The “Iron Law of Evaluation” Reconsidered

Peter H. Rossi

University of Massachusetts at Amherst

Presented at 2003 AAPAM Research Conference

Washington, DC

October 2003

When Professor Besharov told me that he was going to propose a session at this APPAM conference that would discuss my old paper on *The Iron Law of Evaluation,* I had very mixed reactions. On the positive side, I was certainly flattered that anyone would pay attention to one of my old published papers first written several decades ago. On the negative side, that paper was far from being one of my favorites. Indeed it has been the source of considerable embarrassment being easily misunderstood and frequently misused. So Besharov's proposal posed the threat of yet another episode of embarrassment. However, it also offered an attractive opportunity to counteract at least some of the misunderstandings generated by the paper (and by me).

I will start out by providing a brief summary of the paper addressed especially to those who may not have ever read it. Then I will provide my account of how the paper came to be written, providing an historical context. In the last section I will discuss what I would now change in the paper in the light of the current status of social program evaluation.

Brief Summary of the 1983 Paper

There are three main themes in my 1983 paper, as follows:

1. The Iron law states that the typical impact assessment of a public social program finds that the program is either ineffective or only marginally effective. The Stainless Steel Law is that better designed evaluations are more likely to yield such findings.
2. 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 furthermore 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 strategy of try-this-and-try-that with little accumulation of knowledge that might be the basis of social engineering.

3. The major sources of program design failures are: (a) incorrect understanding of the social problem being addressed, (b) interventions that are inappropriate, and (c) faulty implementation of the intervention.

Historical Context

The first version of my “Iron Law” paper was presented around 1972 at an American Academy of Arts and Sciences symposium on poverty research chaired by Daniel Patrick Moynihan. The published version was written in 1982 and incorporated somewhat more up-to-date material. The fact that the paper had any readership over the last three decades is largely due to Moynihan’s repeated references to it when he queried (actually badgered) federal agency staff during his two terms in the Senate.

I published the paper in 1983 after tiring of mailing out Xerox copies in response to requests. In the published 1983 version I tried to bring the paper up to date.

Changes over the Last Two Decades

The laws as stated give the false impression that they rest on empirical data. However, I did not undertake anything that might be remotely called empirical research. I had certainly read a rather large number of evaluations reports and articles. I had formed a strong impression at the time that most evaluations done in the decades before 1982 had found programs ineffective.

I still do not have or know of any empirical studies of the outcomes of all program evaluations or any reasonable sample of program evaluations, although there are a fairly large number of meta-analyses that summarize outcomes of evaluations dealing with some specific substantive areas. Based on my continued reading of evaluations, impressions are now different.

There are quite a large number of well conducted impact assessments that yield statistically and substantively significant effect sizes. I believe that we are learning how properly to design and implement interventions that are effective.

An impressive change in the evaluation field is a considerable growth in the sophistication of evaluators and in the methodology of evaluation. The best of evaluators simply know a lot more about how to design credible impact assessments and have at their command technical tools that make it possible to analyze data in much more sophisticated ways.

At the same time, I also believe that the majority of impact assessments end up with findings of no effect or substantively marginal effects. Disappointment with our ability to find that many programs don’t work has led to the formulation of revisionist alternatives to the prevailing canons of mainline evaluation, a topic that really deserves more treatment that I can give here.

As the use of evaluation research has increased over the past half century an evaluation industry has emerged, composed of a set of differentiated sectors. First, there is a relatively small number of “elite” evaluation organizations, such as MDRC, Mathematica, Abt Associates, RTI, and so on, who have the resources, including the skills, to bid successfully on large scale impact assessment evaluations sponsored by the federal government, large foundations, and some states. The evaluations by these elite research firms are generally quite good, with high statistical power and good generalizability. Second, there is a kind of middle level of evaluation run by smaller firms, academic research institutes, and academics. On this level, evaluation quality is variable, and generalizability is usually less. Third, a much larger collection of public agencies, smaller firms, and individual researchers undertake much smaller scale evaluations, usually of much lower quality, based on inferior research designs, generally underpowered and of quite limited generalizability. Many of the evaluations undertaken by this third sector are pro forma, forced upon local agencies and programs by funders asking for “accountability”. At best, they may serve crude monitoring purposes. At worst, they are meaningless and misleading. As impact assessments, these evaluations are generally useless and sometimes harmful.

Although I do not have any firm data, I believe that the overwhelming number of evaluations is carried out by this third sector. As a consequence, we really do not know whether programs over a wide variety of substantive areas are worth anything. Note that this last statement does not mean that they are in fact worthless, but only that we have no evidence on their worth.

My Current Assessment

It should be quite obvious that currently I believe that the Iron and Stainless Steel Laws cannot be taken seriously as originally stated. There are credible evaluations that show some programs to be effective. There are also credible evaluations, perhaps the majority, which show that the programs evaluated have no effect or substantively small effects. However, I have no firm data on the numbers involved.

Given that the majority of impact assessments are conducted by the least competent and least well-funded sector, I believe that we can make the following generalization: **The findings of the majority of evaluations purporting to be impact assessments are not credible.**

They are not credible because they are built upon research designs that cannot be safely used for impact assessments. I believe that in most instances, the fatal design defects are not possible to remedy within the time and budget constraints faced by the evaluator.

I believe that the evidence supporting this generalization mainly can be found in the evaluations conducted in the third sector discussed previously. Indeed, the field of evaluation would be much better off if we lifted requirements to conduct impact assessments placed on organizations which cannot afford and/or do not have the capabilities to conduct them. If anything, we should encourage them to improve program process and outcome monitoring.

As for the other two major themes to be found in my 1983 paper, each is quite relevant in the 21st century. First, it still remains difficult to design effective programs. We still do not have a good sense of what the engineering side of the relevant social sciences should be like or even whether we should encourage its development. I would welcome the formation of an institute or academy that would take as its charge the assessment of the relative effectiveness of various kinds of interventions.

Second, the 1983 paper’s analysis of why programs fail still applies today. The major virtue of that discussion is as a general guide about what to avoid in the design and implementation of programs, a topic that takes up several chapters in the latest edition of our textbook on evaluation written with Mark Lipsey and the late Howard Freeman.

I have no intentions of writing a revised version of my paper. I believe the evaluation field should move on to address more critical issues concerning how best to conduct in a responsible way evaluative activities short of impact assessment which are credible and useful to the policy community and to those who have to manage social programs.