasoned thought about it.

3. One form that logical arguments about educational practice can take is the practical syllogism. This form usually has three parts: 1) the normative premise(s), i.e., a statement of what is good; 2) the empirical claims or alleged facts in the case; and 3) the value judgments or conclusions about what should be done. It is never explicitly stated in the Bronars article but one possible argument is the following:

Normative premise: Reverence for life is a good thing.

Empirical claims: a) Many elementary school teaching practices in use today do not instill a reverence for life. b) There are educational practices available which do instill a reverence for life.

Conclusion: Adopt these preferred practices.

Does the fact that this argument contains normative judgments make the argument invalid?

2. Is the Bronars article an educational research paper?

Answer:

3. One form that logical arguments about educational practice can take is the practical syllogism. This form usually has three parts: 1) the normative premise(s), i.e., a statement of what is good; 2) the empirical claims or alleged facts in the case; and 3) the value judgments or conclusions about what should be done. It is never explicitly stated in the Bronars article but one possible argument is the following:

Normative premise: Reverence for life is a good thing.

Empirical claims: a) Many elementary school teaching practices in use today do not instill a reverence for life. b) There are educational practices available which do instill a reverence for life.

Conclusion: Adopt these preferred practices.

Does the fact that this argument contains normative judgments make the argument invalid?

3. cont'd

No. It is a valid argument. Since the conclusions follow from the premises, we say that the argument is valid. The facts as claimed or alleged, however, may not be true as stated. (Note: Logical validity is not the same concept as empirical [fact-based] validity. It is unfortunate that the language of research uses the same term, "validity," in two very distinctly different ways.)

Answer:

4. The article contains the recommendation to change the orientation of elementary science programs from experimentation with living things to observation of them. Is this change necessary in the light of the normative premise, "reverence for life is a good thing?"

No. We agree with Bronars who answered our question (personal communication) as follows:

> "The normative premise is not that of unqualified reverence for life but rather the importance of a developed attitude toward nature which involves a sense of purpose and responsibility. The point is not that experimentation should not be carried on, but that when it is carried on it is for the purpose of thoughtfully conceived ends which adults have assumed responsibility for achieving. That is why I am suggesting that the focus be upon observing where children are concerned."

4. The article contains the recommendation to change the orientation of elementary science programs from experimentation with living things to observation of them. Is this change necessary in the light of the normative premise, "reverence for life is a good thing?"

Answer:

5. The investigator considers three assumptions people use in support of practices that "tamper with nature." If we assume that her arguments against them are conclusive, does such a refutation of the assumptions conclusively support her main argument? Why or why not?

5. The investigator considers three assumptions people use in support of practices that "tamper with nature." If we assume that her arguments against them are conclusive, does such a refutation of the assumptions conclusively support her main argument? Why or why not?

5. cont'd

No. They are logically independent. That is, a person could either agree or disagree with each assumption and still either agree or disagree with recommendations for educational practice.

Even if all three assumptions are rejected, as Bronars rejects them, a person could still agree or disagree with her educational recommendations.

Answer:

6. Bronars is concerned with what children learn when they have learning experiences involving the killing of flies and grasshoppers. To what kind of learning might she have appealed to support her argument?

Many empirical researchers and educational thinkers have commented on the notion of incidental or collateral learning. It is always appropriate to ask what else children are learning when we teach them. The Bronars article stimulates us to ask if we are teaching children to disregard reverence for life when we use living organisms as subjects of experiments in school. We would expect empirical research to show that in some cases we do engender the "wrong" belief systems through such experiments.

6. Bronars is concerned with what children learn when they have learning experiences involving the killing of flies and grasshoppers. To what kind of learning might she have appealed to support her argument?

Answer:

7. Does this article contain any data?

Yes. Check page 277 where Bronars reports responses obtained from college students which indicate a continuum of attitudes toward living things, plus the reasons which justify these attitudes. She also quotes a datum from the New York Times about the availability of living creatures from a publishing company. She also reports other information that is properly considered data.

7. Does this article contain any data?

Answer:

8. Bronars takes exception to some of the present classroom practices in elementary school science. a) Upon what sources

8. Bronars takes exception to some of the present classroom practices in elementary school

8. cont'd

of information does Bronars draw to describe these practices? b) Is there any reason to doubt the validity of her description of these classroom practices? c) Is it important that her description be valid?

a) Elementary school science textbooks.

b) Units recommended by textbook writers may not be the ones actually used in the classroom. Observation of classrooms or reports of activities actually taking place in classrooms would be more valid indicators of classroom practices.

c) Yes and no. If the practices Bronars is complaining about occur only infrequently, then the article no longer has much practical significance. On the other hand, as long as some teachers behave as described (which is most assuredly the case), then the validity of the article turns not on the frequency of these "objectionable" practices but on the clarity and coherence of the arguments.

science. a) Upon what sources of information does Bronars draw to describe these practices? b) Is There any reason to doubt the validity of her description of these classroom practices? c) Is it important that her description be valid?

Answer:

9. Does the information in the second half of page 277 help Bronars to reject Assumption #2 on page 276?

The data will help to the degree that the responses made by college students and referred to in the article will generalize to children. Bronars' data have force only to the degree that we assume children would respond in a similar way.

9. Does the information in the second half of page 277 help Bronars to reject Assumption #2 on page 276?

Answer:

10. Write out Bronar's definition of the word "pest." Most primary dictionary definitions call attention to the historical origin of the word, and define "pest" as any organism capable of causing a fatal disease in epidemic proportions. Obviously her definition differs from the primary definition of most dictionaries. Characterize this difference and discuss

10. Write out Bronar's definition of the word "pest." Most primary dictionary definitions call attention to the historical origin of the word, and define "pest" as any organism capable of causing a fatal disease in epidemic proportions. Obviously her definition differs from the

10. cont'd

its importance in terms of Bronars' argument.

Bronars defines "pest" as "something which causes inconvenience to the one employing the term." Thus Bronars treats "pest" as an evaluative term; the primary dictionary definition is descriptive.

Bronars argues that what some adults (e.g., Science Text writers) consider to be pests and worthless, other people (such as teachers and children in their classes) may not consider as pests and, therefore, should not be harmed. This argument requires that "pest" not be considered a descriptive term (in which case there would be widespread agreement). Rather, her argument requires an evaluative definition so that "we cannot describe certain living things as pests per se." (p. 277). Bronars stipulates her definition of pest, and that this definition contains within it the evaluative phrase, "inconvenience to one employing the term." She has chosen one meaning of "pest" over other meanings readily associated with the term without giving explicit reasons for rejecting the alternative (and competing) meanings. The science textbook writers would be equally justified in asserting that for them "pest" is a descriptive term applied to organisms that cause fatal diseases and epidemics.

However Bronars responds:

> "While the primary dictionary definition of the term 'pest' is descriptive I wished to draw attention to the evaluative one. I agree that I should have spelled out my reasons for doing so. In the same way, however, the textbook writers need to explain their use of the term. As the experiment is set forth the fly is not killed because he is a 'pest' (descriptive) but because it is assumed that no one will object to its being used as a victim. There are

primary definition of most dictionaries. Characterize this difference and discuss its importance in terms of Bronars' argument.

Answer:

10. cont'd

other attitudes towards flies, however, as seen in some of the Scientific American articles on their life cycle. Here reference is made to their beauty and to other kinds of characteristics."

We might add to this by quoting Uncle Toby's reaction to flies, from Tristram Shandy:

"-Go- says he, one day at dinner, to an overgrown fly which had buzzed about his nose, and tormented him cruelly all dinner-time - and which, after infinite attempts, he had caught at last, as it flew by him; - I'll not hurt thee, says my Uncle Toby, rising from his chair, and going across the room with the fly in his hand, - I'll not hurt a hair of thy head: - Go, says he, lifting up the sash, and opening his hand as he spoke, to let it escape; - go poor devil, get thee gone, why should I hurt thee? - This world surely is wide enough to hold both thee and me."

The difference, between evaluative and descriptive definitions of terms, isn't very important with regard to Bronars' paper. It is however generally an important point. Too often in educational research, where the value issues continually impinge on every significant problem, we find this slippage between a descriptive and evaluative definition of some key term. The shift of meanings is often very subtle, and is something one should be constantly on guard to catch.

11. Bronars writes, "Pain is a philosophical concept, not a publicly observable phenomenon." (p. 277, paragraph 2). Give reasons for accepting or rejecting this statement.

11. Bronars writes, "Pain is a philosophical concept, not a publicly observable phenomenon." (p. 277, paragraph 2). Give reasons for accepting or rejecting this statement.

11. cont'd

Here is one reason why we might want to reject the statement as it stands: The statement claims that pain is a philosophical concept. It seems to us that pain is no more a philosophical concept than it is a physical concept, or a medical concept, or a concept of ordinary human experience. It is a feeling. There are many different contexts in which the term "pain" is used to refer to this feeling. However we might want to accept the general sense of the statement because we can make a distinction between a concept (and its sign, such as a word, a gesture, a mark) and that to which the concept refers. Concepts which are relatively rich have attached to them a cluster of criteria (sets of meaning) which we use in correctly applying the term. There is an important sense in which it is appropriate to say that we do not"see a concept." We can, however, reach agreement about what it is the concept refers to, i.e., what is observed. Thus, in common medical practice, doctors reach agreement about pain, the threshold of tolerance for pain, the effectiveness of drugs and other treatments to reduce pain, and so on.

Answer:

12. Bronars writes: "All we can do is to state a value position and invite children to consider it. The teacher's right to compel children to accept it is a moral question..." (p. 277, paragraph 1). Yet the tenor of her article suggests that "reverence for life" must be taught to children. Is she logically inconsistent? Why or why not?

At first glance it might appear that she is being logically inconsistent. Bronars states as a fact (p. 277, paragraph 3) that there are a variety of feelings which children have about living things. Thus, presumably, some could have notably tougher ideas about living things than Bronars might wish. To suggest that "reverence for life" must be taught to these tough-minded children implies that the teacher needs to go beyond merely inviting them to consider this value.

12. Bronars writes: "All we can do is to state a value position and invite children to consider it. The teacher's right to compel children to accept it is a moral question...." (p. 277, paragraph 1). Yet the tenor of her article suggests that "reverence for life" must be taught to children. Is she logically inconsistent? Why or why not?

Answer:

12. cont'd

In fact, though, she is not being inconsistent. To suggest that something must be taught in schools does not entail the suggestion that children must be compelled to accept it.

13. Bronars suggests that science study be focused on observation of living things in their natural habitat. She also suggests that learning with actual objects (i.e., living organisms) may not be as effective as learning with representative materials. Is there a contradiction in these two suggestions?

Again, she is not being inconsistent. To explain why we might best quote her own response (personal communication). "Reference is made to the kinds of science study that would best be carried on through the use of field observation techniques (p. 277) and that would best be carried on through the use of representative materials (p. 279). There is no contradiction but rather a reference to different kinds of phenomena."

14. What is Bronars' main question about effects of the educational practices examined in her report? Briefly sketch how this question might be answered empirically.

The main concern of the paper seems to be the relation between certain activities in elementary science practice and two related values: a) attitudes of children concerning reverence for life, and b) attitudes of children toward the balance of nature.

An empirical study comparing these attitudes in children who both have and have not been exposed to the practices of elementary school science which are being questioned here might help determine the effects of these practices upon such attitudes.

13. Bronars suggests that science study be focused on observation of living things in their natural habitat. She also suggests that learning with actual objects (i.e., living organisms) may not be as effective as learning with representative materials. Is there contradiction in these two suggestions?

Answer:

14. What is Bronars' question about the effects of the educational practices examined in her report? Briefly sketch how this question might be answered empirically.

Answer:

14. cont'd

However, it must be stressed that Bronars' arguments cannot be "validated" or "disproved" by any possible result of such an experiment. She wants to argue that classroom activities that involve the heedless and casual killing of living things are wrong in themselves. If we ran a test that discovered killing people did not seem to effect people's attitude to human life we could hardly claim to have shown that killing people is all right.

Appendix X

Edwin M. Bridges, Wayne J. Doyle, and David J. Mahan

Effects of Hierarchical Differentiation on Group Producitivty, Efficiency, and Risk Taking

Administrative Science Quarterly, September 1968, pp. 305-319

Effects of Hierarchical Differentiation on
Group Productivity, Efficiency, and Risk Taking

Edwin M. Bridges, Wayne J. Doyle, and David J. Mahan

Administrative Science Quarterly, September 1968, pp. 305-319

SPECIAL NOTES

In both the Results and Discussion sections, the investigator discusses the use of one-tailed statistical tests (as opposed to two-tailed tests). These two types of statistical tests are frequently used to analyze the type of data presented in this paper. When a researcher tests the statistical significance of the difference of the mean scores for two groups, he calculates those differences (called rejection regions) which, if they occurred, would be so large as to cause him to reject the hypothesis of no difference in group means in the population -- that is, to reject the hypothesis that the differences in means are due only to chance. If a researcher is willing to consider large observed differences as reason to reject this hypothesis of no difference regardless of which group had the higher mean, the researcher is conducting a two-tailed test. (The rejection regions are at the two tails of a distribution of expected differences.) If, as the investigators of this paper have done, only large differences in favor of a specific group will lead the researcher to reject the no difference hypothesis, then a one-tailed test is being conducted. When a one-tailed test is used, finding a difference in favor of the group NOT expected to be superior will not permit the researcher to reject the chance alone hypothesis, no matter how large that unexpected difference is.

There is debate among statisticians over the appropriateness of one-tailed tests. The point to keep in mind is that when one-tailed tests are used, smaller group differences are needed to reject the chance alone hypothesis provided, of course, the differences are in the direction hypothesized. This is true because one large rejection region is used rather than two smaller ones. Had the researchers used a two-tailed test, the differences in efficiency scores (hypothesis 2) and in risk taking scores (hypothesis 3) would not have been statistically significant at the .05 level.

Effects of Hierarchical Differentiation on
Group Productivity, Efficiency, and Risk Taking

Edwin M. Bridges, Wayne J. Doyle, and David J. Mahan

Administrative Science Quarterly, September 1968, pp. 305-319

The article is divided into the following sections: Introduction, Hypotheses, Method, Results, Discussion and Concluding Remarks. Evaluate the article critically, organizing your remarks into six groups to correspond to the six sections of the paper listed above. Be sure to cite strengths as well as weaknesses.

Effects of Hierarchical Differentiation on Group Productivity, Efficiency and Risk Taking

Edwin M. Bridges, Wayne J. Doyle, and David J. Mahan

Administrative Science Quarterly, September 1968, pp. 305-319

A MODEL APPRAISAL

Introduction

1. Although brief, the introduction is a good one. It provides a clear idea of the content of the paper and makes a case for its significance. We believe the general problem is an important one, particularly in the present times of doubt about authoritarian forms in many kinds of organizations -- from communes to private industry, from educational institutions and classroom groups to bureaus of the government. Further, the criteria used for judging forms of organization (i.e., productivity, efficiency and risk taking) are important ones.

2. Student Response. "A peer group is not necessarily undifferentiated. Peer groups hold their internal differentiation and the study does not take this into consideration."

3. Our Reply. The author does write, "undifferentiated groups, i.e., peer groups." We agree completely with the student's remarks and make this point ourselves in another context (see paragraph 15 of our model appraisal).

4. Student Responses. The Introduction is poor in that the investigators, "did not present a review of the existing research," and, "failed to define the hierarchically differentiated and undifferentiated groups."

5. Our Reply. Although we agree that it is important to review existing research and to define key concepts, we do not believe that it is necessary to do these things in the Introduction. The authors do refer through footnotes to the work of others in which the concept of hierarchical differentiation is described.

6. Student Response. "The terminology was so involved that it was difficult to wade through."

7. Our Reply. Many students made similar statements, not only in regard to the Introduction but in reference to other sections as well. The investigators do have an obligation to communicate clearly; but we must remember this article is not meant for consumption by the general public. The language of science cannot be the same as everyday language for the latter is too imprecise. On the other hand, unnecessary jargon can be confusing and some balance is needed.

Hypotheses

8. The hypotheses section of this paper does not merely list the three principal hypotheses which guided the investigators' early work on the problem, but goes beyond to provide a helpful rationale for expecting the results hypothesized.

9. The investigators should be commended for the way in which they used social science concepts and theory to guide their research on administrative problems. This reliance on theory: a) increases the probability that relationships will be discovered; b) provides a way to explain and to account for differences when they do occur; and c) facilitates additional inquiry.

10. The investigators hypothesize (#3) that in differentiated groups the subordinates who generate ideas will hesitate promoting them, and thus fewer of these generated ideas will be presented by the group to the research worker (there will be low risk taking). One could argue the opposite as follows : because of the greater inhibition in differentiated groups, subordinates will only suggest ideas which they feel can be defended; thus the ideas suggested in such a differentiated group are more likely to be accepted by the entire group for presentation to the research worker (there will be high risk taking).

11. Student Responses. "Hypothesis 3 was based on opinion." "The subjective statements used in the explanation of each hypothesis have not been proven."

12. Our reply. These student readers evidently believe it not worthwhile to engage in research whose hypotheses are generated from rationales which are "opinion" and "not...proven." The line of reasoning behind hypotheses can range from radically speculative ideas and mere opinion to coherent rationales and logically tight theories. It may well be true that the payoff of research depends upon the location of the line of reasoning along this continuum. It is our judgment that the investigators utilize a thoughtful (if not compelling) line of reasoning which is much more than unsubstantiated opinion.

Method

13. The investigators chose to use an experimental method rather than a survey or correlational design even though in the field of educational administration the tradition of nonexperimental research is especially strong. A more usual procedure to study the effects of a variable like "hierarchical differentiation" would be to administer an instrument to first identify school groups which differ naturally on this variable, and then to compare these groups with respect to the dependent variables. Our purpose is not to claim that the variable manipulating experiment conducted by the investigators is superior to the more traditional status study (although we suspect it is), but rather to highlight the fact that there is usually more than one way in which a problem can be researched.

14. Sample:

Ten groups, each consisting of a principal and three teachers, were classified as hierarchically differentiated. Ten other groups from the same schools, each consisting of four teachers, were classified as hierarchically UNdifferentiated. Thus, groups were considered hierarchically differentiated or undifferentiated solely on the basis of whether or not the principal was present.

15. Whether or not this distinction (hierarchically differentiated vs. undifferentiated) corresponds to the conceptual definition of "status difference" was unexamined, partly because no conceptual definition was provided. It is quite conceivable that something other than "status" was being manipulated by the investigators, such as "maleness", "personal dominance", "differential familiarity", or "emergent vs. appointed leadership." Since status systems exist within teaching staffs, it is not certain that the all teacher groups, supposedly without status differences, really differed on this dimension from the groups in which the principal was present. The investigators would have been well advised to check the correspondence between the operational and conceptual definitions, perhaps by means of a post-experiment questionnaire or interview.

16. It should be noted that the main comparison was between pairs of groups selected from the same school. Thus, differences between groups could not be attributed to school differences since the groups were essentially matched in this regard. The investigators should be commended for insuring that the basic comparison between the two types of groups was valid, even though results might not be generalizable to all types of school groups in all localities.

17. Procedures:

Under the section, Procedures, the researchers describe the problem to be solved (the doodlebug problem) and the methods of administration. The adequacy of this problem, the decision making procedure, and the role of the experimenter deserve comment at this point.

18. Problem Adequacy. One should note the difference between the doodlebug problem presented to the groups and the range of real life problems to which such groups generally attend. Many of the educational problems faced by teachers have no clear answer as does the doodlebug problem and we may therefore question whether results obtained using this special problem can be made more generally applicable. Closer inspection of the measures generated from the doodlebug problem will reveal that the problem is used to measure the ability to overcome normal beliefs rather than to measure problem solving ability in the usual sense. The doodlebug problem is more like a puzzle than a problem in decision making.

19. Further, the doodlebug problem was too difficult for use in testing differences in productivity in the synthesis phase of the task. A pilot study could have shown this fact. (see p. 4a)

20. Finally, mention of the three beliefs to be overcome (in paragraph 1, p. 310) well before describing them (in footnote 8) is weakness of reporting style.

21. Although the task was, in a sense, artificial and trivial, it does have the virtues of having been thoroughly studied in previous research, and is of such a nature that principals should be equally adept as teachers at solving it. This last point is important, for if the problem were something that principals could be expected to handle more easily than teachers, the group differences could be attributed to the particular skills of principals rather than to the hierarchical differentiation of the group.

22. Although the choice of a suitable problem was a difficult one, we believe the researchers should have chosen one or more tasks more closely related to actual school situations. (see p. 4a)

23. Decision making procedures. For purposes of reaching decisions within each group involved in the problem solving situation, a parliamentarian arrangement in which the majority rules was decided upon. This is an unusual method for school personnel to use for reaching decisions. More likely is the centralist constitutional arrangement in which a group is bound by a decision reached by the person in final authority. As recognized by the investigators themselves (see footnote 14), use of a majority-rule procedure makes it difficult to explain the results. It is when the centralist arrangement is used that status hierarchies in groups are expected to matter most because this form recognizes and utilizes status differences in its operation. Thus, not only is the generalizability of the study weakened by use of an atypical decision making procedure, but the very rationale for expecting status differences to be operating is less applicable to the parliamentarian arrangement and, consequently, interpreting differences to status differences in the groups is very hazardous indeed.

24. Experimenter role. One weakness of the report is its failure to describe clearly or completely the role of the experimenter during the problem solving sessions. It is nowhere indicated how many experimenters were used or the extent to which they had been trained for participation in the sessions. The last sentence in footnote 13 mentions that the experimenter "clarified" ideas. Elsewhere it was stated that the research worker gave immediate feedback (p.308) and could be asked questions (p. 309). All this suggests that the experimenters may have had a more active role in the problem solving sessions than we might believe. It is important for us to know the exact nature of the experimenters' role more accurately to assess possible experimenter bias (or more generally, "instrumentation" effects) and the additional restraints that may have been operating on the behavior of the participants.

Paragraph 19

Note from the investigators: Actually, we did conduct a pilot study, and thirty minutes appeared to be ample time for solving the problem. In seeking to identify a possible cause for the unexpected outcome, we realized that we had blundered. The populations of subjects were different. Whereas the experimental subjects were teachers, the subjects in the pilot study were sophomores in the liberal arts college of a highly selective university. The implication of this situation is unmistakable; the teachers and principals in our sample were not as able as the college sophomores. We chose to safeguard the interests (namely the self-esteem) of our subjects by withholding potentially harmful information. We certainly are not the first researchers who have wrestled with the choice of what to report and what to withhold; this decision frequently arises when the objects of research are human beings.

Paragraph 22

Note from the investigators: We share the reviewers' belief that tasks more closely related to actual school situations should have been used, but feel that they have slighted a persistent dilemma faced by those choosing an experimental approach to research. An experimenter hopes to design a study which has both internal and external validity. A study is said to possess internal validity if the experimental stimulus did in fact make some significant difference in this specific instance. External validity refers to representativeness or generalizability. As Donald T. Campbell ["Factors Relevant to the Validity of Experiments in Social Settings," Psychological Bulletin, 54 (1957), 297-321], has noted,

> Both criteria are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness...If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration.

In selecting the problem, we sought to identify one in which neither principals nor teachers would have an advantage. We were not confident that we could develop a school related problem which could be handled with equal ease or difficulty by principals and teachers. We, therefore, sacrificed external validity in the interests of internal validity, a not uncommon sacrifice at that.

25. Finally, note that it is not made clear how the solutions of the groups were "passed on" to the experimenter. Did the administrator, when present, have any special function in the passing on activity?

26. Student Responses. Many students mentioned the failure of the investigators to give adequate description of the following areas: a) the doodlebug problem; b) method for selecting the principals; c) method for selecting the teachers; specifically if they were volunteers and why there were so many females; d) teacher experience and age; e) effect taping of the sessions had on inhibition; f) fatigue of those meeting in the afternoon sessions; and g) procedures, if any, for checking whether the morning session teachers talked to their afternoon session colleagues.

27. Our Reply. a) In our opinion, the doodlebug problem was adequately described both on page 310 and in footnote 8. Further, an accessible reference where a still more complete description can be found is provided.

28. b) through e). Printing costs are high and there are more papers than scholars have time to read. There facts argue for a judicious choice of those facts and details to be described in the research report itself. Clearly, that information which has the most bearing on the validity of the comparison between the two groups and on the generalizability of the findings should be included. For example, the investigators thought it more important to mention the name of the city than the ages of the teachers. A student argued that if the teachers differed in age they could not be considered "peers" regardless of where they were on the "organizational chart." Previous research can give us clues about what variables are likely to be important and thus worthy of description in the research report.

29. f) and g) Since treatments were randomly assigned to session times, and since small differences between sessions, on the dependent variables were noted, it does not seem important to us that the fatigue and prior knowledge differences of the two groups be described.

30. Student Responses. "Several problems of varying types should have been used."

"A larger cross-section of the population should be used, and not just teaching personnel."

31. Our Reply. These investigators wanted to make very general statements about group structure and group problem solving. It is essential that they design their research in a way that enhances the generalizability of their findings. One way they increased the generalizability of their work was by including several measures of problem solving ability. Had they not given the same problem to all 20 groups and had they used other types of hierarchically differentiated groups (the two student suggestions quoted above) their study

would have had that much more value. We do not believe that most researchers give enough thought or effort to designing studies to maximize generalizability. Ways this can be done without increasing the cost of the research are described by Millman.* A list of ways that research can be said to generalize is presented by Bracht and Glass.**

Results

32. The Results section includes a description of the measures used to represent the dependent variables of production, efficiency and risk taking, as well as a statistical comparison between the two types of groups on these measures.

33. The Measures.

Production. Using the number of beliefs overcome as a measure of production seems reasonable enough, although one could argue that the three beliefs should not be given equal weight.

34. Efficiency. Time to overcome the first belief is a good measure to test hypothesis two since it is in the early stages of group work that relative differences in speed of performance are expected. According to the investigators' predictions, developing the pattern of interpersonal relationships needed for efficient problem solving, "will require more time in hierarchically differentiated groups than in undifferentiated groups." (p. 308) This time consuming process will produce a difference in efficiency more evident in the beginning of the problem solving situation than at the end.

35. The distribution of time to overcome the first belief is likely to be skewed, with a few groups taking relatively a very long time. Such groups will have a disproportionate effect on the mean of all 10 or 20 groups. Further, one could argue that taking an extra minute of time early in the problem solving effort should count more than an extra minute after the group already has worked 15 or 20 minutes. For both of these reasons, it would have been a good idea to use as the index of efficiency not time per se but some function of the time score such as the reciprocal of time (i.e., one divided by the time score) or logarithm of time. Such functions have the desired properties.

* In the Service of Generalization, Psychology in the Schools, 1966, 3, 333-339.

** The External Validity of Experiments, American Educational Research Journal, 1968, 5, 437-474.

36. Risk taking. The risk taking measure used by the investigators is the difference between the number of generated solutions and the number presented to the experimenter. A large difference actually means low risk taking because the group seems unwilling to "risk" presenting solutions to the experimenter.

37. To name such a measure "risk taking" implies there is something to be lost in suggesting inaccurate solutions to the experimenter and that something is being risked in presenting other than the correct answer to the problem. Since the groups were in no way penalized for presenting such incorrect answers, what risk is involved to the group is not clear. The individual is said to risk "failure in the eyes of his superior." But this fear of failure of the individual is not reflected in the group difference score which is used as the risk taking index. Thus, we do not believe this group risk taking index is a measure of risk taking in the usual sense, or in the sense used in organizational theories, but more a measure of how reasonable the suggested solutions seemed to the group involved.

38. The definition of risk taking given by the investigators on page 312 is a stipulated definition and not an operational definition. To be an operational definition, the operations or procedures that must be followed to get the discrepancy index are needed. Of course, a researcher may give a stipulated definition of his operational definition; not all stipulated definitions are operational definitions.

39. Statistical Analysis.

The student should note that although 80 individuals were involved, the investigators correctly compared only the 20 group results. The group, and not the individual, is indeed the correct unit for analysis.

40. The likely skewness of the distribution of the "efficiency" measures has already been commented upon. The "productivity" measure also represents a skewed distribution since most of the groups must have overcome all three beliefs in order for the mean scores to be so close to the maximum score of three. Thus, as was true for the efficiency measure, a few groups which could not get off the ground, so to speak, would have a disproportionate effect on the mean productivity score for all ten groups. The investigators should have presented more of the groups' performance than merely the means.*

---

* Further, because of non-normal distributions and likely large differences in variability between the two types of groups, the mathematical assumptions of normality and homogeneity of variance underlying the proper use of the t test are being violated in the testing of hypotheses 1 and 2. The effect of these violations on the accuracy of the significance test may be quite minimal, however.

Add after footnote page 7

Note from the Investigators: When the assumptions constituting the statistical model for a test are not met, doubt arises concerning the meaningfulness of a probability statement about the hypothesis in question. There is some empirical evidence to show that slight deviations from the assumptions underlying parametric tests may not have radical effects on the obtained probability figure (Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences, New York: Mc-Graw-Hill Book Company, Inc., 1956) and that major effects are likely to occur only when the sample is small (William L. Hays, Statistics for Psychologists. New York Holt, Rinehart and Winston, 1963.) What constitutes a slight deviation or a small sample is unclear, however. In light of the confused picture and to satisfy our o curiosity, we analyzed data by means of the t-test and the Mann-Whitney U test, a non-parametric statistic. The results were identical. As the reviewers noted, th effects may indeed be quite minimal.

41. Since the groups were matched by schools, the appropriate t test involves comparing 10 matched pairs instead of two independent sets of 10 groups each. A different formula for computing the t statistic should have been used.*

42. We also take exception to the use of one-tailed tests. (Recall the special notes in regard to one-tailed tests.) The use of one-tailed tests is most defensible when there is no reasonable way to explain results in favor of the hierarchically differentiated groups. For example, contrary to hypothesis 3, it might be that in hierarchically differentiated groups, generated solutions are more apt to be presented (i.e., greater risk taking exhibited) because subordinates would not want to offend their peers in front of the principal. Had two-tailed tests been used instead of one-tailed tests, the first two hypotheses in the paper would not have been statistically significant. (see p. 8a)

43. Regardless of the t formula used or how many tailed tests were employed, the following interpretations seem reasonable: for each of the three dependent variables there were noticeable differences between the average performance of the groups of each type; it appears unlikely, but still possible, that chance alone accounts for these differences.

44. In the last two paragraphs on page 312, the authors perform two additional analyses. They test whether there is a difference on the dependent variables between the before-school groups and after-school groups, and they compute the correlations among the dependent variables. Had these differences or correlations been large, it would have suggested modifications in the interpretations of their results. The investigators should be commended for taking these precautions and for searching for rival explanations.

45. Student Responses. "I really do not have enough background in statistics to evaluate this section well." "We have not covered this kind of statistics in class." "This section (due to my complete density in the area of knowledge of statistics) is impossible for me to comment on as it was all foreign to me."

46. Our Reply. Of course the kind of discussion we gave in some of the paragraphs 39-44 of the model appraisal does require a statistical sophistication. However, do not be led into thinking that because you lack this sophistication you cannot look at the results of studies critically. The writing of paragraphs 33-38 did not require this sophistication. Without statistical expertise you can still question whether the data presented are relevant to the questions asked. Don't give up too quickly.

*The different formula would have 9 degrees of freedom instead of the 18 reported by the investigators. If, on the average, the two groups from the same school were more alike in their problem solving behavior than differentiated and undifferentiated groups from different schools (as we suspect them to be), then a higher value of t would result. From the data available to us, we suspect that had the investigators used the t formula for matched pairs the results would have been even more statistically significant.

Paragraph 42

Note from the investigators: There are those who, like us, feel that a one-tailed test can be used when there is a theoretical basis for a directional hypothesis (Allen L. Edwards, Statistical Methods for the Behavioral Sciences. New York: Rinehart and Company, Inc., 1958); there are others, however, who feel that the potential for misusing a directional hypothesis is substantial (Gene V. Glass and Julian C. Stanley, Statistical Methods in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1970.) The only statement which can be made with certainty is that a debate over the merits of testing directional versus nondirectional hypotheses has raged for the past twenty years (e.g., see Cletus J. Burke, "A Brief Note on One-Tailed Tests," Psychological Bulletin, 50 (1953), 384-87; and David B. Peizer. "A Note on Directional Inference," Psychological Bulletin 68 (1967), 448).

47. Student Response. "The results didn't allow for different intelligence or personalities of individuals."

48. Our Reply. The student could mean two things by her statement. First, she could mean that the procedures did not equate groups on intelligence or personality. To that we would reply that the random assignment of teachers to groups has the effect that such initial group differences in intelligence or personality would be due to chance alone and they can be estimated by techniqu of statistical inference. Alternately, the student could mean that the results did not provide separate analyses for individuals of different intelligence or personality. To that we would reply that such an analysis would have to be for the group as a whole (criterion scores are for the groups, not individuals within the groups). The small number of groups (10 within each treatment) would make such as analysis of limited value.

Discussion

49. The discussion, perhaps misnamed, consists of the investigators' attempts to provide evidence relevant to three rival hypotheses: 1) the lower proportion of solutions presented to the experimenter in the hierarchically differentiated groups was due to the tendency of ideas advanced by low ranking members to be passed over rather than to a reluctance on the part of subordinates to take risk (pages 313-315, first two lines); 2) a reluctance of subordinates to criticize the ideas of superordinates and/or an uneven distribution of social support was the reason for greater productivity in the undifferentiated groups (p. 315-317); and 3) the curtailment of competition for respect in the differentiated groups was responsible for the differences in productivity between the two types of groups.

50. Some of our objections to what is written in the Discussion section paralle remarks made in connection with our appraisal of the Results section. Our displeasure with the risk taking measure remains. See * below for the remainde of this paragraph.

51. Perhaps most disconcerting is the investigators' belief that the number of ideas initiated as a measure of the degree to which group energies are mobilize is a serious test of the competition for respect explanation. (We wonder why the investigators are so willing to accept Blau and Scott's third explanatory factor after they rejected the first two.)

* The t test should have made use of the fact that the schools were matched. The chi-square tests are inappropriate since the responses of the same person are represented by more than one frequency in the table and thus the independer assumption underlying the proper use of the chi-square test was violated.

principal as an observer. He will have a certain affect on the situation."

53. Our Reply. Recall that the purpose of the additional study was to determine if "...the lower proportion of solutions presented to the experimenter in the hierarchically differentiated groups was not due to a reluctance by subordinates to take risks, but rather to the tendency of ideas advanced by low-ranking group members to be overlooked." (p. 313) To determine which of these is more likely it was necessary, as the investigators did, to design a situation in which the same reluctance by subordinates to take risks was possible (i.e. principal present) but one in which the principal has no chance to overlook subordinates' ideas (i.e., present but no active role).

Concluding Remarks

54. The phrase, "tend to confirm", in the first sentence under the Concluding Remarks section is too strong. "Confirm" suggests that the evidence is now sufficient to warrant acceptance of the conclusion. We do not believe the investigators meant to give such assurance.

55. We commend the investigators for mentioning ways in which the research is still incomplete (e.g., they did not investigate centralist constitutional arrangement or problem solving at the synthesis phase) and for pointing to needed research on the topic.

A Summary of Our Assessment

56. The problem the investigators set out to study is an important one and their study provides a good illustration of the close and sensitive integration of theory and data. We see the choice of the doodlebug problem as an unfortunate one and further object that the researchers offer no evidence that they have successfully manipulated the hierarchical differentiation variable. The investigators did take pains not only to test their predictions but also to examine the assumptions upon which their predictions were based. We believe that the investigators went about their research business in order to protect themselves from improper inference and not just to convince other that they had conducted their study properly.