DOCUMENT RESUME

ED 273 656 TM 860 505

AUTHOR Lynch, Kathleen Bodisch
TITLE Effect Sizes of Programs Applying to the Joint Dissemination Review Panel.
PUB DATE Apr 86
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Effect Size; Elementary Secondary Education; Evaluation Criteria; Evaluation Methods; Meta Analysis; Multiple Regression Analysis; Predictor Variables; *Program Design; *Program Effectiveness; Program Evaluation; Program Proposals; *Program Validation; Success; *Validated Programs
IDENTIFIERS *Joint Dissemination Review Panel

ABSTRACT

Educational programs and evaluations which were submitted to the Department of Education's Joint Dissemination Review Panel (JDRP), in order to be named validated programs, were studied to identify program characteristics associated with large versus small effect size. Effect size was calculated for 165 out of 232 submittals reviewed by JDRP from 1980 through 1983. Results indicated the largest variance in effect size was explained by content area (highest effect size in natural science and lowest in reading, language arts, and mathematics); and secondly, by reported annual operating funds (less than $100,000 had higher effect size). Other program characteristics related to large effect size were gifted participants, regular classroom setting, urban or suburban setting, and behavioral versus attitudinal or affective objectives. Lowest effect sizes were associated with handicapped audiences and special facilities. Locally developed tests, external evaluators, and randomized evaluation designs were associated with higher effect sizes. The combination of program and evaluation features which accounted for effect size were type of test, formula used to calculate effect size, type of objective, and evaluator affiliation. It was concluded that effect size data should not be interpreted simplistically; facile comparisons of the absolute values of effect sizes can tbe misleading. Several tables are provided. The appendixes consist of the JDRP Submittal Analysis Form as well as supplemental instructions for completing the form. (GDC)

Effect Sizes of Programs Applying to the Joint Dissemination Review Panel

Kathleen Bodisch Lynch, Ph.D.
University of Virginia
School of Medicine
Box 382
Charlottesville, VA 22908
804-924-2563

Paper presented at the Annual Meeting of the
American Educational Research Association
San Francisco, CA
April 16-20, 1986

Effect Sizes of Programs Applying to the Joint Dissemination Review Panel

Kathleen Bodisch Lynch
University of Virginia
School of Medicine

The primary purpose of this study was to identify characteristics of educational programs and evaluations which were related to the size of program effects. This was done by calculating effect sizes for programs applying for validation from the U. S. Department of Education's Joint Dissemination Review Panel (JDRP) from 1980 through 1983. The effect size is a standardized measure of program outcomes. After effect sizes were calculated, they were related to characteristics of the programs and evaluations through traditional techniques of data analysis.

The largest proportion in the variance of effect sizes was explained by differences in content area among the programs. Effect sizes were highest for programs in the natural sciences, and lowest for those in reading. Of the evaluation characteristics considered, the type of outcome measure best explained the variation in effect sizes. Higher effect sizes were found for locally developed instruments than for published tests. The type and quality of the evaluation design were also related to effect size, with randomized designs and high quality evaluations being associated with higher effect sizes.

Effect Sizes of Programs Applying to the Joint Dissemination Review Panel

Kathleen Bodisch Lynch
University of Virginia

The primary purpose of this study was to identify characteristics of educational programs and evaluations which are related to the size of effects produced by programs applying for validation from the United States Department of Education's Joint Dissemination Review Panel (JDRP). The JDRP is a panel of experts whose purpose is to review educational products and practices in order to determine whether they are effective. The JDRP makes this determination through consideration of a 10-page written report in which a program describes its goals, activities, costs, implementation requirements, and evidence of effectiveness. Programs also make an oral presentation before a subgroup of three to seven members of the JDRP; decisions concerning approval or rejection are made on the basis of a simple majority vote of this subgroup.

The JDRP considers educational products and practices which it approves to be worthy of nationwide dissemination. For the JDRP, approval of a program indicates that the program has, through a credible evaluation, persuasively demonstrated an exemplary degree of effectiveness in achieving its stated objectives. To be more specific, the evaluation-based

data and related evidence submitted by a program must be valid and reliable, the intervention and its effects should be replicable, and the effects must have been of sufficient magnitude to be considered statistically and educationally important.

While the size of effect that a program produces is only one of the factors that the JDRP considers in deciding whether a program should be approved, it may be one of the more difficult factors to evaluate, according to a former Chairman of the JDRP (J. Schiller, personal communication, July, 1983). Besides the complexity of deciding what constitutes an "educationally significant" magnitude of change, Panel members face the problem of trying to compare the size of effects detected by a wide variety of outcome measures. For example, a claim that a program can produce a mean gain of 20 points on Science Knowledge Test A might be considered remarkable, while a similiar claim based on use of Science Knowledge Test B might be considered trivial. Against what standards can such results be judged?

In this study, techniques of meta-analysis were used to allow comparisons of the magnitude of effects across a variety of program and evaluation characteristics. Specifically, effect sizes were calculated for programs applying for JDRP approval during the years 1980 through 1983. Effect sizes are obtained by transforming the results of an evaluation into a standard score. In the simplest case, for studies involving a

comparison between a treatment and a no-treatment group, the difference between the means of the treatment group and the comparison group is divided by the standard deviation of the comparison group (Glass, 1977). This formula allows different outcome measures to be compared by expressing the difference between groups in terms of standard deviation units, rather than in terms of the original metric of the measuring instrument that was used. After effect sizes were calculated they were related to characteristics of the educational programs and evaluations through traditional methods of data analysis.

Results

During the four years covered by this study, 165 out of 232 submittals reviewed by the JDRP (or 71%) provided the data necessary to calculate effect sizes. Some JDRP submittals reported effectiveness data for more than one content area, target audience, type of objective, type of outcome measure, type of evaluation design, and grade level. Effect sizes were computed separately for each of these variables for each program, and were then aggregated across grade level and type of outcome measure (published vs. locally developed) within programs. Consequently, the number of effect sizes retrieved from each submittal varied, with a total of 263 effect sizes obtained.

The mean effect size over all the programs for which this statistic could be calculated was 0.89. The largest

proportion of the variance in the distribution of effect sizes was explained by content area; effect sizes were highest for programs in natural science and lowest for reading programs (see Tables 1 and 2). Second in importance in explaining the variation in effect size was the amount of annual operating funds reported by the program. Programs with less than $100,000 had higher effect sizes than those with $100,000 or more. Also found to be related to effect size was the type of objective addressed: effect sizes related to behavioral change were higher than average; those related to attitudinal change were lower than average.

Of the evaluation characteristics considered, the type of outcome measure used best explained the variation in effect sizes; locally developed instruments tended to yield higher effect sizes than published tests (see Tables 3 and 4). The type and quality of the evaluation design were also found to be related to effect size, with randomized designs and high quality evaluations being associated with higher effect sizes.

Multiple regression analyses were used to evaluate the separate and collective contribution of various program and evaluation characteristics on effect size. Disregarding the influence of content area on effect size, the factor which explained the largest proportion of the variance in the observed effect size distribution was the type of outcome measure used. Other factors found to contribute at least 1% to the multiple R^2 are listed in Table 5.

A secondary purpose of this study was to examine the ways in which submittals which were approved by the JDRP differed from those which were not approved. During the four years covered by this study, 62% of the programs applying to the JDRP were approved. The difference in mean effect size between approved and not-approved programs was not very large: 0.92 vs. 0.83, respectively. However, when programs which received unanimous decisions for approval or rejection were compared, a large difference in effect size was observed: .91 vs. .34, respectively. When programs were further categorized according to content area, in eight out of 13 areas where comparisons were possible, approved programs had higher mean effect sizes than not-approved programs (see Table 6). (However, it is interesting to note that the single variable which best discriminated between submittals which were approved or not approved by the JDRP was unrelated to effect size, program characteristics, or evaluation characteristics: it was a subjective rating of the overall quality of the written submittal--how easy it was to read and understand.)

If one were to compose a profile of the most successful programs reviewed for this study, in terms of magnitude of effect sizes produced, the following features would be included:

1. content area of natural science, general cognitive skills, motor skills, social science, or health/physical education (all had effect sizes

greater than 1.00)

2. target audience of gifted students
3. regular classroom setting
4. urban or suburban location
5. objectives related to behavioral changes
6. annual cost per student less than $84 and annual operating funds less than $100,000.

Programs for which the lowest effect sizes were obtained, on the other hand, had the following features:

1. content area of reading, language arts, or math
2. target audience of handicapped students
3. special facilities required
4. rural location
5. objectives related to attitudinal or affective change
6. annual cost per student greater than $84 and annual operating funds of $100,000 or more.

Characteristics of the methods used to evaluate the educational programs were also found to be differentially related to effect sizes. The most important of these in explaining the variation in effect sizes was the type of instrument used to measure program outcomes; locally developed instruments were associated with higher effect sizes than published tests. A profile of evaluation features typical of programs with higher effect sizes would include the following:

1. locally developed outcome measure
2. independent evaluator

3. randomized evaluation design
4. evaluation design rated as high quality
5. zero or one problem in data analysis
6. effect size calculated according to the basic effect size formula.

Evaluations of programs which had lower effect sizes tended to have the following characteristics:

1. published tests used as outcome measures
2. combination of staff and independent evaluators
3. norm-referenced evaluation design
4. evaluation design rated as low quality
5. two or more problems in data analysis
6. effect size calculated by use of a variation of the basic formula.

Finally, when program and evaluation features were considered jointly, and disregarding content area, the factors which in combination best explained the differences in effect sizes were, in order of importance:

1. type of measuring instrument used
2. type of formula used to calculate effect size
3. presence of an attitudinal objective
4. presence of a behavioral objective
5. evaluator affiliation.

For these five factors, higher effect sizes were associated with the use of a locally developed instrument, use of the basic effect size formula, program objectives related to

changes in behavior, and the presencr of an independent evaluator. Conversely, lower than average effect sizes were observed when program objectives involved changes in attitudes or affect.

Discussion

The results of this study have practical as well as theoretical implications. In the context of the JDRP, knowledge of the typical effect sizes achieved in the various content areas could assist Panel members in judging whether a program has produced a change that is "large enough" to be considered exemplary. For example, an effect size of .75 might be considered large for a reading program (since mean ES for reading = .56), while it would be considered small for a natural science program (mean ES = 1.32). Such use of effect size data could enhance the JDRP's decision-making process by allowing systematic comparisons to be made between current outcomes and outcomes typically achieved in a particular content area. In fact, these types of judgments are now made implicitly by the JDRP. The deliberate consideration of mean effect sizes would make this aspect of the assessment of educational significance more explicit, without compromising the Panel's capability of taking into account the many other factors which affect their decisions to approve or not approve particular programs.

Beyond the identification of differences in effect sizes for different content areas, this study found that certain

characteristics of educational programs and evaluations cut across the variety of content areas to influence outcomes (a potential contribution of meta-analysis noted by Pillemer and Light, 1980). For example, the use of locally developed instruments was found to be associated with higher effect sizes than the use of published tests. Because published tests are designed for maximum applicability across a wide range of educational experiences, they are more effective at measuring general achievements than specific learnings (Ball, 1981). As the match between the measuring instrument and specific program outcomes improves, other things being equal, the size of effects detected will increase - even if the programs being compared are "in reality" equally effective.

Other features which in this study were shown to be related to effect size were the type and quality of the evaluation design used. This finding has implications for the interpretation of individual programs' effect sizes. Recall that the magnitude of an effect size is dependent on two factors: the difference between the treatment and comparison groups, and the amount of variance that exists within the study. To the extent that the evaluator can reduce extraneous variance through increased precision of measuring instruments, or more careful planning and implementing of the evaluation design, the size of the effect detected will increase, other things being equal (Hall, 1980; Sechrest & Yeaton, 1982). While these features are certainly desirable for all

evaluation research, when they do not exist consistently throughout a collection of studies, the interpretation of effect size for any given program is confounded with the quality of the evaluation design.

Several useful products could be developed from the results of this study's analyses. For example, the effect sizes which were calculated could be assembled into what Cline (1976, p. 374) has called "a directory of effects...[for] the educational consumer." These data could be included in the JDRP publication of Educational Programs That Work, a catalogue of one-page summaries of all the educational programs that have been approved as effective by the JDRP. This would give potential adopters additional information to consider when evaluating the merits of installing one program as opposed to another. For example, if Program A has demonstrated a larger effect size than Program B, but is comparable in other respects, then Program A could be adopted. Alternatively, if Programs C and D produced approximately equivalent effect sizes, then the decision could rest strictly on cost or ease of implementation or attractiveness of materials or whatever happens to be important to the consumer.

Besides producing a list of effect sizes achieved, this study began to provide some answers to a question about current practice in education, which was posed by Chelimsky (1978, p. 16): "In those places where we can find demonstrated results, what are the common elements?" Larger

effect sizes were found in certain content areas (e.g., natural science), for certain subgroups of students (e.g., gifted), in certain settings (the regular classroom), and in certain locations (urban and suburban). More detailed analyses of educational programs represented by these subgroups--along with systematic comparisons of their features with those of the least effective programs--might further our understanding of why some educational programs are so much more effective than others.

Conclusions

This study used a sample of programs applying for JDRP approval to identify characteristics of educational programs and evaluations which were differentially related to effect size. The calculation of effect sizes for educational programs has some clear advantages for assessing program outcomes. The chief one of these is that, through statistical techniques, a standard metric for the size of effect produced is generated, thereby allowing comparisons among programs which have used diverse outcome measures.

The most important lesson to be learned from the results of this study, however, is that effect size is a statistic which should not be interpreted simplistically. Because effect size <u>is</u> differentially related to various characteristics of programs and evaluations, facile comparisons of the absolute values of effect sizes can be misleading. Meaningful interpretations of effect size require

careful consideration of the influences on effect size which are not causally related to program processes, in particular, any features of the evaluation which work to affect the amount of variance in the study.

Programs which apply to the JDRP for validation represent some of the finest efforts being made in education in the United States today. Continued systematic study of these programs will contribute to our understanding of what makes educational programs effective.

References

Ball, S. (1981). Outcomes, the size of the impacts, and program evaluation. New Directions for Program Evaluation, 9, 71-86.

Chelimsky, E. (1978). Differing perspectives of evaluation. New Directions for Program Evaluation, 2, 1-18.

Cline, M. G. (1976). The "what" without the "why" or evaluation without policy relevance. In C. C. Abt (Ed.), The evaluation of social programs (pp. 367-374). Beverly Hills, CA: Sage.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Hall, J. A. (1980). Gender differences in nonverbal communication skills. New Directions for Methodology of Social and Behavioral Sciences, 5, 63-77.

Pillemer, D. B., & Light, R. J. (1980). Synthesizing outcomes: How to use research evidence from many studies. Harvard Educational Review, 50, 176-195.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. Evaluation Review, 6, 579-600.

Table 7

Mean Effect Size by Program Characteristics

Program Characteristic	n	M	SD
Year of Review			
1980	46	.93	.64
1981	69	.72	.64
1982	79	.90	.67
1983	68	1.05	1.14
Objective Type			
Cognitive	226	.88	.77
Behavioral	16	1.21	1.36
Attitudinal	20	.68	.64
Content Area			
Math	58	.64	.45
Reading	55	.56	.41
Career education	24	.95	1.14
Natural science	20	1.32	.76
General cognitive skills	20	1.26	1.05
Language arts	19	.66	.56
Social science	18	1.15	.76
Health/physical education	10	1.13	.56
Motor skills	10	1.22	1.63
Other	28	1.19	.74

Table 7 (continued)

Program Characteristic	n	M	SD
Target Audience			
Students (regular)	197	.87	.75
Handicapped students	37	.67	.98
Gifted students	6	1.63	.90
Teachers	12	1.23	.92
Other	10	1.13	.88
Classroom Type			
Regular	199	.94	.86
Special Facility	61	.72	.62
Setting			
Urban only	64	.94	.88
Rural only	53	.75	.67
Suburban only	26	.93	.89
Combination	79	.94	.84
Annual Operating Funds			
Less than $100,000	123	1.00	.83
$100,000 or more	108	.72	.62

Table 2

Amount of Variance in Effect Size Distribution Explained by Program Characteristics

Program Characteristic	n	R^2
Objective type	262	.01
Content area	262	.19
Target audience	262	.04
Classroom type	260	.01
Setting	231	.02
Operating funds	231	.05
Installation cost	213	.01
Continuation cost	213	.01

Table 3

Mean Effect Size by Evaluation Design Characteristics

Design Characteristic	n	M	SD
Evaluator			
Independent only	137	.99	.91
Staff only	18	.91	.67
Combination	39	.77	.75
Instrument Type			
Published	135	.67	.53
Locally developed	93	1.25	1.05
Other	22	.80	.54
Design Type			
Norm-referenced	56	.59	.34
Quasi-experimental	162	.92	.85
Experimental	40	1.13	1.01
Design Quality			
Low/very low	32	.67	.67
Medium	84	.89	.92
High/very high	146	.93	.78
Data Analysis Problems			
0 or 1	215	.93	.85
2 or 3	47	.68	.56
Effect Size Formula			
Regular	187	1.02	.91
Other	75	.58	.29

Table 4

Amount of Variance in Effect Size Distribution Explained by Evaluation Design Characteristics

Design Characteristic	n	R^2
Evaluator type	194	.01
Instrument type	250	.12
Design type	258	.04
Design quality	262	.01
Data analysis quality	262	.02
Effect size formula	262	.08

Table 5

Results of Stepwise Regression of

Selected Variables on Effect Size

Step	Characteristic Entered	R^2	Increase in R^2	Direction of Influence
<u>Without Cost Variables Included</u>				
1	Local instrument	.113	.113	+
2	Basic ES formula	.137	.024	+
3	Attitudinal objective	.152	.015	-
4	Behavioral objective	.163	.011	+
5	Independent evaluator	.173	.010	+
<u>With Cost Variables Included</u>				
1	Local instrument	.143	.143	+
2	Attitudinal objective	.159	.016	-
3	Basic ES formula	.173	.014	+

Table 6

Mean Effect Size by Content Area and JDRP Decision

Content Area	Approved		Not Approved	
	n	$\overline{ES}$	n	$\overline{ES}$
Reading	40	.58	15	.50
Math	37	.74	21	.44
Career education	10	1.32	14	.68
Natural science	11	1.33	9	1.31
Social science	14	1.20	4	1.00
Computer literacy	3	2.19	1	.56
Writing	3	.62	2	.58
Special education	2	1.20	2	1.34
Language arts	15	.64	4	.73
Health/physical ed.	7	1.08	3	1.26
Motor skills	8	.74	2	3.15
Gen. cognitive skills	11	1.20	9	1.34
Other	5	1.50	2	1.26

Appendix A

JDRP Submittal #:_____
JDRP Session Date:_____
Coder's Initials:_____

JDRP SUBMITTAL ANALYSIS FORM

Program Title/Location:_____
_____

I. PROGRAM DESCRIPTION AND BACKGROUND

1. Content area for which claims of effectiveness are being made (check all that apply)

___arts and humanities
___basic skills
___reading ___writing ___math
___bilingual education
___career education
___vocational education
___teacher education
___other (specify):_____
___No information
___health/physical education
___migrant education
___natural science
___social science
___special education
___gifted education
___Cannot tell

2. Target audience for which claims of effectiveness are made (check all that apply)

___students
___teachers
___other (specify):_____
___No information
___administrators
___adult learners
___Cannot tell

3. Educational level of target audience (check all that apply)

___preschool
___K - 12: specify grade level(s)_____
___community college
___graduate school
___inservice/staff development
___other (specify):_____
___No information
___four-year institution
___adult education
___continuing education
___Cannot tell

4. Years of intervention's existence

_____ to _____
month, year month, year

___other_____
___No information
___Cannot tell

JDRP No._____

5. Funding source (check all that apply)

___ESEA Title I
___ESEA Title III
___ESEA Title IVC
___Vocational Education Act
___state funds
___local school funds
___private individuals/institutions/foundations
___other (specify):_____
___No information
___Cannot tell

6. Average annual operating costs

___<$50,000
___$ 50,000 - 99,999
___$100,000 -149,999
___$150,000 - 199,999
___$200,000 - 250,000
___>$250,000
___No information
___Cannot tell

7. Cost/student (round to nearest dollar amount)

$_____for installation
$_____for continuation
$_____installation/continuation not specified
___No information
___Cannot tell

8. Evaluator's affiliation (check all that apply)

___on staff of program
___consultant
___academic
___research firm
___other (specify):_____
___No information
___Cannot tell

9. Types of objectives for which claims of effectiveness are made (check all that apply)

___cognitive
___behavioral
___attitudinal
___other
___No information
___Cannot tell

10. Setting for intervention (check all that apply)

___regular classroom
___special facility
___other (specify):_____
___No information
___Cannot tell

11. Location of intervention (check all that apply)

___urban
___suburban
___rural
___No information
___Cannot tell

12. Duration of intervention (specify #)

_____
___No information
___Cannot tell

JDRP No._____

II. EVALUATION

A. MEASUREMENT FEATURES

[Use separate sheets and complete sections IIA, IIB, IIC for each objective that the program addresses.]

13. Briefly specify objective:_____
_____
Type:___cognitive ___behavioral ___attitudinal ___other

[Supply the following information about each instrument used to measure accomplishment of the specified objective. Use additional forms if necessary.]

14. Name of instrument_____

15. Derivation (check one)

___published test
___measure developed for this project specifically
___measure adopted from another source but modified
___other (specify):_____
___No information ___Cannot tell

16. Administration

For norm-referenced tests (check one):
___(some) testing not done at empirical norming dates
___testing done at norming times
___No information ___Cannot tell ___Not applicable

For treatment-comparison groups (check one):
___groups tested at widely differing times
___groups tested at/near the same time
___No information ___Cannot tell ___Not applicable

17. Validity data provided (check all that apply)

___face ___content ___construct ___criterion
___other (specify):_____
___No information ___Cannot tell

18. Reliability data provided (check all that apply and supply values)

___stability_____ ___equivalence_____
___interrater_____
___internal consistency (specify type)_____
___other (specify)_____
___No information ___Cannot tell

JDRP No._____

Objective_____

B. DESIGN FEATURES

[Complete this section separately for each objective addressed unless the same design(s) applied to all objectives. Numbers in parentheses refer to Campbell and Stanley's (1963) tables of designs.]

19. Type(s) evaluation design used (check all that apply)

<u>One Group: No Control or Comparison Group Established</u>
___post only (1) ___pre-post (2) ___time series (7)
___comparison against goals (criterion-referenced)
___other_____

<u>More Than One Group: Non-randomized Assignment to Groups</u>

___untreated or ___alternate treatment comparison
___post only (3) ___pre-post (10)
___national or __local norms
___multiple time series (14)
___regression discontinuity (16)
___other_____

<u>More Than One Group: Randomized Assignment to Groups</u>
___untreated or ___alternate treatment comparison
___post only (6) ___pre-post (4)
___multiple time series (14)
___other_____

___ <u>Qualitative Design</u> (specify):
_____
_____

___Other_____
___No information ___Cannot tell

20. Features affecting the comparability of the treatment (T) and comparison (C) groups (check all that apply)

___participants volunteered ___instructors volunteered
___intact group chosen because of similarity to treatment group
___pretest scores of T and C groups significantly different
___demographic characteristics for T and C groups dissimilar (e.g., SES, age, sex, race, school size)
___other_____
___No information ___Cannot tell ___Not applicable

JDRP No._____

Objective_____

21. Review the list of threats to internal validity (see Appendix B), and Campbell and Stanley's tables of designs. This study's degree of internal validity is (check one)

___Very High: with a reasonable degree of certainty, all of the applicable threats to internal validity can be ruled out
___High: with a reasonable degree of certainty, all but 1 or 2 of the applicable threats can be ruled out, and these are not "fatal flaws"; the evidence that the program caused the observed results is believable
___Medium: at least half of the applicable threats can be ruled out, and there are no "fatal flaws"; the evidence is ambiguous--neither totally convincing nor totally unconvincing
___Low: fewer than half of the applicable threats can be ruled out but there are no "fatal flaws"; the evidence is not very convincing
___Very Low: at least one threat is a "fatal flaw", i.e., is a compelling and plausible rival explanation for the observed results; the evidence that the program caused the observed results is not at all convincing

22. Are the program components clearly described?

___all are ___most are ___some are
___No info ___Cannot tell

23. Means of monitoring program implementation (check all that apply)

___instructor self-monitors
___program staff monitors
___directly ___ by reviewing instructor's records
___No information ___Cannot tell

24. Was evidence supplied to indicate that the intervention's effects were replicated? Put a check mark in all categories where non-aggregated data were presented and an 'A' for aggregated data.

___instructors ___classrooms ___grade levels
___schools ___settings ___time periods
___other (specify)_____
___No information ___Cannot tell

JDRP No._____

C. DATA ANALYSIS FEATURES

25. Years for which evaluation data were provided:
_____ to _____
month, year month, year
___No information ___Cannot tell

26. Data analysis procedures used (check all that apply)

___descriptive statistics ___zero-order correlations
___t-test ___ANOVA ___ANCOVA ___multiple regression
___nonparametric statistics ___ARIMA time series
___content analysis ___qualitative analysis
___other:_____
___No information ___Cannot tell

27. Features of data analysis procedures or presentation affecting statistical conclusion validity (check all that apply) *[Review threats and A Guide for Selecting Statistical Techniques.]*

___inappropriate or inadequate analysis procedures (specify)_____
___omission of some relevant outcome data
___omission of information about analysis procedures used (e.g., name of statistical test used)
___other_____
___No information ___Cannot tell

III. REPORT

28. Clarity: Number of "Cannot tell" responses_____
Total number of items_____

29. Completeness: Number of "No information" or "omission" responses_____
Total number of items_____

30. Overall quality of report (check one)

___Good: easy to read and understand; well organized; to the point
___Fair: average in readability and presentation
___Poor: difficult to read and understand; disorganized; irrelevant information presented; relevant information lacking

31. JDRP vote:____approved ____not approved ____abstained

Comments:

JDRP No.____

Objective_______

32. Information for calculation of effect sizes (ES):

[Use additional sheets if necessary.]

Instrument:____________________

Group	N	Mean	SD	Stat. Sig.	ES

Instrument:____________________

Group	N	Mean	SD	Stat. Sig.	ES

Instrument:____________________

Group	N	Mean	SD	Stat. Sig.	ES

Instrument:____________________

Group	N	Mean	SD	Stat. Sig.	ES

Appendix B

Supplemental Instructions for Completing the JDRP Submittal Analysis Form

1. Content area. Mark only those categories for which claims of effectiveness are being made.

Marine, ecology, environmental education = natural science.
Nutrition, family life, cancer education = health/physical.
Law, consumer education = social science.
Teacher education = preservice only (inservice and staff development have a separate category).

2. & 3. Target audience & educational level. Mark only those categories for which claims of effectiveness are made, even if the program narrative refers to other categories. If claims are not stated specifically enough, review data presentation and mark only those categories for which data were provided. Do not include grades which represent follow-up data only, unless specific claims were made regarding the long-term effects of the intervention.

4. Years of intervention's existence. Do not necessarily record what the submittal reports under Years of Development: check submittal for dates indicating that the program existed before the reported beginning date or following the reported ending date.

6. Average annual operating funds. Mark "cannot tell" if a total amount is reported, but it is not clear if the figure is a total across the years or for one year only.

7. Cost/student. Record dollar amounts as reported in narrative. If a total amount is reported, along with the total number of participants, calculate cost by dividing total amount by total number of participants.

8. Evaluator affiliation. Do not record as evaluator those whose sole tasks were to administer tests or review tests for adequacy. Do record those who designed evaluations, constructed tests, analyzed data, and wrote evaluation reports.

9. Types of objectives. Mark only educational objectives for which claims of effectiveness are made, even if the program narrative refers to other types of objectives. Mark "other" for artistic, physiological or other types of objectives which defy categorization.

10. Setting. If the program is a pull-out program

(requires student's removal from regular classroom), mark "other."

11. Location. Mark only those locations for which data have been reported and claims of effectiveness are being made. Do not check if claims were made that program effects were replicated in different settings, but data were not provided in support of the claim.

12. Duration of intervention. Specify amount of time (in hours, minutes, etc.) per day, week, etc. that participants spend on program activities.

14. Name of instrument. Report only on instruments measuring outcomes for which claims of effectiveness have been made. Do not report on tests used for selection or to establish comparability of groups.

15. Derivation. Mark "other" for tests such as state-mandated achievement tests which are not published or generally available.

16. Administration. For norm-referenced testing, mark "no information" only if no information at all was supplied regarding when the test was administered. Mark "cannot tell" if the test administration time was mentioned but it was not specified whether this corresponded to the empirical norming schedule. Mark "Not applicable" if normed tests were used, but the norm-reference evaluation model was not.

17. & 18. Validity and reliability. Check "unspecified type" if reliability and validity are referred to but specific types are not discussed. Report ranges if more than one value is given.

19. Evaluation design type. Record what the actual design was, not simply what the submittal termed it. Mark all statements which describe the design.

Mark "untreated comparison" when a comparison group received no program or received the traditional or regular curriculum. Mark "alternate treatment comparison" when two or more program approaches were being compared. Examples: Scores on a reading achievement test made by students in the reading program apppying for JDRP approval compared with scores made by students receiving the regular reading curriculum = untreated comparison. Scores on a test of knowledge of careers compared for students receiving a career education program and students receiving no systematic career information = untreated comparison. Math

achievement test scores of students receiving an innovative math program with or without instruction in the use of calculators compared against each other and against scores of students in traditional math classes = both untreated and alternate treatment comparisons.

20. Group comparability. If the design was norm-referenced or one-group only, mark "not applicable."

22. Program description. Program components are sometimes clearly identified but not clearly described. For this item, consider the descriptions.

23. Implementation monitoring. Routine observations made by supervisors should not be considered evidence of implementation monitoring unless it was specified that that is what the purpose was. If this was not specified, mark "cannot tell."

24. Replication evidence. If data were aggregated across elements in a category, mark 'A' for the category. For example, if 3 schools were involved in the program, but only one statistic (e.g., one mean) is reported, put an 'A' on the line before "schools." If data were not aggregated across elements in a category, put a check mark on the line next to the category. For example, if statistics were reported separately for grades 6 and 7 (e.g., a mean for grade 6 and a mean for grade 7), put a check on the line before "grade levels."

27. Data analysis features. Note use of gain scores and grade equivalent scores when these were the sole form of data presented. Note inappropriate uses of ANCOVA. The appropriate use of ANCOVA requires random assignment of subjects to treatment, or the strong presumption that nonrandom assignment is random in effect. It is not appropriate to use ANCOVA to provide statistical adjustments for differences between groups arising from the essential nonequivalence of the groups themselves. Both of these should be marked as "inappropriate or inadequate analysis procedures." In addition, if only gain scores were reported (without posttest scores), mark "omission of some relevant outcome data."

28. & 29. Clarity and comprehensiveness. For total number of items, count the numbers for which "cannot tell" and "no information" were possible responses. Note that "omission" responses in item 27 count as "no information" responses. Total number of items will differ according to the number of instruments and evaluation designs included in the submittal.

32. Effect sizes. Compute effect sizes using the formula: post treatment mean minus post comparison mean divided by the post standard deviation of the comparison group. Do not calculate effect sizes if only percentiles were reported. If NCEs were reported, use 21.06 as the standard deviation. If only adjusted posttest scores were reported, and standard deviations were given, use these. If only gains were reported, along with their standard deviations, use these. Make note of which formula was used (refer to Appendix C).