

Program Evaluation as Quasi-Experimental Research Bernie Fabry, 12/97

It has been pointed out elsewhere (Meadowcroft, 1997) that program evaluation most often takes the form of program monitoring rather than scientific level experimentation. At this level, it does not attempt to include the very technical and rigorous procedures of experimental control to separate the wheat from the chaff and rule out all threats to the valid inference of cause and effect. But that does not mean that program evaluation is not research at some level. Hawkins and Hursh (1992) have identified three levels of research that progress from simple monitoring of clinical outcomes to scientific quality research. The same three levels can be used as a classification scheme for program evaluation to clarify the distinction between monitoring and research, and to help evaluators identify the incremental steps that move program evaluation activities along a continuum of increasingly more rigorous scientific practice.

Level 1 - Monitoring for Accountability

The first discernable level on a continuum of research involves measuring changes in at least the single most important clinical outcome while providing a specific service, treatment, or educational intervention for a group of clients. This typically involves identifying and defining the desired outcome of treatment, and measuring change over time in some relatively objective fashion while implementing a program of services. In individual-subject research (see for example, Barlow, Hayes, & Nelson, 1984) this would be referred to as a “B only” or “treatment only” research design because measurement occurs only for the intervention phase which typically is the second, or B phase, of a research study. For this level of research the goal is to monitor a program’s outcomes as a form of accountability to the clients, the program staff, and funding sources. From an historical perspective, this has been the most common level of evaluation.

To illustrate, we started our program evaluation project by using a group consensus procedure (described in Fabry, 1997) to identify outcomes that were most important to stakeholders. The programs we were interested in served troubled and troubling kids in out-of-home placements. The stakeholders wanted to know where kids were living, where they were working or going to school, how they played, and how they felt about their lives after leaving our treatment programs. We developed simple, direct definitions for each of these four areas (Fabry, Hawkins and Luster, 1994) and measured our outcomes each year. As an example, one outcome we reported was the percent of kids who were not living in any kind of treatment program a year after leaving our programs. Over eight years of annual evaluations anywhere from 60% to of 77% of our kids were not living in any kind of treatment program a year later. The graph we reported simply showed the trend year after year. By monitoring our outcomes we were able to show that most kids were doing extremely well after leaving our programs.

Level 2 - Semi-Scientific Evaluation

The second level of research includes a comparison of an intervention or treatment with some different condition. In individual-subject research the typical practice is to measure performance during a baseline or “A phase” prior to the start of an intervention or “B phase.” This provides a

comparison of a no-treatment condition with a treatment condition and is referred to as an “AB” research design. Level 2 research also would include a corresponding increase in the rigorousness of the measurement process. This could be achieved by using a standardized measurement tool and/or adding a reliability checking procedure to minimize biasing of data. The purpose of Level 2 research is to persuade others of the effectiveness of a service, education, or treatment strategy relative to some other condition. For instance, we might treat a group of kids and show improvement for each one of them (Level 1 research); but if we also showed that the improvement was significantly better than some comparison service (Level 2 research), our treatment would have greater significance and value.

Level 2 research is actually quite easy to accomplish. To illustrate, many of our kids come to us with a history of psychiatric hospitalizations. When different stakeholders were asked what outcome they most wanted to see for these kids, they voted for minimizing psychiatric hospitalizations for reasons too numerous to list here (in hindsight this seems so obvious, but “back then” it wasn’t - not even to the stakeholders). To measure baseline performance, the official reports of prior placements were reviewed and the number of days in psychiatric hospital settings prior to coming to our program were added up. This was the A phase of a Level 2 research project. Then, to measure performance in the B, or intervention phase, the number of days in psychiatric hospital settings after admission to our program were counted. A graph showed that during the three years prior to admission to our program the number of days in hospital settings was increasing dramatically over time. In contrast, by the end of the first year in our program the days were substantially lower, almost reaching zero.

A replication of this positive outcome was obtained by documenting the same pattern of increasing numbers of days of hospitalization prior to admittance, and a sharp drop after admission for each new kid. While replication of outcomes is a hallmark of Level 3 research, it was relatively simple to achieve in this case.

When working with groups of people, the comparison needed for Level 2 research also can occur between one group receiving an intervention and a different group exposed to something else. This has been referred to as a quasi-experimental research design (Cook & Campbell, 1979). For example, in our evaluation project we were able to compare the percent of kids who had dropped out of school after discharge across each of our programs. In addition, we reported the national drop out rate for kids in grades 7 through 12 as another point of comparison. Of course these are not scientific-quality comparison groups with random assignment of equivalent clients to treatment and no-treatment groups, but the comparisons provide valuable information that becomes cumulative with repetition of the findings. We also increased the rigorousness of our data collection by adding a reliability checking procedure. We used an interview process that required us to ask at least two people to report on each kid.

Another class of Level 2 research that is valuable to program evaluators has been called exploratory data analysis (Tukey, 1977). The approach is described as detective in character. It is a search of data for clues. Exploratory data analysis techniques focus on ways to look at data, ways to pull the data apart and look for relationships and patterns in the data. Its concern is with appearance and detection of possible hypotheses in contrast with the confirmation of hypotheses for Level 3 research.

To illustrate from our own work, we had been reporting the percent of kids living in situations that were less restrictive than our residential treatment programs a year after discharge from those programs. The question arose about whose outcomes were we measuring and reporting. Kids who left our programs could be doing well in some less restrictive program a year later and would be counted as our “successes” by the definition we were using even though they might have left our programs as “unsuccessful discharges.” So we pursued an exploratory analysis of our data. We separated kids according to whether they were successful or unsuccessful discharges from our programs, and then further separated them according to the level of restrictiveness of their living situation (Hawkins, Almeida & Fabry, 1992) one year after discharge. The analysis revealed among other things that some of our unsuccessful discharges were actually doing quite well a year later in non-treatment settings. Overall, we concluded that we needed to include a stricter definition of success in our reports and to be less concerned about whether kids left our programs meeting our criteria of success. In addition to reporting the percent of kids living in less restrictive settings, we then began reporting the percent of kids who were not living in any kind of treatment setting a year after discharge.

This is only a brief description of exploratory data analysis. It can be the most rewarding level of research because it can suggest possible strengths and weaknesses in our programs rather than just monitoring and comparing outcomes.

Level 3 - Scientific-quality research

This is the level of research published in most refereed journals. Typically it includes a number of refinements all designed to eliminate threats to an inference of cause and effect. Classic texts describing this level of research (as well as the approximations defined by Levels 1 and 2) include Barlow, Hayes & Nelson (1984), and Cook and Campbell (1979).

Level 3 research strategies initially may seem too arduous, rigorous and/or compulsive for program evaluators, and may seem to create a formidable gap between the work habits of evaluators in service delivery settings and the work habits of scientists in their tightly controlled settings. However, there are a number of evaluators who have been able to bridge the evaluation-research gap. It is interesting to discover that this typically has occurred through persistence and gradual refinement in the evaluators’ work habits. Many evaluators have started with the ‘keep it simple’ principle in mind and followed Level 1 research practices, then gradually added more rigor and organization by first developing Level 2 practices and then occasionally taking advantage of Level 3 opportunities. To add scientific rigor a bit at a time is not as hard as you might think.

References

- Barlow, D.H., Hayes, S.C., & Nelson, R.O. (1984). The scientist practitioner: Research and accountability in clinical and educational settings. New York: Pergamon.
- Cook, T.D., & Campbell, D.T. (1979). Quasi-Experimentation: Design & analysis issues for field settings. Chicago: Rand McNally College Publishing Co.

Fabry, B. D., Hawkins, R. P., Luster, W. C. (1994). Monitoring outcomes of services to severely disturbed children and youths: An economical follow up procedure for mental health and child care agencies. Journal of Mental Health Administration, 21(3), 271-282.

Hawkins, R.P., Almeida, M.C., & Fabry, B.D. (1992). A scale to measure restrictiveness of living environments for troubled children and youths. Hospital and Community Psychiatry, 43, 54-58.

Hawkins, R.P., & Hursh, D.E. (1992). Levels of research for clinical practice: It isn't as hard as you think. The West Virginia Journal of Psychological Research and Practice, 1, 61-71.

Meadowcroft, P. (1997). Using follow-up measures for outcome monitoring.
<http://www.standardsandoutcomes.com>.

Tukey, J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley Publishing Co.