Increasing External Validity:
Program Evaluation’s Path to Greater Policy Relevance?

Douglas J. Besharov
Douglas M. Call
University of Maryland

During his distinguished career, Peter H. Rossi made many important contributions to the field of program evaluation. Perhaps his most famous (or infamous) were his Metallic Laws of Evaluation, first mentioned in a 1978 paper and then formally published in a 1987 article, “The Iron Law of Evaluation and Other Metallic Rules.”¹ (See box.) The Metallic Laws were meant as an informal (and humorous) overstatement of an exceedingly serious point. As Rossi explained in 2003:

The Iron Law: “The expected value of any net impact assessment of any large-scale social program is zero.”

Stainless Steel Law: “The better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero.”

Brass Law: “The more social programs are designed to change individuals, the more likely the net impact of the program will be zero.”

Zinc Law: “Only those programs that are likely to fail are evaluated.”

At a 2003 APPAM panel in his honor, Rossi lamented that the attention his Laws received was “the source of considerable embarrassment being easily misunderstood and frequently misused.”³ Nevertheless, their underlying message struck a deep nerve in the social policy and evaluation community. His essential message, echoed, for example, in many of the presidential addresses of


³Ibid, 1.
the Association for Public Policy Analysis and Management, was that the results of program evaluation were not helping guide the development or operations of large social programs.

In the fifty years since, program evaluations have grown vastly in number, size, and sophistication. Major advances have been made in estimating the outputs and outcomes of particular projects or individual program sites (“internal validity”). Much less progress, however, has been made estimating systemic or program-wide impacts, which require more generalized measures of effects (“external validity”). And yet, external validity is the second essential quality for an evaluation to be helpful in policy formulation.

The reasons for limited policy relevance are widely appreciated:

- No matter how well-implemented, limited-site RCTs (unlikely to be representative of the entire program) lead to tentative results, at best. Their replications regularly fail, sometimes because of the initial evaluation’s weaknesses but, also, because of unrepresentative sites, populations, or conditions;

- Observational studies that use large data sets may be more representative of a program (and thus have greater external validity), but the trade-off is substantially weakened causal validity; and

- Mixed methods designs may provide additional points of interpretive reference, but they rarely lead to definitive conclusions.

Thus, as Randy Brown said in accepting the 2020 Rossi award for contributions to the field of program evaluation: “While evaluations have shown some programs to be successful over the years, the experience of legions of policy researchers suggests that the essence of Rossi’s Iron Law—that relatively few public social programs are found to be effective—continues to hold true.”

Rossi attributed this continuing failure to identify successful “large-scale social programs” to a failure of “social science engineering.”

The major reason why public social programs fail is that effective programs are difficult to design. Those who typically dominate in designing programs often do not have the social science skills and knowledge needed. Basic social science nonetheless is not advanced enough to provide strong guides to designing effective programs. The consequence is that the designing of social programs has been a kind of trial and error

---

4Randall S. Brown (remarks, Association for Public Policy and Management annual meeting, November 12, 2020).

strategy of try-this-and-try-that with little accumulation of knowledge that might be the basis of social [science] engineering.\textsuperscript{6}

To encourage the “accumulation of knowledge that might be the basis of social [science] engineering,” Rossi conjured “a new profession of social and organizational engineering devoted to the design of human services delivery systems that can deliver treatments with fidelity and effectiveness.” That does not seem to be in the offing.

On the other hand, something should be done. The current lack of sufficiently reliable impact evaluations of large-scale social programs that can be used for social science engineering should not be acceptable. The problem has become more pressing as evaluation results have become more important inputs to policy making and program planning, often called “evidence-based” decision making. Already, we are seeing the weaponization of competing research “findings”---even in the face of false positives and false negatives.

The contributors to this collection believe that a sustained effort to increase the generalizability of evaluation results is an important next step. Their papers explore the issue from separate but related perspectives:

- “Learning about External Validity from Multiple Literatures” (Jeff Smith, University of Wisconsin);
- “Differing Program Environments and Limited Replications Constrain Generalization: The Case of Published Evaluations of School Finance Policies” (Danielle Handel and Eric Hanushek, Stanford University);
- “How Mixed-Methods Research Can Improve the Policy Relevance of Impact Evaluations” (Burt Barnow and Sanjay Pandey, George Washington University);
- “The Logic of Generalization from Systematic Reviews to Policy and Practice” (Julia Littell, Bryn Mawr College); and
- “Improving the Usefulness and Use of Systematic Reviews and Meta-Analyses to Inform Policy and Practice” (Rebecca Maynard, University of Pennsylvania).

\textsuperscript{6}Rossi, “Iron Law, Revisited,” 3.