Renzulli, J. S. (1972). The confessions of a frustrated evaluator. *Measurement and Evaluation in Guidance, 5*(1), 298–305.

The Confessions of a Frustrated Evaluator

Joseph S. Renzulli*

Anyone planning to deal with a theoretical or practical problem in educational evaluation these days cannot help but be overwhelmed by the massive and often conflicting body of literature that has grown up around this topic. With the possible exception of the three subjects—segregation, sex education, and student power—it is difficult to find an educational issue that has generated more rhetoric and greater controversy than the current concern for *evaluation* and the related issue of *accountability*.

As educators with dollar signs in their eyes busily scrambled to scoop up their fair share of the federal windfall that began with the enactment of the National Defense Education Act (NDEA), only the very foolish were not haunted by the inescapable realization that sooner or later they would be held accountable for the brave but often vague promises and grandiose schemes that were unblushingly written into thousands of proposals by well-intentioned but often starry-eyed educators who gave little or no serious attention to the problems of evaluation. Thus, the schemes that promised to provide solutions to schoolmen's prayers turned out, in most cases, to have little or no significant impact on the education system in general; and although the trimmings in today's schools may be a little fancier, the main course that is served to most youngsters is not very different from the educational menu of two or three decades ago. The main question is, of course, why is this so? Why in the face of unprecedented financial support for educational programs and projects have we been unable to demonstrate in any definitive fashion the effectiveness of proposed innovations in the system? The challenge of answering this question must be laid squarely at the feet of the educational evaluator. Although great strides have been made in the science of evaluation in recent years, our inability to master certain basic problems still prevents us from bringing about the often discussed but still illusive transformation of the schools.

The early chaos and misapplication of many federally financed programs have undoubtedly contributed to our heightened concern for evaluation and accountability, and this concern has given rise to a new methodology or second generation of educational evaluation. Unlike the first generation that seemed to be hung up on psychometrics and problems related to the measurement of individual differences, the second generation is concerned with system or program evaluation, and its main purpose is simply to find out how specific modifications in system inputs will bring about specific changes in system outputs. Although this methodology has focused so far mainly on special projects and externally financed programs, it does not take a great

* Joseph S. Renzulli is Associate Professor of Educational Psychology and Director of the Teaching the Talented Program, University of Connecticut, Storrs.

deal of imagination to envision the day when this rapidly expanding technology is refined to the point where it can be brought to bear on virtually every activity that parades under the banner of education.

But before we can reach the seemingly impossible dream of isolating cause-effect relationships in education, certain basic and essentially unresolved problems in evaluation must be mastered. This article attempts to provide a structural overview of some of the major problems and issues that continue to plague educational program evaluation. By isolating and pointing out the main dimensions of these problems, we hope to provide some direction for future efforts and to bring the problems into sharper focus so that the second generation of evaluators will be better able to resolve them.

Politics of Educational Evaluation

The first problem is concerned with the relationship that exists between the evaluator and the program being evaluated—that is, the politics of educational evaluation. Educational enterprises, and especially those that involve large amounts of funds from external sources, are managed by people and institutions that stand to gain certain benefits if their projects prove successful.

Although the main criteria for the success of any educational program should be in terms of benefits realized by students, we cannot deny that institutions stand to gain such benefits as continued funding and prominence from a successful program, and that the persons operating these programs stand to gain job security, prestige, and power in the form of decision-making authority. With these stakes consciously or unconsciously in mind, the administrators of programs and projects seek personnel who are euphemistically dubbed “independent external evaluators.” In most cases the project managers have complete freedom in selecting the evaluator, and it is from their budget that the costs of evaluation are paid.

Notwithstanding the fact that project administrators are genuinely interested in finding ways in which their programs can be improved, we would be deluding ourselves if we believed that they are not primarily interested in getting a good overall evaluation, or at least avoiding an evaluation that might conclude with such statements as: “This project is a complete waste of time and money, and shows no noticeable benefits for the students” or “The project director is incompetent and should be fired immediately.” Please remember that this is the same project director that hired the evaluator, squired him through several three-martini lunches, and suggested that he publish the results of his evaluation in a professional journal.

If project administrators’ motives, so far as evaluation is concerned, appear to be somewhat less than virtuous, we should keep in mind that they have been forced into a situation over which they have had little control. In most cases, program evaluation is mandated by higher-level administration or funding agencies, and as such, is usually considered a necessary evil by project administrators and staff members who are asked to avail themselves of the evaluator’s instruments of oppression so that their competence can be judged.

A few years ago I was involved in the evaluation of a large compensatory education program for inner city youngsters. On one occasion a large group of teachers and administrators were called together for a briefing on evaluation and the distribution of evaluative materials. As the group began to leave the auditorium, a quiet chant began to reverberate throughout the room—"Two, four, six, eight—we don't wanna evaluate!" That was among some of the kinder things that were said throughout the course of this particular evaluation.

It is difficult to understand how the need for evaluation is almost universally accepted at the cognitive level, and yet the efforts of the evaluator are likely to be greeted with all the warmth and understanding of a husband who finds his wife in bed with his best friend. Where have we gone wrong? Must practitioners feel threatened by evaluators and ambivalent about taking part in an evaluation? Must the role of the evaluator bring him into conflict with the role of the practitioner? If the answers to these questions are anything less than a resounding "No," what can we do to gain the understanding and support of persons on whom we must eventually pass judgment? A closer look at the evaluator may help us isolate a few more problems.

First, the evaluator usually stands to make a financial gain from his involvement with a funded project. While there is nothing inherently evil about turning a reasonable profit from one's labors, there are a few inherent subtleties that all of us who have played the evaluation game are aware of. If we come up with a generally favorable evaluation we are likely to be rehired, and under these circumstances it is difficult to dismiss from our minds the great temptation on the part of the evaluator to report some needed improvement in minor areas while at the same time pointing out the general value and effectiveness of the bulk of a program.

A second problem is somewhat less embarrassing to talk about; however, it certainly must be considered as a source of evaluative bias. Good evaluation theory tells us that the evaluator should be involved in a project from the outset. He should work closely with the administrators and staff from the proposal-writing stage to the preparation of the final report. He is a full-fledged member of the team and, as such, he is quite likely to develop both a close personal relationship with project personnel and a strong emotional tie with the project itself. He starts to talk about our students, and our control group, and how much money we have left in *our* budget. Under these circumstances, we must raise the question: Is the independent external evaluator sufficiently detached from the project to take an objective look at it, or is he likely to approach evaluation with some of the same positive biases as a person who sets out to evaluate his own program?

We have somewhat of a dilemma here. If we were able to create a completely independent cadre of external evaluators—political untouchables with all the consumer concerns of a group like Nader's Raiders, would we not also widen the gap between the evaluator and the people whose honesty, trust, and cooperation are needed to mount an effective evaluation? In one situation where this kind of independence existed, teachers prepared students for a post-test by teaching lessons directly from the test booklet.

As long as factors such as funding, prestige, and power are involved, the political relationship between the evaluator and those being evaluated will not be an easy problem to solve.

One Irresistible Force Meets Another

Two irresistible forces in education seem hell-bent on a collision course, and I am afraid that our friend the evaluator is going to be caught squarely at the point of impact. The first irresistible force is the *behavioral objectives movement*. Although there is some growing controversy about the role of behavioral objectives in curriculum planning and evaluation, one cannot deny the value that it has had in helping to build evaluation and accountability models and to advance the science of education beyond the vagueness and lack of specificity that seems to have made us the step-child of the sciences. Further, some of the new experiments in education such as performance contracting and the voucher plan, and systems analysis approaches such as the planning, programming, and budgeting systems (PPBS) could not have been implemented had we not been able to specify in precise behavioral terms the objectives toward which our efforts are directed. Some educators, however, may be carrying their concern for behavioral objectives a bit too far. Nevertheless, this movement has been nothing less than a blessing to the evaluator who must know what is *supposed* to happen before he can begin figuring out how to measure it.

There is still another irresistible force growing in education today—a renewed concern for the total development of the individual as a human being, dealing with such difficult-to-measure objectives as self-actualization, consciousness III, and sociability. This force is somewhat more obscure than the behavioral objectives movement, but is nevertheless growing in force and magnitude.

As youngsters on college campuses across the country began to revolt against curricular irrelevance and lack of concern for the affective domain, educators from primary grades through graduate school began to give serious attention to the benefits that might be realized from a much more informal and humanistic education—one that replaces punch-card relationships with primary-group experiences, classroom activities and didacticism with experiences in the real world, and punitive testing and grading practices with concerns that focus on individual satisfaction. Supporters of this humanistic point of view remind us that a large proportion of man's ,behavioral repertoire is *not* acquired in formal learning situations, and that although the school can claim credit for giving youngsters the three R's and other skills and information, they have done so at the expense of other equally important objectives. Another argument is that this lack of humanistic concern for individuals has made schools essentially oppressive places and learning essentially coercive. Education, the humanists say, has degenerated to a process of conformity-oriented hurdle-jumping, and they point accusingly to the behavioral objectivists for making the process even more mechanistic.

Those with a humanistic orientation are also concerned with the objectives of education, but they do not agree with the behaviorally oriented about the precision with which the objectives can be measured. How can we determine in an objective and

scientific manner, argue the behaviorists, when an educational experience has contributed to the development of a fully functioning, self-actualized, socially conscientious human being?

Herein lies the problem for the evaluator. Some truly relevant experiments are taking place in education today—experiments such as open classrooms modeled after the British primary system, alternate learning centers that do not have a formal curriculum, and experiences in group dynamics that are designed to improve race relations and to narrow the generation gap. Recently I read a very exciting Title III proposal designed to advance innovation and creativity in education. The guidelines for evaluating this proposal had a strong behavioral-objectives orientation, and, using these criteria, I was forced to give the evaluation design of this particular project a very low rating. Although objective tests were written into the evaluation design of the proposal, what was clearly written between the lines was a ritualistic compliance with state department of education regulations, a compliance that I believe would yield little useful information about the true objectives of the proposed program. Further, if this proposal should be funded, determining its effectiveness will be nothing less than an evaluator's nightmare.

I wish that I could offer some concrete suggestions for softening the impact of the collision between the two forces. The usual impassioned plea for research into the measurement of affective processes can only be reiterated, but perhaps we should also make a plea to funding agencies that will result in some relaxation of the rigid behavioral-objectives model that governs so much of the thinking so far as evaluation is concerned.

What Is Wagging What?

Another major problem with which the front-line evaluator is concerned is the distinction between *evaluation* and *research*. How much of the rigor and control that one finds in traditional research design is necessary in order to carry out a "respectable" evaluation? Like the researcher, the evaluator does not want to be accused of employing an inappropriate or sloppy design, and since many second generation evaluators have entered this field with a strong research background, they are constantly haunted by the fear that they will lose the respect of their more rigorous colleagues.

Unfortunately, the theorists and model builders in evaluation have shown us little consistent direction in this regard. For example:

> One of the more obvious blots on the otherwise nearly-clean escutcheon of the educational research community stems from its ill-fated involvement with evaluation. In responding particularly to Federal mandates for better evidence of program success, educational practitioners have sought the assistance of educational researchers in designing and carrying out evaluative studies. The resulting effort has been a failure so conspicuous that I regard it as unnecessary to attempt to document it. That failure, I contend, has, as one of its chief (but not

only) roots, the mistaken assumption that the research paradigm is appropriate to evaluative inquiry. Nothing could be further from the truth. The unfortunate marriage produced by this error in judgment has left irremedial scars on both parties [Guba, 1969, p. 4].

Contrast this statement with the following comments in a recent issue of *The Urban Review* (1969) which was devoted solely to program evaluation:

> Both [research and evaluation] should he serving the same function of supplying information to planners and policy makers about what does and does not work … but accurate information can stem only from rigorous experimental design and data collection techniques, whether in research or evaluation [Hawkridge & Chalupsky, p. 8].

> There are no formal differences between "basic" and "applied" research or between "research as such" and "evaluation research." Research designs, statistical techniques, or data collection methods are the same whether applied to the study of the most basic principles of human behavior or to the most prosaic of social action programs [Rossi, p. 17].

> From a technological viewpoint evaluation is a fundamental research activity. It can be conducted with as much precision as any other form of research, though it presents certain specialized problems such as controlling for the amount of attention given to the experimental group or that of executing a scientific design under naturalistic conditions [Mann, p. 12].

These last three statements are in sharp contrast to the first by Guba, and his divergence of opinion raises some significant problems for the evaluator. Should the evaluator insist on research that is rigorous in helping people to evaluate educational programs or should he evaluate the program in its naturalistic setting? Is the tail not wagging the dog when we change a program so that it will conform with an experimental design?

Some prominent researchers (Mackie & Christensen, 1967; Ottinger, 1969) have told us that "a very small percentage of findings from leading research studies are useful, in any direct sense, for the improvement of training and educational practice" (Mackie & Christensen, 1967, p. 5) because conditions employed by the research bear no determinable relationship to conditions outside the research setting. Commenting in a similar fashion in a recent issue of the *Review of Educational Research*, Cohen has expressed the opinion that:

> Experiments with decentralization, tuition vouchers, doubling per-pupil expenditures, and radical changes in secondary education have two salient attributes in common: to have meaning they would have to be carried out in existing schools, and few schools would be likely to oblige [Cohen, 1970, p. 233].

Thus we are faced with another dilemma. If our evaluations do not respect some of the mandates of good educational research, then they are unlikely to hold any water and the evaluator may be accused of not really demonstrating the effectiveness of the program under consideration. This is especially true when the evaluator depends heavily on soft data such as interviews and observations, data that usually cannot be treated with sophisticated statistical techniques as easily as hard data. Although the soft-nosed evaluator does not completely negate the value of statistics, he argues with great vehemence that the sights, sounds, and smells of the real classroom are easily hidden by statistics. Further, if the evaluator imposes strict controls over the program he is evaluating, he may be accused of ignoring the real world in favor of a design that has scientific respectability. Not only may the tail be wagging the dog in this case, but we must also question the impact that three-decimal research has had in bringing about major changes in education to decide whether or not the evaluator with a research bent wants to play the same game. Change is often governed by the heart as well as the head and, whether we like it or not, I think that much of the impact of a book like *Crisis in the Classroom* is the result of its heat-rendering anecdotes rather than tables of means, standard deviations, and correlation coefficients.

When in Doubt—Give Another Test

The last problem discussed here concerns the role of testing and especially achievement testing in the evaluative process. If we were fortunate enough to have at our disposal a series of standardized measures that accurately reflect the stated objectives of our program, the problem would be quickly solved. But because of the vast changes that are taking place in the curriculum these days this is seldom the case. Thus we are left with the choice of either using standardized tests that are inappropriate in varying degrees, or attempting to develop our own tests. The latter choice, however, presents us with another kind of problem. If we are going to build our own measuring instruments, we must also deal with the psychometric problems with which the first generation of educational evaluators were concerned. Without some assurances that our homegrown instruments possess reliability, validity, objectivity, and practicality, our results are likely to be viewed with suspicion. Building in these scientific requirements is of course within the realm of possibility; however, it is often a luxury that the evaluator of a relatively small project cannot afford.

The use of standardized tests in evaluation also poses another kind of problem. A growing body of research findings indicates a somewhat limited relationship between the typical achievement criteria for program success and the presumed adult consequences of education, such as better jobs, and higher income. This non-relationship is particularly apparent among black students. A study by Blau and Duncan (1968) showed that once inherited status is controlled for, years of school completed is only moderately related to adult occupational status, and the relationship between education and occupation is much weaker for blacks than for whites.

It appears then that a word of caution is in order here for at least the evaluators of compensatory programs who use school achievement as a proxy for the long-range criteria of success in adult life. We might even raise the question of whether

standardized achievement batteries have outlived their usefulness, a usefulness that a few brave souls in education have always questioned. With the trend moving away from grouping (one of the major uses of standardized achievement tests), the limited relationship between these tests and curricular experiences, and the dangers of self-fulfilling prophecies always hanging over our heads, perhaps the evaluator can take the lead in questioning the usefulness of those tests.

References

Blau, P., & Duncan, O. (1968). *The American occupational structure.* New York: Wiley.

Cohen, D. (1970). Politics and research: Evaluation of social action programs in education. *Review of Educational Research, 40*(2), 213–238. https://www.jstor.org/stable/1169534

Evaluating educational programs: A symposium. (1969). *The Urban Review, 3*(4), 4–22. https://doi.org/10.1007/BF02322249

Guba, E. G. (1969). Significant differences. *Educational Researcher, 20,* 4–5.

Mackie, R. R., & Christensen, P. R. (1967). *Translation and application of psychological research.* Goleta, CA: Human Factors Research.

Ottinger, A. G. (1969). *Run, computer, run.* Cambridge, MA: Harvard University Press.