REFERENCES

Harshbarger, D. (1974). The human service organization. In H. W. Demone and D. Harshbarger (Eds.), A handbook of human service organizations. New York: Behavioral Publications.

Harshbarger, D. (1983). Executive effectiveness: What the books don't tell us. Invited Address, Annual Meeting of the Association for Behavior Analysis, Milwaukee, Wisconsin.

Peters, T. J., & Waterman, R. H., Jr. (1982). In search of excellence, New York: Harper & Row.

Sommer, R. (1963). Expertland. Garden City, NY: Doubleday.

Evaluation at the Frontier: Some "Timely" Comments for Future Use

Paul M. Wortman
ERS Presidential Address
University of Michigan

Occasions such as this are times when one reflects on the past and ponders the future. It was only twenty years ago that the now classic monograph by Campbell and Stanley (1963) was first published. It was only fifteen years ago that the RFP to evaluate Head Start was issued. It was only ten years ago that the first government-sponsored training program in evaluation research admitted its first students. And, of course, both of the organizations meeting here have been in existence a shorter time than that. The question we have gathered together to address is whether a field with such a short history has a long future before it. And, if so, what does it look like?

TIME

Evaluation has always been concerned with time or—perhaps more accurately—timeliness. The logic has been that if evaluation reports are

not available at the proper time, they will not be utilized and the whole exercise will be for naught. The consequence of failing the timeliness test is a distraught, anxiety-ridden, neurotic evaluator. Or worse yet, a distraught, anxiety-ridden, neurotic and unemployed evaluator. Of course, those of you who have had government evaluation contracts know that even meeting deadlines often results in distraught, anxiety-ridden, neurotic evaluators.

If there is one commandment that guides evaluation it is this: "Thou shalt have thy final report in on time." Cronbach (1977), in his remarks to the inaugural meeting of the Evaluation Research Society in 1976, said: "Data are no good if the report on them is ready too late." He went on to say a lot of other things with which I also disagree, but timeliness borne of utility formed the central rationale for his comments. Others such as Weiss (1977) and, most recent, Chelimsky (1981), in an address to this same forum, have echoed the need for timely evaluation results to ensure proper utilization. I would like to offer a dissenting opinion. I believe evaluation has for too long been crucified on a cross of timeliness.

One has only to look back at the Head Start evaluation to see the problems generated by a so-called timely evaluation. Datta (1976) has provided an excellent history of that evaluation and its impact. The Westinghouse-Ohio University evaluation report was submitted on time within a year of the contract award. And it was utilized—both in an enlightenment or conceptual way, but also in an instrumental way. Head Start was labeled a failure and became grist for the intellectual mills viewing early educational interventions as based on improper theory. Such thinkers as Hernnstein, Jencks, and Moynihan were in the forefront of this effort. In commenting on the impact of the report, Datta (1976) stated:

> Few articles on early childhood education written for social policy fail to mention the report as reasonably conclusive evidence that, if long-term effects on school achievement are the goal, Head Start alone is not the way to achieve it [p. 161].

The program itself also suffered directly. The summer programs were terminated, and funds for the year-long programs were frozen at a constant level.

What does this have to do with the inadequacies of timeliness? As many of you know, the Westinghouse-Ohio report was incorrect. While it was considered a "long-term" evaluation because it collected data from students in grades 1 through 3 who had previously been in Head Start, it did not assess important long-term effects. According to Datta, it was only through the personal efforts of then HEW secretary Elliot Richardson that the program survived at all. Nevertheless, I was somewhat surprised when I encountered a researcher a few years ago who claimed that he personally had saved Head Start. Despite being taken aback at his immodest introduction, I did subsequently read a copy of his evaluation study (Lazar et al., 1977).

The evaluation collected uniform follow-up data from 14 "experimental infant and preschool programs" that used Head Start or Head Start-like curricula. With but one exception, all the programs had been conducted prior to the late 1960s. The follow-up data collected in 1977 thus provided a real long-term assessment of high-quality early education interventions of the Head Start variety. The results, in contrast to the Westinghouse evaluation, found "a significant increase in IQ when compared to a control group that lasted for as long as three years." However, the follow-up data indicated that the IQ difference was not maintained. The most startling results from the study were that early education results in a significant reduction in the number of children needing to be either held back a grade or placed in special education classes. The authors of this study concluded that "the findings of this report now leave no reasonable doubt that in the main, programs which had deliberate cognitive curricula had a significant long-term effect on school performance."

The findings of this report make one wonder—and made me wonder—about the logic of the policymaker-utilization-timeliness linkage. Riecken, Boruch, and associates (1974) had said that "evaluation is a political act," and Cronbach, in his adddress and subsequent book *Toward Reform of Program Evaluation* (Cronbach et al., 1980), had amplified and extended this theme. In essence, it is a so-called policy-shaping community that drives the evaluation enterprise, according to Cronbach. Chelimsky, in her ERS presidential address, discussed a new process "which begins with the user and moves to an evaluation." But all of this smacks of the tail wagging the dog. Policymakers come and go, but the same old social problems remain. The current debate over the

quality of education and students' ability to function in an increasingly technological age is a reprise of the post-Sputnik angst that shook the educational system in the late 1950s and early 1960s and resulted, in part, in programs such as Head Start.

I know, all too well, the problems in generalizing from a single case. I had a similar "Aha!" experience in another report that I was inadvertently forced to read. My colleague, Lee Sechrest, had on numerous occasions spoken enthusiastically about the work of Roland Tharp and Ronald Gallimore and their general approach to evaluation described in their chapter in the 1979 *Evaluation Studies Review Annual* (Tharp & Gallimore, 1979). As with most busy academic evaluators, the paper soon was buried in my "must read when I have some time" pile. As fate would have it, I one day assigned the preceding chapter in the *ESRA* volume to my class and the secretary mistakenly photocopied and attached the Tharp and Gallimore chapter as well. Remembering the sage advice of my former colleague at Northwestern, Bob Boruch—"Waste not, want not"—and considering the substantial copying charges, I immediately assigned Tharp and Gallimore as well.

They describe an ecological approach to evaluation called the "climax" model that entails a lengthy period of research and development until the final stable program is attained. In all honesty, I cannot say that I was immediately enlightened by the chapter. In fact, I found it somewhat anticlimactic. (Well, last year we had subpoena envy.) My initial cynical response was that such an evaluation was nice work if you could get it, and that it provided full employment for evaluators. However, I was skeptical that an outcome evaluation would ever occur given the changing nature of innovations and that none was presented by the authors even after a number of years of evaluative work. But with time my attitude changed. As I began to study innovations in medicine, I saw quite clearly that even surgical and drug technologies change with time and that an early evaluation can be premature. I was all the more convinced when I saw that most major medical evaluations take five years or more. The recently completed medical social experiment examining primarily behavioral factors in reducing heart disease, called the Multiple Risk Factor Intervention Trial or MRFIT (1982) took ten years to conduct at a cost of $115 million (Kolata, 1982).

What does all this mean? I want to make it perfectly clear: I am not advocating that you withhold your final reports. However, I do find

fault with the current accepted logic that determines when a final report is due. Certainly a medical researcher would laugh at any policymaker who demanded an early report or preliminary findings. And most policymakers would not dare to ask. But there are numerous cases of this going back to the evaluation of Head Start. So why does this problem exist in education but not in medicine? And how does one determine the proper length of an educational evaluation study?

As to the first question, two related issues—equity and importance—seem to account for some of the difference in temporal approaches. Ill health, unlike poor education, seems to know no class or socioeconomic boundaries. While the more affluent can afford good education, they cannot easily buy their way out of cancer, heart disease, and other major illness. As a consequence, health may be seen as more important to more people. For drug and some medical devices true experiments or randomized clinical trials (RCTs for short) are mandated by law to determine their effectiveness and safety before they can be marketed. No such similar policy is deemed necessary in educational innovation. Again, this is perhaps an indication of the relative differential importance accorded the two areas. Certainly the status our society accords physicians and teachers would support that, as would the social expenditures for the two areas. Policymakers may feel comfortable in dictating to educators, given that it is largely a governmental activity from funding to actual delivery, but they do not, nor have they been very successful in controlling the medical care system.

The second question concerning the length of an evaluative study is a bit more problematic. It is perhaps simple and convenient to say the evaluation should be as long as the policymaker wants. But that ignores the internal dynamics and theoretical rationale for an intervention. Not only do programs take considerable time to develop, as Tharp and Gallimore note, but the program's effects may be small, incremental and, with luck, cumulative. This is not a new point—John Heilman said as much in 1980. As Heilman (1980) noted, such "programs are likely to have measurable effects only over a period of months or years." This is the general case in medicine and, I believe, in education as well. In a paper Bob Boruch and I wrote a few years ago (Boruch & Wortman, 1979), we observed that program developers too often either looked for large effects or promised them to funders and policymakers to obtain needed support. This raised expectations too high and set evaluations to

detect large, immediate effects. Under such circumstances just about any methodological approach would work. Mosteller (1981) and his associates (Gilbert et al., 1977) have shown, however, that the vast majority of innovations fail. Those that work often have small effects that are hard to detect without good methods such as RCTs with sufficient statistical power. For example, Bill Yeaton and I have been examining the research literature on coronary artery bypass graft surgery—the so-called CABG procedure. We found an average surgical benefit of only 4% greater survival than medically treated patients (Wortman & Yeaton, 1983). A study would require over 1100 patients to detect this effect reliably, and none of the 9 RCTs conducted included this many patients.

Well, I have been trying to answer the question indirectly. To be more specific, some time for program development and implementation must be allocated. This formative process usually takes a few years. Second, some realistic estimate of the expected effect must be made. From this the length of the evaluative study can be derived. Thus the utilization-driven timeliness approach must be reversed.

It would seem reasonable for evaluations to seek long-term effects allowing for some program development and the cumulative impact of incremental small effects. Perhaps evaluation should be "given away" as a former head of the American Psychological Association recommended for applied psychological research in his presidential address. At the very least, evaluators could leave easily maintained data archives that could be reviewed at appropriate intervals. I fully recognize that things can get out of hand and that evaluation reports may never get written. Clearly, there are limits. As John Maynard Keynes said, "In the long run we're all dead." Cronbach (1975) has worried about hidden temporal interactions that may change any relationship in the long run. We obviously have to be judicious with our scarce evaluative resources, and there is no guarantee that the results will differ from those found in the short run, as the recent follow-up report on psychotherapy found (Nicholson & Berman, 1983). But I believe inexpensive archiving may be one way to accomplish this. Thus we probably need an additional commandment: To paraphrase Paul Masson, we shall conduct no evaluation before its time.

METHOD

The timing of an evaluation has, I believe, direct implications for the methods employed. In this I am also in agreement with Heilman (1980) and Tharp and Gallimore (1979). It is time to end, once and for all, our methodological bickering and paradigmatic power struggles. There is a time and place for all methods. Depending on the circumstances, one or another may be the most suitable, and often multiple methods will be useful.

The early developmental stages of a program are best suited to the richly textured qualitative methods advocated by many (see Patton, 1978; House, 1980), as are unique situations often occurring in complex organizations. I recently *completed* a government contract to evaluate the National Institutes of Health Consensus Development Program (Wortman et al., 1982)—a unique approach to medical technology assessment using an evaluative concept similar to the "science court" (Kantrowitz et al., 1976). The program was one of a kind and just beginning its third year when the evaluation study began. This clearly was not the time or the place for a randomized experiment. Instead, we spent a lot of time observing the process, interviewing various participants, and reading archival material. That does not mean that other methods were inappropriate, only that it took considerable time to develop a conceptual framework that made such activity meaningful.

We did do a considerable amount of "number crunching" as well. We designed a survey instrument for conference participants; we performed a kind of content analysis on the consensus statements—the jury verdicts on the effectiveness, safety, and appropriate conditions for using the technology; and we did a small-scale randomized experiment comparing high-quality news reports to the consensus statements.

I mention this experience because as President of ERS I am representing an organization that has what Don Campbell would call a "hard science" image. Moreover, as a former Northwestern evaluator, I personally am associated with such methodological proclivities. Speaking for both, there is room and a clear need for methodological ecumenicism.

Incidentally, if I may return to the NIH evaluation for a moment, it is interesting to contrast the ease of conducting more formal evaluative studies to those in other areas of health. The Office of Management and

Budget guidelines prohibit formal experiments containing more than nine experimental subjects unless prior clearance is obtained. Since the clearance process takes at least nine months, it effectively eliminates methods in such short-term, timely evaluation. So our study of the utility of the consensus statements was limited to 18 physicians—nine were randomly assigned to read a consensus statement and nine read a report from a journal like *Science*. As I have already noted, such randomized studies are routinely required for drugs. In fact, NIH spends well over $100 million annually on RCTs such as MRFIT. In health it is not unusual to find various statutory requirements waived to allow such studies.

In education this does not appear to be common practice. Boruch, Cordray, and associates (1983) made just such a recommendation in their recent report to the Holtzman committee on federal education evaluation practice. This is a reasonable and sensible suggestion. In health it is also not uncommon to find a number of hospitals combining to conduct a single study. Such so-called multicenter clinical trials provide the large number of subjects needed to test an innovative program or technology in a short period. This avoids the temporal interactions feared by Cronbach and the technological change feared by medical researchers. For example, the MRFIT study involved 12,866 men recruited at 22 different clinical centers over a two-year period.

In contrast, the follow-up study of Head Start-like experimental programs was a retrospective, patched-up multicenter evaluation. While participants did agree on a common posttest protocol, there was no common pretest protocol, uniform eligibility requirements, or random assignment. In fact, the researchers had difficulty obtaining funding given the preconceptions derived from the Westinghouse-Ohio evaluation. As far as I know, there have been no randomized multicenter educational evaluations.

Randomized studies are the Cadillacs of research methods, and they carry a price tag that may not sell in today's economic market. Other methods that have recently become fashionable are quite cost-effective and have achieved some interest in the governmental-evaluation sector. I am speaking of research synthesis methods such as Glass's meta-analysis. Chelimsky, in her presidential address, described the General Accounting Offices's plans to employ "evaluation synthesis" as a timely, policy-relevant, inexpensive method. Since then GAO has conducted a

number of these syntheses—all of which have been qualitative, by the way. My colleagues Fred Bryant, Bill Yeaton, and I have been conducting quantitative syntheses in education and health.

In education, Bryant and I have been examining the literature on school desegregation and academic achievement (Wortman & Bryant, in press). Despite the policy relevance and social resources and conflict engendered by this issue, it is a most desultory literature. We were able to locate over 100 studies, of which only 19 were usable for such a synthesis. There were no usable randomized studies and most were weak quasi-experiments. Surprisingly, only two of the usable studies had been published. Most were unpublished hard-to-locate dissertations—a truly, and perhaps properly, fugitive literature.

By contrast, Yeaton and I uncovered 25 high-quality studies out of the 90 we found on the CABGS procedure (Wortman & Yeaton, 1983). Nine were RCTs and the remainder were strong quasi-experiments. All were published in good medical journals. Again, this points up the variation in methodological quality by area. The irony is that an impoverished area such as educational evaluation produces studies of such poor quality as to make cost-effective procedures such as research synthesis of questionable value.

Again, it is interesting to speculate on the reasons for the difference in methodological orientation toward evaluation in the two areas. The same factors noted earlier seem to be important. In addition, the life-threatening aspects of more concrete, tangible health interventions appear to play a role. After all, it was the thalidomide scare in the early 1960s that led to the requirement for "adequate and well-controlled investigation." These have been interpreted to be two well-done RCTs. While ease of conducting such evaluative studies may partly explain the difference, other factors seem to be at work as well. Those noted earlier seem important. In addition, NIH has funded numerous complex, expensive RCTs. NIH has escaped much of the political turmoil that has characterized education, perhaps due to its reputation as a hard science, research organization. The National Institute of Education never had the funding or the staff leadership to accomplish this. Only last week I learned that NIE would not publish a report of a consensus-like conference examining the desegregation issue even though the panel was stacked to find with no effects or harmful effects of desegregation. Of course, it did not, and the political hacks running NIE have jettisoned

the report and fired the project officer. While NIH is by no means free of petty bureaucrats and politics, there is more scientific freedom to be controversial.

RECOMMENDATIONS

The question I posed at the onset concerned the future of evaluation research. What, then, are the implications of my intervening comments for the profession?

On the temporal aspects of evaluation, there are at least two points to be drawn. First, the evaluator and program designer have to play a more central role in educating the funder, and others in the "policy shaping community," on the proper design of an evaluative study. In health a biostatistician is typically given a central role and carries out this function.

Second, the utilization-driven approach is largely a government evaluative contract activity. However, the portfolio of evaluative activities must now be expanded beyond the federal government. The theme of this meeting is to explore not only ways to make this relationship work better but, more important, to explore other areas requiring evaluative skills. Private business is one major area, especially in the areas of market research. Here many of the issues and methods of evaluation, including timeliness, will be similar. However, the problems may be more diverse, the bureaucracy less rigid, and the rewards more enriching. Health is another area, and one that has been used throughout this address to illustrate the potential for evaluation in other areas. This does not mean that health is an evaluator's Nirvana. As Tom Chalmers has written, there is great room for improvement. However, the foundations are there for significant evaluative work without the hassles noted earlier. In either case though, the evaluator—you and I—must play a more active educative role.

Concerning the methodological issues, there are also a number of points. Multiple methods are appropriate to the proper conduct of evaluations, even within a single study. Evaluators have to recognize that their specific skills may be inadequate for a particular study. This can be handled by learning new methods—the function of the presessions held here yesterday—or bringing together a team of people with

the requisite skills. There are, as the current program and recent texts indicate, many relevant new skills to be mastered: Cost-benefit analysis, structural equation models, survey methods, and quality-of-life measures are among the many new and emerging skills needed.

There are, in addition, different methods that may be most appropriate at different times, as noted earlier. Thus the evaluator may have to apply the temporal perspective and constraints to the methods available. A quick-and-dirty RCT or outcome evaluation, such as the Westinghouse study, is no longer "kosher" by today's standards. Evaluators must temper their need for economic survival with methodological honesty.

In sum, we must believe that there is a fruitful future for evaluation. As long as society has the resources and the will to cope with evolutionary challenges through active interventions, there will be a need for well-thought-out and well-conducted evaluative studies.

REFERENCES

Boruch, R. F., Cordray, D. S., et al. (1983). Recommendations to Congress and their rationale: The Holtzman project. *Evaluation Review,* 1983, *7,* 5-35.

Boruch, R. F., & Wortman, P. M. (1979). Implications of educational evaluation for evaluation policy. In D. C. Berliner (Ed.), *Review of research in education.* Washington, DC: AERA.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on training. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Chelimsky, E. (1981). *Designing backward from the end-use.* Paper presented to the Evaluation Research Society Annual Meeting, October.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30* 116-127.

Cronbach, L. J. (1977). Remarks to the New Society. *Evaluation Research Society Newsletter, 1,* 1-3.

Cronbach, L. J. & associates (1980). *Toward reform of program evaluation: Aims, methods, and arrangements.* San Francisco: Jossey-Bass.

Datta, L.-E. (1976). The impact of the Westinghouse/Ohio evaluation on the development of Project Head Start: An examination of the immediate and longer-term effects and how they came about. In C. C. Abt (Ed.), *The evaluation of social programs.* Beverly Hills, CA: Sage.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977). Progress in surgery and anesthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, and F. Mosteller (Eds.), *Costs, risks and benefits of surgery.* New York: Oxford (1980). University Press.

Heilman, J. G. (1980). Paradigmatic choices in evaluation methodology. *Evaluation Review, 4,* 693-712.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: Sage.

Kantrowitz, A. et al. (1976). The science court experiment: An interim report. *Science, 193,* 653-656.

Kolata, G. (1982). Heart study produces a surprise result. *Science, 218,* 31-32.

Lazar, I., Hubbell, V. R., Murray, H., Rosche, M., and Royce, J. (1977). *The persistence of preschool effects: A long-term follow-up of fourteen infant and preschool experiments.* Washington, DC: Government Printing Office.

Mosteller, F. (1981). Innovation and evaluation. *Science, 211,* 881-886.

Multiple Risk Factor Intervention Trial Research Group (1982). Multiple Risk Factor intervention trial: Risk factor changes and mortality results. *JAMA, 248,* 1465-1477.

Nicholson, R. A. & Berman, J. S. (1983). Is follow-up necessary in evaluating psychotherapy. *Psychological Bulletin, 93,* 261-278.

Patton, M. Q. (1978). *Utilization-focused evaluation.* Beverly Hills, CA: Sage.

Riecken, H. W., Boruch, R. F., & associates (1974). *Social experimentation: A method for planning and evaluating social intervention.* New York: Academic Press.

Tharp, R. G., & Gallimore, R. (1979). The ecology of program research and evaluation: A model of evaluation succession. In L. Sechrest and associates (Eds.), *Evaluation studies review annual (Vol. 4).* Beverly Hills, CA: Sage.

Weiss, C. H. (1977). Introduction. In C. H. Weiss (Ed.), *Using social research in public policy making.* Lexington, MA: D. C. Heath.

Wortman, P. M., & Bryant, F. B. (in press). School desegregation and black achievement: An integrative review. *Sociological Methods and Research.*

Wortman, P. M., Vinokur, A., Sechrest, L., & associates(1982). Evaluation of the NIH consensus Development Process—Phase I: Final Report. Ann Arbor, MI: Institute for Social Research.

Wortman, P. M. & Yeaton, W. H. (1983). Synthesis of results in controlled trials of coronary artery bypass graft surgery. In R. J. Light (Ed.), *Evaluation studies review annual (Vol. 8).* Beverly Hills, CA: Sage.

Comments on Thomas C. Chalmers's Address: Evaluating Clinical Trials

William Yeaton
University of Michigan

Those ERN/ENet members who remained for Saturday morning's featured speaker were richly rewarded by Dr. Thomas C. Chalmers's