DOCUMENT RESUME

ED 252 460 SO 016 112

AUTHOR Shaver, James P.
TITLE The Law-Related Education Evaluation Project: A Methodological Critique of the "Impacts on Students" Findings.
PUB DATE 16 Apr 84
NOTE 21p.; Paper presented at the "LRE Evaluation Project: A Symposium on Methods, Findings, Conclusions," at the Rocky Mountain Regional Conference of the National Council for the Social Studies (Phoenix, AZ, April 11-14, 1984). For related documents, see ED 247 164 and SO 016 110-111.
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Delinquency; *Educational Assessment; *Legal Education; Outcomes of Education; *Reliability; Research Design; Research Methodology; *Research Problems; Research Reports; Secondary Education; Statistical Analysis; *Validity
IDENTIFIERS *Law Related Education

ABSTRACT

The major finding of the Law-Related Education Evaluation Project report for Year 1 (1981), that law-related education courses can reduce juvenile delinquency, is of limited use to educational decision makers and could be misleading. The research design leaves much to be desired; however, that fact must be considered in light of the difficulty of structuring educational research to meet the demands of experimental designs. A major disappointment is that the project assessed delinquency only through student self-reports of behavior. That assessment, without supportive indicators of delinquent behavior, vitiates the study's major finding. LRE will, under certain circumstnace, be associated with changes in student reports of delinquent behavior, but it is not clear if the reports validly represent actual behavior. The failure to deal with the importance of the results other than in terms of statistical significance or to report the information (correlation coefficients, means, and standard deviations) that readers could use in deciphering the results is also a major shortcoming of the report. An analysis of the second year LRE Evaluation Project, which was supposed to provide methodological improvements as well as replication of data relevant to the first year's results, shows that it contains many of the same methodological shortcomings as the first study. Educators are urged to be cautious about relying upon these reports to advocate law-related education. (RM)

THE LAW-RELATED EDUCATION EVALUATION PROJECT: A METHODOLOGICAL CRITIQUE OF THE "IMPACTS ON STUDENTS" FINDINGS*

James P. Shaver
Utah State University

The first findings from the Law-related Education Evaluation Project, funded by the National Institute for Juvenile Justice and Delinquency Prevention, became available in December 1981 (Hunter & Turner, 1981). Soon afterwards, based upon that report, claims about the efficacy of law-related education in the reduction of juvenile delinquency began to appear in LRE publications--as illustrated by the following headlines: Street Law News (Spring 1982), "Law-related Education Emerges as a Useful Tool to Deter Delinquency"; LRE Report (Winter 1982), "Study Indicates That LRE Can Reduce Juvenile Delinquency"; and, the LRE Project Exchange (Winter 1982), "Two-year Study Indicates that LRE Can Reduce Juvenile Delinquency". The optimism of those articles flowed directly from the LRE Evaluation Project report.

Would the same claims have been made if the report had been accompanied by a careful methodological critique or if claims in the report had been couched in language appropriate to its methodological shortcomings? And, does the report of the second year of the LRE Evaluation Project provide valid support for the first year conclusions? These questions provided the setting for this paper.

Methodological Concerns With the Year One LRE Evaluation Project

Educational research, particularly as applied to the evaluation of curricular programs such as law-related education, is no easy endeavor. The difficulties of arriving at firm empirical results when faced with the problems of arranging for research in real-life settings which make laboratory controls impossible have long been a concern among educational researchers, among those who lament the lack of accumulative knowledge from educational research (see, e.g., Kerlinger, 1977; Shaver 1979), and among those who anguish over research results and evaluation reports as they attempt to make decisions about practice in the schools. Clearly, then, it would be unrealistic to expect perfectly valid research in an effort to evaluate law-related education. For that reason, the focus in the paragraphs that follow will be less on ways in which the

*Paper prepared for presentation at a symposium, The LRE Evaluation Project: Methods, Findings, Conclusions, at the Rocky Mountain Regional Conference of the National Council for the Social Studies, Phoenix, April 12, 1984.

research methodology could have been improved than on how the research was interpreted and presented. The critique is not meant to be all inclusive. Rather, the intent is to indicate some major concerns in regard to the report and the conclusions drawn prematurely, I believe, from it.

The Assessment of Juvenile Delinquency

The LRE Evaluation Project has emphasized juvenile delinquency as its major dependent variable (i.e., the major variable upon which the effects of LRE were to be assessed). Although that focus may seem questionable to some educators, it does make sense in light of the funding of the project by the National Institute for Juvenile Justice and Delinquency Prevention, U.S. Department of Justice. In addition, as Hunter and Turner (1981) pointed out, juvenile delinquency is relevant to citizenship education because "law abiding and delinquent behavior are positive and negative indicators of citizenship" (p. iii). Of greater concern to a methodological critique of the project is how delinquency was assessed and the way in which that variable is referred to throughout the report.

The LRE Evaluation Project assessed "delinquency" only through student self-reports of behavior. Educational and psychological researchers tend to be dubious of self-reports, especially if the responses have social desirability attached to them and there is either a lack of substantial evidence of validity or a lack of independent data consistent with the self-report results.

In sociology, self-reports of delinquent behavior have become a standard assessment technique, in large part because police and court records have often put ethnic minorities and those from lower socio-economic strata in an unjustifiably poor light because their delinquent acts tend to place them more often in contact with the police and are more likely to be recorded in police records and result in court action. Such records have been used as the basis for discriminatory conclusions about the relative prevalence of crime among ethnic minorities and those from lower socio-economic classes. Incidence surveys and self-reports were introduced as corrective measures; and the results with them have indicated that whites and those from higher socio-economic classes are more often involved in crimes, even though perhaps different types of crimes, than the official records indicated.

Despite their widespread use, the validity of self-reports has been (Clark & Tifft, 1966), and continues to be (Empey, 1982, pp. 122-124; Jensen and Rojek, 1980, pp. 90-96), a concern among sociologists. The preferred methodology is to use a combination of two or more types of assessments of delinquent behavior--perhaps self-reports, police and court records, and/or incidence reports--to see if, in the sociologist's terminology, they "triangulate", that is, converge on the same conclusion. Educational researchers would tend to treat the matter as one of using data from other sources to establish the validity of the self-reports.

The use of self-reports of juvenile delinquency is of particular concern in an educational study, such as the LRE Evaluation Project, where both social desirability effects and experimenter effects are likely. "Social desirability" refers to the tendency to give responses deemed to be socially desirable, especially if the respondents believe that confidentiality may be breached. "Experimenter effects" refers to the tendency of those who administer experimental treatments (in this case, the teachers) to convey their expectations, implicitly or explicitly, to the subjects (here the students) and thus affect the subjects' responses. If teacher-student rapport is good, the students may respond in ways to support the teacher's expectations; if rapport is not good, subjects may respond contrary to what they perceive as the experimenter's desires.

The experimenter effect variable is, for example, an alternative explanation for instances of increased reported delinquent behavior among students in LRE classes. As the report (Hunter & Turner, 1981) indicates, based on the 11 types of offenses that were assessed:

> The predominant result in four LRE classes . . . was a reduction in delinquency [sic], compared with the control classes. The result in three classrooms . . . was a pronounced increase in delinquent behavior [sic] among LRE students, while the remaining three classrooms . . . showed a slight increase or no change. (p. 14).

The term sic (an indication that a quoted passage is reproduced precisely) is used in the above quotes to indicate that, lacking substantiating data, there must be serious concern about the extent to which actual delinquent behavior was assessed by students' self-reports of that behavior. The report itself should have been couched in terms of "student self-reports of delinquent behavior". That would not only have been consistent with the data obtained, but would have encouraged readers of the report, especially those who wanted to interpret the findings for school people, to use properly qualified terminology in referring to the evaluation results.

Conclusions About Causality

As mentioned above, designing applied evaluation studies to be conducted in the field is no easy task. There is no pretense in the LRE Evaluation Project's first year report that a perfect design was set up. The design is quasi-experimental, as students or teachers were not assigned randomly to treatments. The experimental and control groups were intact classes, with the experimental groups selected purposely to yield information relevant to the research questions (Hunter & Turner, 1981, p. 5). How the control classes were selected is not clear in the report, but a reasonable presumption is that they were selected by school administrators, as was the case in the second year of research. Selection is, of course, a matter of concern, because the factors involved might rival LRE as alternative explanations for any results.

To make cause and effect claims from such a design is most questionable. Yet, throughout the report reference is made to "impact"

(for example, to "impact findings", p. 13), variables which were "affected favorably and unfavorably" (e.g., p. 14), and "affects on students' skills" (e.g., p. 22).

In addition, correlational data were interpreted causally. There is, for example, the following discussion:

> Knowledge gained in a structured law-related education class using any of the three curricula involved in this evaluation was significantly correlated (p=.05 or better) with a reduction in [reported] infractions of school rules, property offenses at school, violence against students, public disorder, and drinking . . . . If all else is equal, the greater the knowledge gain, the fewer delinquent acts committed. Since knowledge gained is not correlated significantly with any other factor used in this study as a predictor of delinquency, its affect on behavior appears to be direct. While knowledge gained in and of itself can have a favorable impact on behavior, unfavorable change in one or more of the other factors can offset the behavioral benefits of increased knowledge. (p. 13)

It is a common dictum in statistics that correlation does not establish cause and effect. There is, of course, the possibility that a variable which has not been assessed bears a common relationship to the variables for which a correlation has been obtained, thereby accounting for the relationship. It could, for example, be that students who have higher levels of I.Q. not only learn LRE content better but become less likely to report delinquent behavior. Moreover, even if one were to assume that causality underlay a correlation coefficient, there is no information in the coefficient about the direction of influence or causality. Rather than increases in knowledge causing a reduction in reports of delinquent behavior, it could be that those who reduce their delinquent behavior have more motivation or time to study, thereby increasing their knowledge--a causal relationship that runs opposite to the interpretation in the report.

What Effects? In any event, it appears that LRE had little effect. Because of the lack of random assignment of students to treatment groups, it could be argued that the appropriate unit of analysis was the classroom and not individual students. Interestingly, the report indicates:

> Based on a count of the number of dimensions affected [sic] favorably and unfavorably, five of the ten LRE classrooms . . . showed a net improvement in knowledge and predictor variables relative to the respective controls. Three classes . . . displayed a net deterioration in the same dimension, and two . . . showed slight improvement or no net change. (p. 14)

It would be plausible to interpret a result in which 5 observations were positive, 3 negative, and 2 showed no improvement, to be about what one would expect by chance. That interpretation can be verified intuitively by thinking of it as a coin toss problem--i.e., consider a gain to be a head and a no gain or loss to be a tail. Five out of 10

would be the most likely chance occurrence. A consistent result is obtained if one is somewhat more discriminating and considers three events: A positive result, a negative result, and no difference. The probability of a 5-3-2 occurrence, as a departure from the even split among the three possibilities that would be expected if there was "no effect", is .50, using chi-squared. That, again, is well within what one would expect by chance.

The report also indicates (p. 14), as noted above, that, in regard to "the 11 types of offenses", compared to control classes there was "a reduction in delinquency [sic]" in only four out of ten LRE classrooms, an increase in reported delinquency in three of the ten LRE classrooms, and no increase or only a slight increase (apparently meaning, "not statistically significant") in three LRE classrooms. Put more strongly, in six out of ten comparisons the control classes' reports of delinquent behavior were equal to or better than those of LRE classes. Again, these patterns (4-3-3; or, 4-6) are about what one would expect by chance. From this perspective, the picture of LRE "impact" is not strong.

The Use of Statistical Significance

The heavy reliance on tests of statistical significance in the report is bothersome for two reasons—one having to do with assumptions about causality, the other having to do with the educational significance or importance of the results.

There is a tendency in the report to rely on statistical significance as an indicator of the "impact" of LRE. Aside from the questions raised above, the reader who lacks a statistical background in statistical inference needs to be cautioned that tests of statistical significance do not address questions of causality. A test of statistical significance answers only one question, how likely is it that a result could have occurred by chance? That is, in the case of comparing means, if one were to draw two random samples from the same population*, how likely is the difference in means that was obtained? The result of the test of significance, therefore, is no more than a statement of probability. If a result is statistically significant, e.g., at the .05 level, that means only that if one drew random samples under the null hypothesis (i.e., assuming that the samples were coming from the same population), this result would occur by chance five times or fewer out of 100.

Clearly, then, a test of significance does not speak, even indirectly, to the question of what, other than chance, might have caused a difference between groups.

Note, too, that a test of statistical significance does not tell us that a particular result (e.g., a difference between means or a correlation coefficient) is not a chance occurrence. Even if we drew random

---

*Or, alternatively, if the samples were drawn from two populations with equal means (technically, mu's).

samples from the same population and compared means, using the .05 level of statistical significance we would conclude five times in 100 that our results were not chance occurrences under the null hypothesis, clearly an erroneous decision (called by statisticians, a Type I error). A statistically significant result, then, does not tell a researcher what might have produced a result or whether a particular result is a chance occurrence or a "real" difference. (See, e.g., Carver, 1978, and Shaver, 1980, for further discussions of the limited implications of statistical significance.)

It is of particular interest that attaining statistical significance is directly related to the researcher's sample size. This makes conceptual sense because, roughly speaking, the larger one's sample, the more likely it is that statistics computed for it will approximate the values of the population. However, with very large samples, differences that are statistically significant may be educationally trivial. In the LRE Evaluation Project, for example, with ten experimental and ten control groups and assuming an average class size of 30, approximately 600 students were involved in the analyses. As an illustration of the effects of sample size on statistical significance vis-a-vis educational importance, consider that with a sample size of 600, a correlation coefficient of .08 would be statistically significant at the .05 level. The coefficient of determination (r^2) would be .0064, indicating the proportion of variance which the two variables have in common (.0064 multiplied by 100) is only .6%.

When pairs of means are being compared, an estimate of the amount of variance in scores on a dependent variable associated with membership in the treatment or control group can be obtained by computing a point biserial correlation coefficient and squaring it. If the difference between means is statistically significant, the point biserial coefficient will be, too. With a sample size of 600, an r_{pb} of .08 is also statistically significant at the .05 level. If one considers r_{pb}^2, a statistically significant result could be one in which less than one percent (.6%) of the variance in scores on a dependent variable is associated with membership in the treatment or control group—hardly likely to excite one as an educationally significant result.

Another way of construing the same information is in terms of Effect Size. An r_{pb} of .10 (statistically significant with N=600) would correspond to an Effect Size of about .2 (Cohen, 1977, p. 22). That means that if you subtracted the control group mean from the experimental group mean and divided by the control group standard deviation, you would obtain a value (the Effect Size) of .2. This Effect Size can be interpreted, using the values of the normal curve, as indicating that the mean of the treatment group (assuming that it is the higher one) exceeds the scores of 58 percent of those in the control group. Of course, if there were no difference between the means, you would expect the treatment group mean to exceed 50 percent of the scores in the control group. Again, the Effect Size for that statistically significant result would not likely be deemed an indication of educational importance. (For an introduction to Effect Size, see Borg & Gall, 1983.)

This elaboration is made here because the relationship between statistical significance and educational significance or importance is rarely discussed in statistics courses or in the educational research literature. Yet, it is important to keep in mind that, with the sample size for the LRE Evaluation Project, statistical significance can occur with trivial results.

Unfortunately, and this may seem anticlimactic after the rather extended discussion above, the report (Hunter & Turner, 1981) does not present correlation coefficients, means, or standard deviations, so the reader is unable to assess the educational importance of the statistically significant results. (Some of the correlation coefficients were sent to me. Those that were statistically significant for LRE knowledge gain and reported delinquent behavior ranged from .09 to .13, emphasizing the need for caution in interpreting the project's statistically significant results.) I am forced to the conclusion that the report is basically uninterpretable from the perspective of school people seeking information as to the effects of law-related education.

In a recent article in The Educational Researcher, Shapiro (1984) discussed the differences in philosophical underpinnings for research in econometrics and in ed-psychometrics that result in different emphases, for example, on internal versus external validity and in the interpretation of unexplained variance. Perhaps what we see in this report is a difference between the perspectives of educational evaluation and sociological research, with the latter more attuned to general trends and statistical significance. In that sense, the report reaffirms that while educational, psychological, and sociological research have much in common with educational evaluation, they are not completely corresponding domains. Educators are concerned with educational significance, and a report that relies on statistical significance as the only indicator of important findings is not useful.

From Findings to Recommendations

It will perhaps be evident from the discussion to this point why the following paragraphs from the report (Hunter & Turner, 1981), presented under the heading "Implementation", are of concern:

> The findings of the LRE classroom impact evaluation are clear in supporting the underlying theory of law-related education [sic]—supportive in showing that when LRE is implemented in accordance with prescriptions for the development of sound LRE programming, the classroom learning experience favorably affects factors which are directly related to socially approved behavior, namely those described in the introduction to this report: commitment, attachment, involvement, belief in the moral validity of social rules, equality of opportunity, and positive labeling. This in turn effects a reduction in the delinquent behavior of students exposed to the class. (p. 33)

The report goes on to make a curious statement:

> Delinquency [sic] was not reduced in every LRE class studied. Had that been the case, it would have been extremely difficult to isolate the critical features which appear to make a difference in the capacity of law-related education to have the effects sought in the impact research design. (p. 33)

In fact, why investigate the "critical features" of LRE when its effectiveness, in contrast with that of the control groups, had not been established? It would have made as much sense to study the critical features of the control groups, and more sense to look at critical features across control and LRE groups.

The questionnaire data had been combined with ethnographic data in an analysis that, according to the report, "revealed six features of LRE programs in the sites studied that seriously affected [sic] whether the law-related education class had a favorable impact [sic] on factors associated with delinquent behavior [sic] and, by extension, on delinquency" (p. 20). And the report states:

> The six features which appear to differentiate successful LRE classrooms from less successful or unsuccessful classrooms (by the criterion of delinquency reduction) translate into familiar prescriptions for successful implementation of LRE. The more of the features that are present, the more likely an LRE class is to have desired behavioral outcomes. (p. 33)

It is then recommended that a "prototype implementation model for LRE [be developed] which emphasizes and explains the necessity for the presence of all prescribed features" (p. 36). Aside from the questionable validity of drawing conclusions from a potentially invalid dependent variable, and from findings that are not sound indicators of causality and which may be statistically significant yet educationally unimportant, there has been a general reluctance in teaching methods research to go from correlational, quasi-experimental data to recommendations for change. For example, research findings that teacher enthusiasm is correlated with student learning in natural settings do not mean that when teachers are trained to be enthusiastic and/or try to be more enthusiastic, the same result will occur. There may be other correlates of enthusiasm that are not affected by training or by the conscious effort to be enthusiastic. Equally important, when we try to manipulate instruction, the results may be different from the effects of natural variation. The same is true for program settings. It seems highly likely, for example, that the effects of "active involvement of building level administrators" (one of the recommended features of successful LRE classrooms, p. 35) in a voluntary natural setting may not be reproduced if principals are forced into participating in an LRE program.

As noted in the prior section on Causality, the LRE Evaluation Project's findings hardly constitute the rousing endorsement of law-related education that is implied by the claims made in the report and in the headlines cited in the opening paragraph of this paper. The conclusion (Hunter & Turner, 1981, p. 33) that "the findings of the LRE

classroom impact evaluation are clear . . . in showing that when LRE is implemented in accordance with prescriptions for the development of sound LRE programming, the classroom learning experience favorably affects factors which are directly related to socially approved behavior . . . [which] in turn effects a reduction in the delinquent behavior of students exposed to the class", could just as well been stated: "Without law-related education curricula, teachers who properly implement accepted methods of instruction will favorably affect factors directly related to socially approved behavior, which will in turn effect a reduction in the delinquent behavior of their students". Indeed, the impact data suggest that the latter may be the more legitimate conclusion.

From this view, it is also important to put in appropriate perspective the claim (Hunter & Turner, 1981, p. iii) that, "the research shows that the knowledge gained in skill acquisition in this arena [LRE] affects directly the students' adherence to the law—a small but statistically significant effect". It would be more appropriate to refer to "a statistically significant effect which appears to be small and trivial" (which cannot be detected from the report itself because of the lack of adequate information).

Summary

To sum up, the Law-related Education Evaluation Project report for Year 1 is of limited use to educational decision-makers, and could be misleading. The design leaves much to be desired, but that must be considered in light of the difficulty of structuring large, geographically-spread educational research in which it is extremely difficult to meet the demands of experimental designs. Nevertheless, a major disappointment is the unjustified use of self-reports of delinquent behavior. That use, without supportive indicators of delinquent behavior, vitiates a major proported finding of the study. We know that LRE will, under certain circumstances, be associated with changes (positive or negative) in student reports of delinquent behavior, but we do not know if the reports validly represent actual behavior.

Unfortunately, the use of language in the report does not reflect this very serious limitation. That is, "juvenile delinquency" and "behavior" are referred to throughout without use of the important qualifier "self-reported". Such language is likely to mislead readers, as is the uncritical and unjustified use of terms such as "impact" and "effect" to refer to what are at best statistically significant associations.

The failure to deal with the importance of the results other than in terms of statistical significance, or to report the information (correlation coefficients, means, and standard deviations) that readers could use in deciphering the results, is also a major shortcoming of the report. Because very small correlation coefficients or mean differences would have been statistically significant with the size of the sample, the lack of sufficient information makes the report basically meaningless as a decision-making document.

In addition, there is the unflattering evidence for LRE when the results for the LRE and control classes are contrasted with the probabilities of such results occurring by chance. The translation of the findings as reported into recommendations for teachers or school districts is of dubious validity.

It is a cardinal rule of educational evaluation that reports should address questions of interest to the anticipated audience, with findings reported in a form that is understandable and interpretable by that audience. If the intended audience was public school educators, this report receives a low grade.

The Second Year Report

It might be argued that the 1981 report is not of interest, now that it has been succeeded by another report (SSEC-CAR, 1983). However, many people are still referring to the injudicious conclusions drawn from the first report in LRE publications. That is part of the reason for the elaboration of concerns above. That elaboration is also intended to serve as a context for discussing the report of the second year of research, which was intended to provide methodological improvements as well as replication data relevant to the first year's results. It is pertinent to ask at this point, how does the second year report stack up?

Methodology

A number of methodological questions about assessment and research design were raised above in regard to the 1981 LRE Evaluation Project report. Were those methodological shortcomings also present in the second year study?

Assessment. Considerable concern was expressed above about the use of students' self-reports of behavior as indicators of actual delinquent behavior. That same assessment procedure, without supportive evidence, was used in the second year research. And, terminology to make it clear that self-reports, rather than assessments of actual behavior, were used is also absent from the second year report.

In addition, in both years, changes in students' skills (including "those related to basic communication, such as writing, reading, speaking, and listening; analytic thinking skills, such as identifying alternatives, identifying consequences, and making decisions; and social skills, such as working cooperatively with others and relating to law and justice personnel", (Hunter & Turner, 1981, p. 22; SSEC-CAR, 1983, pp. 4-2 to 4-3) were assessed by asking LRE teachers to estimate program effects. The validity of gross, high inference ratings of specific behaviors has been seriously questioned by researchers. Moreover, such ratings are likely to be contaminated by the teacher's own volunteer and enthusiastic participation in the project. It would have been most surprising had the teachers indicated anything other than what they did--that is, that the program did have positive effects. Again, a finding in an important outcome area is obscured by the lack of adequate assessment.

Another aspect of assessment has to do with verification of the independent variable. In educational research, a basic question is whether the independent variable of differing instruction was actually implemented (see e.g., Shaver, 1983). One common approach to this problem, when addressed at all, is to gather observational data along dimensions of behavior considered crucial to the independent variable. The data are then analyzed to determine whether expected differences were present and the extent to which the lack of those differences—or, put differently, the amount of variability among experimental and/or control group teachers—may have invalidated the research.

Such verification of the independent variable was not undertaken in the first study, although classrooms were visited to obtain estimates of the likelihood that what was going on would build positive attitudes toward the law, increased attachments to the school, and favorable peer relationships (Hunter & Turner, 1981, p. 19). In the second year study, LRE classrooms were observed three times (SSEC-CAR, 1983, p. 3-7). However, control group classes were observed only once, and so were not included in the analysis of data (p. 5-2). In fact, inadequate attention to the reliability and validity of the observational data (see pp. 3-13, 3-14) make it dubious that such comparisons would have been worthwhile. Consequently, little is known about the differences and similarities in instruction in experimental and control group classes.

Correlational analyses were conducted to investigate what characteristics of LRE classrooms were associated with outcomes. This was a pilot effort, not a part of the original research proposal (SSEC-CAR, 1983, p. 1-8). Ratings on 11 classroom observational variables (based on class characteristics that were "thought to be relevant to the teaching of the law and to the reduction of delinquent behavior", p. 5-12) were correlated with scores on student outcome variables. The correlations varied greatly in size, ranging from .00 to .55 in size. There was an inconsistent pattern of correlations, and the coefficients did not in general confirm the predictions based on the theory of delinquency causation. Interestingly, several pages were spent in interpreting correlation coefficients of .19 or less between classroom observational variables and reported student behavior (pp. 5-45 through 5-51), although those results are perhaps best summed up by the report's own words, "One might reasonably say that little or nothing should be made of these associations, as most are weak" (p. 5-45).

In short, as is not uncommon in educational research (Shaver, 1983), we cannot be sure what treatments were compared or what variations among treatment and/or control groups might have affected the results. Indeed, given the findings to be discussed below, it may well be that a "non-study" occurred, in the sense that the independent variable was not implemented.

Design. Nearly every educational field study can be criticized on methodological grounds. The intent in this critique is not, therefore, to pick at details in order to discredit the research (especially as compared to other such research), but to provide readers, particularly those who are not well-trained in statistics and research design, with a properly circumspect perspective from which to view the findings and the recommendations based on them.

As already noted, the design of the first study was quasi-experimental; the selection of both experimental and control groups was based on conscious as well as implicit factors which might have affected the results. The same is true of the research for the second year, with one striking exception, to be noted shortly. The LRE classrooms in the national study came from sites at which LRE "could be tested under the most favorable possible conditions" (SSEC-CAR, 1983, p. 3-2). The schools selected were in districts which were willing to cooperate in the evaluation, which were already using the curriculum materials, which would send teachers to receive training, and in which there was already evidence of strong support for the program by the building administrators. Comparison classes were selected by the building administrators based on their estimates that those classes approximated the LRE classes in their schools (SSEC-CAR, 1983, pp. 4-3 to 4-4). Principals' judgments about such matters are not of unquestionable validity. And, it is commonly agreed that initial assignments of students to classrooms often involve, intentionally or not, factors that will be associated with later student performance. Keeping these selection factors in mind is important not only for evaluating the validity of the research findings, but in deciding on the extent to which they might apply to schools in general or to the school or school district with which one is personally concerned.

The second year study did include a junior high school in Colorado in which it was possible to randomly assign all 9th grade students to either LRE or conventional civics classes, with three sections of each and with the civics classes serving as control groups. The junior high school is noted in the report as having a strong record of training its teachers to use innovative teaching strategies and encouraging the use of those strategies in classrooms. To the project evaluators, then, the use of this school provided an opportunity "to assess the unique impact of LRE over and above the impact of superior instructional strategies" (SSEC-CAR, 1983, p. 4-4, underlining in original).

The use of such a school, of course, creates problems of external validity—that is, of the extent to which its results are generalizable to other schools. Strongly positive instructional environments and teachers with high levels of training and instructional competence are likely to be present in high socio-economic or otherwise educationally-oriented communities. And the students, too, are likely to reflect such community norms. Unfortunately, no data are reported that would allow the consideration of representativeness of the students in the Colorado school or the comparison of their characteristics with those of students in the 19 schools in California, North Carolina, Michigan, and Illinois which made up the national study. Nor is any information given about the community in which the school is located, although one might well suspect that it is suburban and well-to-do. The extent to which the findings from the Colorado site are generalizable to other schools is an important question that cannot be addressed because of the lack of data describing the sample.

Analysis of Data. As would be expected, at the sites other than the one in Colorado, the comparison classes were often not equivalent, at the beginning of the research, to the LRE classes on age, self-reports of delinquent behavior, and the other variables assessed in the research.

Typically, analysis of covariance is used in such situations, although theoretically it also calls for random assignment, to control for initial group differences. Multiple regression, which is based on the same least squares mathematical model, is an alternative approach (Cohen & Cohen, 1975, Ch. 9). In multiple regression, the variability in mean scores on the dependent variable due to initial group differences on the control variables is accounted for by first entering the control variables in the regression equation and then entering the treatment variable to determine whether the amount of additional variance associated with treatment group membership is statistically significant. It is this multiple regression approach which was the basic analysis for assessing treatment "effects" in the second year study.

Perplexingly, this approach is referred to as a "conservative estimate of effects" (SSEC-CAR, 1983, pp. 4-9 to 4-10). Reference is also made to a "soft" estimate, in which pretreatment differences are not considered in analyzing change scores—clearly an unacceptable approach. Strangely, the researchers combined the "conservative" and "soft" estimates, giving the results of the conservative estimates "twice the weight" of the results of the soft estimates. They say:

> We offer the combined estimates as an approximation of findings that a conservative estimate alone would have yielded had there been equivalence at time-1 between students enrolled in LRE classes and those enrolled in comparison classes (p. 4-10).

In fact, their conservative estimate is an estimate of the results that would have been obtained had there been time-1 equivalence (Cohen & Cohen, 1975, Ch. 9). The so-called conservative analysis of the findings is the one that merits attention, although, ironically, the combined estimate does not produce results which differ substantially.

Educational Importance of the Findings

As in the report of the first year's findings, the emphasis in the report of the second year's finding is on tests of statistical significance. One improvement is that correlation coefficients for pairs of variables are reported for the national study data. However, in the analyses of primary interest—that is, comparisons of the LRE and control groups—only means, B-weights (which indicate the amount and direction of change in raw scores on a particular dependent variable depending on LRE or control group membership), and statistical significance are reported. The correlation coefficients which are "normal by-products" of the multiple regression analysis, and "allow for a more deeply etched portrait of the phenomena under study" (Cohen & Cohen, 1975, p. 314), were not reported. Standard deviations for the comparison and control groups were not reported either. Consequently, educational significance could not be checked by computing Effect Sizes. If t-ratios for the B-weights had been provided, a "natural measure of effect size" (Cohen & Cohen, 1975, p. 348) could have been computed: coefficients indicating the proportion of the variance in a dependent variable, adjusted for initial differences, that was associated with LRE and control group membership.

If a number of statistically significant results had been found favoring LRE, this lack of information would have raised perplexing questions of interpretation. Given the mixed findings, which will be discussed in the next section, this lack of information is somewhat less perplexing. Rather than being the basis for puzzlement as to whether the differences favoring LRE were educationally important, it leads to puzzlement as to whether the results were even less impressive than they appear to be.

From Findings to Conclusions

The report for the second year of the LRE Evaluation Project (SSEC-CAR, 1983) concludes, in regard to "program impact on students", that:

> In sum, strong and defensible findings from the Colorado site indicate that LRE is capable of reducing delinquent behavior and favorably affecting most of the correlates of law-abiding behavior that were measured. The less persuasive, suggestive evidence from the national sites points to the same conclusion. (p. 4-50)

The Colorado Site. As noted above, it is difficult to interpret the results of the project in regard to LRE-control group outcome differences. Nevertheless, it is worthwhile to ask whether what information can be readily gleaned supports the report's optimistic conclusion. The findings from the Colorado site merit attention first, because of the quality of the design at that site.

Delinquent behaviors were again the central concern of the project. A summary of the "effects" at the Colorado site (SSEC-CAR, 1983, pp. 4-26 to 4-27) indicates that out of the 10 categories of reported delinquent behavior, there were statistically significant results for two of the 10 comparisons in the direction of the LRE students; for four out of the 10 comparisons, results in the right direction were not statistically significant; and, for four out of the 10, there was no discernible difference between the groups. In sum, only two of the 10 results were statistically significant.

Another set of measurements that are particularly relevant to law-related education fall under the category of "Belief" in the report--an incongruous title because each of the four variables actually deals with attitudes: Attitudes toward police, toward deviance, toward personal violence, and toward rationalizations that deviance may sometimes be all right. For these four variables (p. 4-25), there was statistical significance in the direction of LRE on one (attitudes toward police) and no statistical significance on the other three.

With these sets of findings, it hardly seems worthwhile to ask whether the statistically significant results were educationally important. In any event, it would be difficult to answer the question because standard deviations were not reported, although means were.

These results hardly seem to constitute "strong and defensible findings . . . that LRE is capable of reducing delinquent behavior and

favorably affecting . . . correlates of law-abiding behavior . . . ." School people will be interested to know, however, that there was (p. 4-28) a statistically significant gain in knowledge for the Colorado LRE groups as compared to the control groups, although because neither means or standard deviations were reported, it is, again, not possible to evaluate the educational significance of that result.

The National Study. What about the "less persuasive", but "suggestive", evidence from the national study sites. It is worth noting initially that with 27 LRE classes and 24 measures, 648 differences were tested, and only 12.4 percent (80) yielded a statistically significant result. This was "barely more than the 10 percent expected to occur by chance at the significance level chosen (.05, one-tailed = .10, two-tailed*)" (p. 4-30). Indeed, "on the longitudinal measures [that is, the pre-post assessments, which include reported delinquent behavior, "beliefs", and knowledge], there were 35 favorable and 45 unfavorable effects" (p. 4-30).

A summary table for the comparisons of the 27 LRE classes with their control classes (p. 4-31) yields additional information for the conservative estimate, which is the appropriate one. Focusing again on a central emphasis in the report's conclusions, with 10 categories of reported delinquent behavior and 27 LRE-control comparisons, there were 270 differences to be tested. Of these, 20 of the comparisons yielded statistically significant** results favoring the control groups, 240 yielded no statistically significant differences between LRE and control groups, and only 10 yielded statistically significant differences favoring the LRE groups (fewer than the 13 or 14 expected by chance at the .05 level of statistical significance). For the "beliefs" (actually, attitudes) variables, the results were similar. Out of 108 comparisons for the four variables, seven comparisons favored the control groups, 98 yielded no difference, and only three favored the LRE classes (fewer than the five expected by chance at the .05 level of statistical significance).

Even using the results of the unacceptable analysis in which "conservative" and "soft" analyses for the national sites were combined, the results are not strikingly positive. For the 270 comparisons involving the 10 categories of delinquent behavior across the 27 LRE-control group comparisons, 90 favored the control groups, 65 yielded no difference (a total of 155 comparisons which either favored the control groups or yielded no difference), and in only 115 instances was the difference in favor of the LRE groups. For the four "beliefs" (attitude) scales, out of 108 comparisons, 28 favored the control groups, 54 yielded

---

*As a point of statistical convention, it is not clear why the .05 probability was doubled. Once a directional alternative hypothesis (a one-tailed test of significance) is specified, differences in the wrong direction should be considered statistically nonsignificant.

**The summary table refers only to "favorable, zero, and unfavorable impact". In the overall context of the report, it is assumed that, e.g., "favorable impact" is meant to refer to a statistically significant result in favor of LRE classes.

no difference (a total of 82 comparisons which did not favor the LRE groups), and only 26 favored the LRE groups. Hardly rousing support for LRE.

It is of interest to school people to note that, as one might expect, in the national study explicit LRE instruction did again produce statistically significant higher mean scores on tests of knowledge of the law and judicial processes. In 24 of the 27 comparisons in the national study, the LRE class had a statistically significant higher mean (at the control mean. Again, it is unfortunate that the information is unavailable to determine whether these differences can be considered educationally significant.

Theory Testing

Readers of the two LRE Evaluation Project Reports (Hunter & Turner, 1981; SSEC-CAR, 1983) will note a heavy emphasis in both on a theory of delinquency causation, pulling together "control theory, strain theory, and labeling theory" as a basis for instrumentation and analysis decisions (Hunter & Turner, 1981, pp. 2-4; SSEC-CAR, 1983, e.g., pp. 1-2 to 1-3, 1-5, 3-4, 3-13). Substantial portions of the findings sections are devoted to tests of theory, and those who have read the reports may wonder why this critique has not referred to the theory-testing to this point. There are several reasons for that apparent oversight.

In the first place, the primary purpose of educational evaluation, as distinct from social science research, is to provide evidence and conclusions in regard to whether an educational practice "works" and under what conditions. The resources of an evaluation project should be directed primarily toward that end. Questions about theory are interesting, but secondary; obtaining information on theory, when possible, is a bonus, and should not be a central concern.

Secondly, evaluation reports, as I have noted earlier, should speak to the concerns and needs of the audience. If that audience is taken to be LRE educators, it is crucial to note that law-related education is notably atheoretical. To the extent that LRE has been successful (and perhaps its greatest success is indicated by the LRE Evaluation Project findings that students reported LRE classes to be "better than other classes", Hunter & Turner, 1981, p. 19; SSEC-CAR, 1983, pp. 4-21, 4-30), it has been due, in my opinion, to an intuitive correspondence with a rationale for instruction with which John Dewey (e.g., 1916, 1933) would be very comfortable: That is, law-related education is activity oriented and it involves students in areas of concern that are relevant to their own lives. To say that the LRE Evaluation Project investigated "the theoretical premises upon which LRE is based" (SSEC-CAR, 1983, p. 1-5) is simply not accurate. The theory tested is one constructed independently of law-related education, not one which has formed the basis for law-related education. The results pertinent to that theory are, therefore, likely to be of little relevance to school people and curriculum developers in the field.

That relevance is vitiated even more by the general status of, and attitudes toward, theory building in education. To say that the results of efforts at educational theory building have been unproductive is no exaggeration (see, e.g., Shaver, 1982), nor is it an exaggeration to indicate that teachers have generally not found the results of theory-oriented research to be useful (Eisner, 1984). The general skepticism about research efforts to build theory and about the usefulness of such theory in educational practice suggests that the discussion of theory and of results relevant to theory in the two LRE Evaluation Project reports will serve primarily as distractors for school people who try to ferret out what the results of the Law-related Education Evaluation Project might tell them about effective LRE.

A reading of the "recommendations for improved implementation (SSEC-CAR, 1983, Ch. 7) will not allay the skepticism about theory and the usefulness of attempting to translate theory into practice. The recommendations are vague, in part reflecting contradictory findings (e.g., p. 7-2). It is of interest, along those lines, that increased small group work in the second year, based on a first year recommendation for more active student participation, was negatively associated with both attitudes and nondelinquent peer relations, apparently because of inadequate directions given to students, the use of unsuitable exercises, and the failure of the exercises to produce "true task interdependence" and "explicit reward interdependence" (SSEC-CAR, 1983, p. 7-6).

In any event, as one might expect, the results in regard to the theory were not clear-cut. Some of the hypotheses based on the theory were not supported by the data (SSEC-CAR, 1983, p. 5-16, 5-31, 32), including a nonsignificant correlation that is inconsistent with the hypothesis that law-related education would exert an effect on delinquent behavior through gains in knowledge of the law. As a result of the weak and inconsistent associations, the project had to resort to "after-the-fact interpretations of findings" (1983, pp. 5-16, 5-49), not a strong confirmation of the theoretical base.

Ironically, recommendations for practice were once again made, based on the theoretical model (pp. 5-50, 5-51). But they are not likely to be of great or unique use to teachers. One example will illustrate the point. From the report (p. 5-50):

> Student attachment to the teacher is a powerful tool for building belief in the moral validity of law in influencing delinquent behavior. Attachment can be built by interactive and well-paced teaching; by sharing instructional objectives with students, and by preparing students mentally to receive instructions; by striking a skillful balance between adequate concreteness and detail in the time available for the instruction; and particularly by checking frequently for student understanding during instruction and during student practice using information gained to adjust instruction accordingly.

This sounds like a general recommendation for good teaching, regardless of the theoretical proposition of student attachment to the teacher. Moreover, it lacks the specificity to be of much assistance to a teacher deciding what to do in the classroom.

In short, this methodological critique has largely ignored the theory and the findings related to it because they seem to have little relevance to the primary evaluation question about the effectiveness of law-related education, because the theory is imposed upon LRE rather than growing out of it, and because there is little of interest or use to school people in the "theory-related" findings. The general unproductivity of efforts to build theory with relevance to classroom practice, as well as the inconsistent support of the delinquency theory by the findings of this study, do not argue for serious consideration of theory as an evaluation, as contrasted to a social science, concern. Again, educational-social science research and educational evaluation are not identical fields, even though they have elements in common. My preference would have been for greater attention to the assessment, analysis, and reporting issues raised in this critique, so that the evaluation results would have reflected legitimate concerns about educational significance or importance, at the cost of the time and effort spent on theory.

Conclusions

The motivation to initiate this symposium came from a deep concern over the misinterpretations of the report of the first year LRE Evaluation Project findings that I saw in LRE publications. At the time that I proposed the symposium, I had not seen the report for the second year of the project. Once I learned that it would be available prior to the symposium, it was my hope that the second year findings would clarify questions and resolve doubts which I had about the results of the first year project. Unfortunately, as my comments above indicate, that has not been the case.

Most educational evaluation research can be subjected to criticism. I have already noted the difficulty of conducting studies of this sort. This paper is not intended as a "hatchet job", nor as carping about incidental aspects of the research and report by the LRE Evaluation Project. School people need to be aware that there are fundamental questions about the interpretability of the Project results, because of how the research was conducted and the findings reported. Moreover, the findings do not square with the rather optimistic conclusions.

The non-researcher who attempts to sift out legitimate conclusions from the summary of the first year report (Hunter & Turner, 1981) that was distributed (which lacks adequate detail to understand what was done or to comprehend the interpretations presented) or from the second year report (which is heavily detailed and statistical) faces a monumental task. There may be more that could be discerned from the Year 2 report (especially in Ch. 3, which describes teachers' reports of the difficulties in implementing the three LRE curricula), and some parts may deserve more attention than could be given to them in this symposium paper. Nevertheless, a critique focused on the "impact" findings is particularly appropriate for an educational conference such as this one. I would urge educators to be cautious about relying on the report to

advocate law-related education, and not to feel too discouraged if they find it difficult to relate the delinquency causation theory findings to the practice of law-related education.

REFERENCES

Borg, W. R., & Gall, M. D. (1983). Educational research: An introduction (4th ed.). New York: Longman.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Clark, J. P., & Tifft, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. American Sociological Review, 31, 516-523.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. New York: Wiley.

Dewey, J. (1916). Democracy and education: An introduction to the philosophy of educaton. New York: Macmillan.

Dewey, J. (1933). How we think: A restatement of the relation of reflective thinking to the educative process. Boston: D. C. Heath.

Eisner, E. W. (1984). Can educational research inform educational practice? Phi Delta Kappan, 65, 447-452.

Empey, L. T. (1982). American delinquency: Its meaning and construction. Homewood, IL: Dorsey.

Hunter, R. M., & Turner, M. J. (1981). Law-related education evaluation project: Final report, phase II, year 1. Boulder, CO: Social Science Education Consortium and Center for Action Research.

Jensen, G. F., & Rojek, D. G. (1980). Delinquency: A sociological view. Lexington, Massachusetts: D. C. Heath.

Kerlinger, F. N. (1977). The influence of research on education practice. Educational Researcher, 6, 5-12.

Shapiro, J. Z. (1984). On the application of econometric methodology to educational research: A meta-theoretical analysis. Educational Researcher, 13(2), 12-19.

Shaver, J. P. (1979). The productivity of educational research and the applied-basic research distinction. Educational Researcher, 8(1), 3-9.

Shaver, J. P. (1980, April). Readdressing the role of statistical tests of significance. In W. R. Borg (Chair), Tests of significance: Readdressing their role. Symposium at the meeting of the American Educational Research Association, Boston.

Shaver, J. P. (1982). Reappraising the theoretical goals of research in social studies education. Theory and Research in Social Education, 9, 1-16.

Shaver, J. P. (1983). The verification of independent variables in teaching methods research. Educational Researcher, 12(8), 2-9.

Social Science Education Consortium and Center for Action Research. (1983). Law-related education evaluation project: Final report, phase II, year 2. Boulder, Colorado.