From the Great Society
to
Continuous Improvement Government

The Hopes of an Impatient Incrementalist

Douglas J. Besharov

APPAM Presidential Address

November 7, 2008

American Enterprise Institute for Public Policy Research

Welfare Reform Academy
University of Maryland
American Enterprise Institute
1150 Seventeenth Street, N.W.
Washington, D.C. 20036
www.welfareacademy.org

University of Maryland 1856

Douglas J. Besharov is the Joseph J. and Violet Jacobs Scholar in Social Welfare Studies at the American Enterprise Institute, a professor at the University of Maryland School of Public Policy, and the president of the Association for Public Policy Analysis and Management.

Thanks, Kathy.

I am delighted to be here today and to have had the opportunity to serve APPAM, an organization I respect and care deeply about. Over the years, I've learned a lot from APPAM, from its meetings, from JPAM, and from participating in its governance. So, thank you all very much.

I also want to congratulate president-elect Kathy Swartz, the Program Committee, and our stalwart executive director, Erik Devereux, for a great conference. Thanks for all your hard and very good work.

New Orleans after Katrina (2005)

Many of those in the room (dare I say most?) were cheered by Tuesday's results. After years of questionable policies and catastrophic management, there are, I fear, many hidden Katrinas yet to surface, as evidenced by the current financial crisis. Even those who did not vote for Barack Obama should wish him—and our nation—a successful presidency.

In important respects, my talk today is based on my own hopes for Obama's presidency. (Yes, I did prepare two drafts of this talk, but I spent more time on this one.)

Policy and management intertwined

The program says that my topic today is: "From the Great Society to Continuous Improvement Society," but, as I refined this talk, I realized that I wanted to emphasize the connection between sound policy and strong management. Call it the effect of Katrina on me. So I changed the title to: "From the Great Society to Continuous Improvement Government."

With his mandate—and the large Democratic majorities in Congress—President Obama should have the political running room to engage in a candid appraisal of current domestic programs and take steps to improve them—as he promised in the campaign. (You could liken it to Nixon going to China.) And, judging from his campaign team, he should have the talent pool of appointees needed to do so.

I am, however, worried that an Obama administration may approach government improvement as solely a management issue. Obama, for example, pledged to create a White House "SWAT team" made up of government professionals to review programs for waste and inefficiency. After such reviews, he said, "We will fire government managers who aren't getting results, we will cut funding for programs that are wasting your money, and we will use technology and lessons from the private sector to improve efficiency across every level of government—because we cannot meet 21st century challenges with a 20th century bureaucracy."


Toppling Saddam, Baghdad (2003)

Sound policy matters as much as proper management, which, sadly, is a major lesson from the invasion of Iraq. Most programs cannot simply be managed to better performance; there must also be major changes in how they operate. Learning what changes are needed is the province of policy analysis and evaluation.

Policy and management are, of course, APPAM's two dimensions, and they are connected in many ways and at many levels. In preparing for this talk, I re-read prior presidential addresses and—besides being completely and utterly intimidated—I noticed how most of them, in their own way, grappled with both of these elements of public policy. So, today, I will do the same, but with my focus being on the policy side of program improvement.

And, in doing so, I will concentrate on the need to speed up the process of knowledge building, that is, the process of hypothesis identification and testing.

Here to stay

In a nutshell, here's my argument: In the 1960s, various social programs were started (like Head Start) or dramatically expanded (like AFDC). Loosely, we call that period of expansion the Great Society. Most of these programs are still with us today (although many have changed in important ways) and most address, or at least seek to address, important social needs. (Some programs, like Model Cities, did not make it, of course.)


Mrs. Johnson at a Head Start Center

But too many Great Society programs have been disappointments—at least when compared to the high hopes of the 60s. As many of us in this room have proven all too often, the sad reality is that most domestic programs don't work nearly as well as we would hope, let alone as well as their advocates claim.

Some programs, like Model Cities, did not last, but most of these programs are surely here to stay. How many of us really think there could be an America without a social safety net? Sure, some programs have been trimmed or modified, like AFDC/TANF, but cash welfare is still with us, as are food stamps, Medicaid, child welfare services, and, yes, even job training for troubled youth. The list is long—and, of course, many think it should be longer.


Angora goat

Even if these programs did not serve important social functions, which they do, it would be nearly impossible to dismantle them. Do you remember that 1950s' relic of the Cold War, the mohair subsidy, designed to ensure a source of wool for military uniforms decades after they were made of everything but wool? It was ended in 1995, only to be reinstated four years later through the good work of industry lobbyists. Like this Angora goat, the source of mohair, the subsidy lives on. So, perhaps Model Cities will be back.

Some Great Society programs are not well-designed for contemporary problems, and, in fact, many were probably not the right approach even back in the 1960s. But, forty-plus years later, we should declare that most elements of the Great Society are as permanent as any government programs, and that it is time for us to step up to the plate and do the unglamourous work of program improvement. In the title, I use the phrase "continuous improvement government" not because I like buzzwords, but to emphasize that the effort will be a step-by-step process—that combines policy and management tools.

Head Start

Head Start, considered by many the gem of the Great Society, is a prime example of the need to improve an ongoing program. Since that first summer of 1965, about twenty-five million children have passed through the program, at a total cost of about $145 billion, and yet we are still arguing about whether Head Start "works." I've contributed some to this argument, so I know it well.


Romulus and Remus

Of course it matters how children are raised. Romulus and Remus were suckled by a wolf, and they founded a city that became a great empire. The rest of us, though, need much more care and nurturing to reach our full potential. The real policy question is not whether there was a proper random assignment of those 123 Perry Preschool children back in 1962, but, rather: How much can government programs do to compensate for family and community deficits?

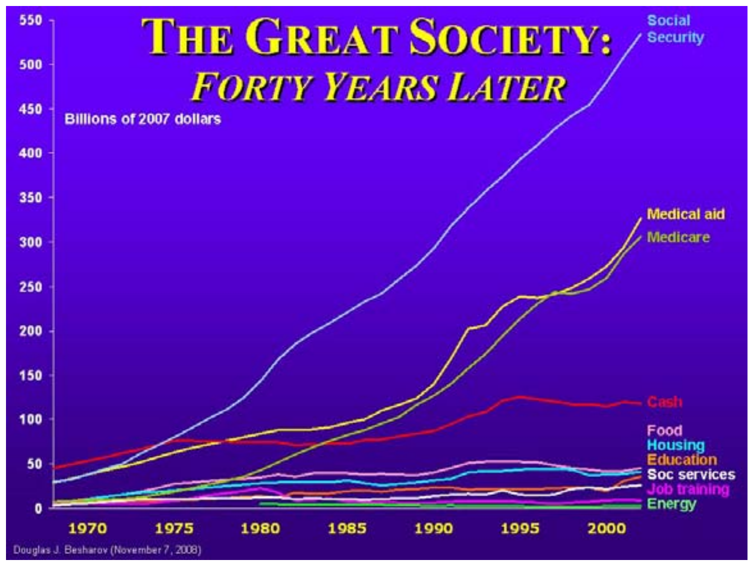
Spurred by a 1997 U.S. Government Accountability Office report concluding that there was *"insufficient" research to determine* Head Start's *impact*, in 1998, Congress required the U.S. Department of Health and Human Services to conduct the first rigorous national evaluation of Head Start. To its credit, the Clinton Administration took this mandate seriously and initiated a 383-site randomized experiment involving about 4,600 children. (In fact, throughout his presidency, Bill Clinton and his appointees were strongly supportive of efforts to improve Head Start, even to the point of defunding especially mismanaged local programs.)

Confirming the findings of earlier, smaller evaluations, the Head Start Impact Study (released in June 2005) found that the current Head Start program has little meaningful impact on low-income children. For example, even after spending about six months in Head Start, 4-year-olds could identify only two more letters than those who were not in the program, and 3-year olds could identify one and one-half more letters.

There is, as some of you know, an argument about how to label the effect sizes involved. But I think that argument loses sight of the larger disappointment involved. No *socially* significant gains were detected on a host of measures. Surely reasonable people on both sides of the political spectrum can agree that these outcomes are simply not good enough—even for government work.

Insufficient funding

Many observers say the problem is insufficient funds. Money is certainly an issue. Just a cursory look at this graph tells the story: spending on Social Security and medical entitlements (Medicare, Medicaid, and all other forms of medical aid) is way up, as is spending on cash programs (many of which are also entitlements). But spending on service- or treatment-oriented programs has been comparatively flat for three decades.



For years, the entitlement crisis (and it should be called that)—coupled with repeated tax cuts—has been eroding discretionary spending at all levels of government. Across a whole range of activities, the political pressure to feed Social Security, Medicare, and Medicaid (many would want to add defense spending to this list, even though it has been a declining percentage of GDP) has undercut government's ability to think, to plan, and to do. A government that underinvests in maintaining bridges and in the SEC's oversight of financial institutions is probably doing a pretty bad job maintaining social welfare services for the disadvantaged, let alone improving and expanding them.

So, more money would undoubtedly help, assuming it was spent wisely. A giant assumption, though. Head Start already costs about 50 percent more than high quality, state-run pre-K programs—with much poorer results.

Improvements made

As many of you know, I start with a preference for small government; I think that government too often tries to do what it cannot do successfully. But I have little doubt that most domestic programs could be vastly improved. The question is: How?

I read the evidence to say that government programs are most likely to improve through a series of ongoing adjustments in how they operate. A few of these changes will be large, but most will be small. If that makes me an incrementalist, so be it.

I am, however, an impatient incrementalist. Sure, progress is being made in many fields, but from the point of view of those being served by these programs, it is excruciatingly slow. I cringe when responsible researchers say that it may take a generation to get things right. Most of us will be retired by then, and some of us will be dead. We need to change how we do business.

There are many reasons for weak or ineffective programs. I will focus on only one: the slow pace of knowledge accretion and program development.

One of the truly great accomplishments of modern social science has been the use of the randomized experiment to address important policy questions. For good reason, it is widely referred to as the “gold standard.” In program area after program area, well-planned and well-implemented randomized experiments have made major contributions to public policy.

In recent decades, such APPAM stalwarts as Abt, MDRC, and MPR (often funded by HHS), more lately joined by the Department of Education’s Institute for Education Sciences, have shown time and again that rigorous social experimentation is both possible and fruitful.

There were many other players. For example, the RAND Health Insurance Experiment (HIE) established that increased cost sharing by clients led to reduced medical care usage without any widespread effect on health status. Although there remains some disagreement about the methodology and implications of the findings (but isn’t there always?), most of us are members of health insurance plans shaped by the RAND findings.

These very real successes, however, should not obscure how much is left to do.

Knowledge building is too slow

Our current approach to R&D is too slow, too haphazard, and too often fails to factor in the dead ends that are inevitable in policy research and program development. For the past year and a half, I have had the privilege of being the editor of JPAM’s Policy Retrospectives section, and I have seen this glacial slowness in area after area.

It took, for example, more than seven years (ten years if you include when the thirty-month impacts were released) for Abt's very fine evaluation of the Job Training Partnership Act (JTPA) to conclude that the program failed to accomplish many of its goals. By the time the results were released, the JTPA program had been changed considerably, in an attempt to improve its effectiveness. But none of these changes, including the creation of a separate youth program and targeted services to those with multiple employment barriers, were assessed by Abt before the program was terminated.

Ten years later, we are only now beginning an evaluation of JTPA's replacement, the Workforce Investment Act. Final results are not scheduled for release until 2015—six years from now. That's half-way through Barack Obama's second term, assuming that there is one. I ask you, will he wait until then before deciding whether to put more money in the program or to radically restructure it?

We urgently need to speed up and expand the processes of program design and testing, program implementation and evaluation, and, when necessary, program redesign, as the process continually repeats itself over time.

Find more promising ideas to test

Far too many evaluations, however, obey Pete Rossi's "Iron Law of Evaluation," namely, that: "The expected value of any net impact assessment of any large scale social program is zero."

Sometimes, the experiment is underpowered, or poorly designed or poorly implemented. There's often a more fundamental problem, though, which Pete, a college-aged Trotskyite turned social liberal (no Neocon, he), hated to acknowledge: Sometimes the program being tested is simply a bad idea.

Frequently, the political and administrative process that leads to some research designs seems, well, wrong-headed. Consider, for example, the Comprehensive Child Development Program (CCDP), meant to test the effect of well-coordinated services and parental education on the growth and development of young children.

Hopes were high for this $300 million program that served low-income, single mothers for as long as five years. It spent about $19,000 per family per year (that's on top of AFDC, food stamps, WIC, and other safety-net programs), and about $58,000 total per family. But a closer look at the project design suggests that it never had a good chance of succeeding:

- Program sites were all but prohibited from funding their own services—because this was a test of the impact of using free services from existing community sources,

- Center-based child care was not authorized unless the mother was working—because the program sought to teach parents how to educate their children, and

- The sites were also prohibited from changing their approach even as their experience suggested that a mid-course correction was urgently needed—because there was to be no deviation from the planned intervention.

I could go on.

When Abt announced that the program had had no positive impact on the young mothers or their children, many people concluded that these mothers were beyond the reach of our programs, rather than that the CCDP was a bum intervention. (Of course, some advocates just blamed the evaluators, but that's nothing new, either.)

Was the $300 million spent on the CCDP a waste of money? You betcha.

I don't mean to single out the CCDP for special criticism. There are many other examples. It just happens to be an experiment that I lived through, and have the scars to show for it.

At the risk of offending, let me say that, as a field, we are not very good at coming up with good ideas to test. It's one thing to advocate for a new "program" to combat a particular serious social problem. It's quite another to specify what particular elements should be in the program. Truth be told, in many areas, we suffer a dearth of good ideas that can be appropriately tested.

I don't want to exaggerate. There are many untested ideas deserving of serious examination. But, really, what are the new approaches to reducing teen pregnancy that should be tested? To job retention for welfare leavers? To helping troubled youth? To making schools "work"?

The chances are small that some government or foundation committee will come up with a program that will work much better than all the other programs that have gone before. For example, we could decide that a major cause of teen pregnancy is unprotected sex, and we could then decide that there should be a program to encourage safe sex. But we'd have great difficulty coming up with reasonable approaches to doing so—that have not already been tried, and been found wanting.

Innovators and outliers

We should be looking—systematically—for successful program innovators—or outliers (as I will describe in a moment)—and then try to learn what they are doing that seems to work.

That's essentially what happened in the welfare reform experiments of the 1980s and 1990s. In a multi-step process, MDRC first found that the labor force attachment strategies followed in Riverside County, California (mostly job search and diversion), stood out as outliers (in earnings increases and caseload declines) compared to other programs in the GAIN experiment. Then, in a series of randomized experiments, MDRC found that Riverside-like approaches outperformed human capital strategies. Other random assignment studies confirmed this finding, as did simple pre/post analyses.

And, of course, these kinds of "work first" strategies now characterize most welfare programs.

Call it the bottom-up generation of ideas: a process that identifies promising ideas from the frontline that higher-level planners might never have imagined. Before the Riverside findings, few experts proposed work first strategies (some exceptions are in this room), and instead emphasized job training, mandatory work experience, and time limits.

The learning can go beyond just statistics. MDRC staff noted that the Riverside staff rang a bell whenever a welfare recipient found a job. When the bell rang, all staff would stop what they were doing and would clap. Recipients meeting with staff would usually be told what was going on, in the hope that this would "send a signal to all clients in the office about a key goal of the program."

Welfare reform tools: bell and incense

On the other hand, the job club leaders in another welfare office burned incense in order to get in the mood to serve clients. Don't ask…

So you need to be somewhat selective in the ideas you pursue. (An important point that I will return to in a moment.)

Recognizing this dearth of testable, good ideas, recent R&D planning efforts have actively sought to develop new ideas—usually by reaching out to a broad array of experts and program operators, often using a snowball technique.

This is a good, but, I think, insufficient process. For it depends on new program ideas having been noticed by those being interviewed, and on their ability to identify what works by more than appearances and reputation. We need to expand and sharpen the process of finding promising programmatic ideas. The search for good ideas must be much more systematic—wholesale rather than retail.

In this regard, I have high hopes for the development of outcome-oriented performance management systems—like that of the Workforce Investment Act—that are capable of monitoring not just program outputs (like clients trained) but also short-term outcomes or impacts (like clients employed after a certain period of time).

Key to this approach is the fact that, in most programs, there is a fair degree of deviation from the mean. It's true, some Head Start centers are really good. To the extent that such performance management systems could identify outliers, they could become engines of "continuous program improvement." (Such performance management systems need not be national. Statewide and even local systems have been successfully used to identify outliers.)

Up to now, the tendency has been to use performance management systems to identify the outliers on the left hand of the distribution—and then work to either improve or defund them. That kind of high stakes management has not made these systems popular with most service providers. Who likes to be judged, especially if the yardstick seems unfair?

Such systems, however, can—and, increasingly, are—also being used to identify outliers on the right hand of the distribution. These outliers should then be studied to see what it is about them that seems to work better than average.

Flexibility to innovate

Essential to the bottom-up generation of ideas is the ability of individual programs to innovate, or at least, to do things a little differently from the other guys. Too many programs, however, are straightjacketed by rules and regulations—usually to control program costs and sometimes to protect the program from the alleged predations of conservative administrations.


Egg McMuffin

The original Egg McMuffin was a violation of McDonald's rigid rules about only serving lunches and dinners from pre-approved menus and precise recipes. It was developed surreptitiously by one franchisee—who then tricked Ray Kroc, legendary president of McDonald's, into tasting it. What if Kroc had refused to taste it?

Within reason (and with all appropriate safeguards), program flexibility should be encouraged. In the coming years, I hope that the Obama administration can revisit the issue of waivers, which proved such a powerful tool in welfare reform. Even for the Obama team, though, that won't be easy.

Through much of the period since the Johnson presidency, attempts to get programs to cooperate in efforts to evaluate and improve them have been stymied by the unfriendly political atmosphere of recent decades. Putting aside the danger that an outside evaluation might get it wrong (a real problem, we must acknowledge), they rightly feared that any negative findings could be used to defund the program (as happened to the JTPA youth program after the evaluation). Hence, for too many years and for too many programs, there has been an entirely understandable tendency to defensively circle the wagons.

In 2003, for example, the Bush administration proposed an eight-state waiver experiment that would have explored different approaches to integrating Head Start into the wider world of child care by giving states control over Head Start funds. Even with the most stringent controls, the Republican congress refused to approve even this limited experiment, because of vociferous opposition from the Head Start lobby and its allies. In one particularly colorful phrase, Congressman George Miller (D-CA) said that handing control of Head Start over to states was "like handing your children over to Michael Jackson."

In effect, they wouldn't taste the Egg McMuffin.

Why was there so much opposition to an experiment in even a few states—unless they feared that the experiment would be successful, and would demonstrate a better model for providing early education to disadvantaged children?

Perhaps, just perhaps, the Head Start community and the Democratic Congress will trust a President Obama more than they have trusted President Bush. I wouldn't count on it, though. Just remember what happened when Jimmy Carter suggested moving Head Start to the new Department of Education. Ah, the Congress…

Attributing causation

As the smoking incense reminds us, the process of searching for good ideas must be selective. Identifying what seems to be a promising approach is only step one. Appearances can be deceiving.

Next comes the challenging task of attributing causality to any apparent programmatic differences. This is not easy, and many prefer to go with their own, often ideologically tinged, reactions. Think of the controversies over "schools that work." People see what seems to be a successful school, but there is no way to tell how much (or even whether) the school itself is contributing to the success of students.

Statistical techniques can often establish that the program is providing a "value added" (to a sufficient level of approximation, at least), but more often than impatient advocates and policy makers would like, a definitive determination require rigorous and, I am afraid, time-consuming experimentation.

Learning from failure

That first Egg McMuffin that Kroc tasted was the result of almost one year's worth of recipe testing—which brings me back to the process of testing program or service ideas. The tendency to test only one program idea at a time (as happened in the JTPA and CCDP evaluations) elongates the learning process from years to decades—as we test and fail, and test and fail again.

R&D strategies should be planned with failure in mind, and they should be structured so lessons can be learned from those failures. But can you imagine telling a foundation official or political appointee that you really don't think this idea has more than a 50/50 chance of working? (The true figure, of course, is much lower.)

We should do more testing of multiple ideas at once, perhaps through planned variation experiments. They test more than one idea at a time—thereby increasing the likelihood of finding positive impacts in a shorter period of time.

Consider the Department of Education's recent random assignment experiment to study the effectiveness of sixteen (yes, sixteen) educational software products. In this congressionally mandated study, a group of experts selected the sixteen software products on prior evidence of effectiveness. Then, Mathematica Policy Research and SRI tested the products in 33 school districts and 132 schools, with 439 teachers participating in the study. Within each school, teachers were randomly assigned to a treatment group, which used the study product, or to a control group, where teachers were to teach reading or math the way they otherwise would have.

After one year, there were no statistically significant differences in test scores between classrooms using the selected reading and mathematics software products and those in the control group. These are unhappy results, but imagine where we would be if only one or a few products had been tested.

Not all programs or questions lend themselves to planned variation, and the practical and cost requirements of mounting a successful planned variation have, no doubt, discouraged more attempts. But I hope we see more planned variations within existing programs, as it becomes progressively more difficult to isolate a meaningful, zero services control group. In 2003, for example, about 25 percent of the control group in the Head Start Impact Study was in some other form of center-based care.

Diversity of methods

Finally, after years of extolling the virtues of randomized experiments over other forms of evaluation, it's time to be more explicit about their frequently serious limitations, including limited generalizability, inability to capture community effects, contamination and substitution, randomization bias, only testing the intent to treat, high cost, and so forth. (Did I mention that they, as currently conducted, they often take a long time?)

Randomized experiments may be the "gold standard," but as the last few months have reminded us all too well, diversification can be crucial for long-term returns. Judy Gueron made this very important point earlier today. Hence, before closing, I would like to recognize the growing contribution of nonexperimental methods in the policy process.

Charles Atlas advertisement

For many questions, a randomized experiment may be the only way to approach a definitive answer. Sometimes, however, even a simple before and after demonstration works just fine, as Charles Atlas reminded countless scrawny boys in the 1950s.

Recent years have seen an explosion of more sophisticated nonexperimental work, much of it as helpful as the average randomized experiment (and sometimes more so)—especially since it usually is much more timely. In the right hands, much can be learned from a growing number of nonexperimental techniques. Think about Brian Jacob's fine work that was presented in yesterday's Kershaw Lecture.

In the future, we can expect significant payoffs as analysts use nonexperimental techniques to plumb evolving performance management systems, such as that of the Workforce Investment Act. The Department of Labor has funded Carolyn Heinrich of the La Follette School of Public Affairs and her colleagues to use propensity scoring to explore four-year employment and earnings impacts for early cohorts of WIA participants. Their report should be available in December, 2008—as opposed to 2015 for the randomized experiment, and at much less cost, too.

If used responsibly, nonexperimental approaches can hasten and enrich the development of knowledge. I apologize for the qualifier, but nonexperimental evaluations can raise special worries. Replication is usually difficult, and the analysis is often less than transparent. Put bluntly, the low barriers to entry (all one really needs is a laptop, a data set, and a rudimentary knowledge of statistical analysis) invite mischief: An advocate group can issue a report (or at least a press release), and be in the newspapers and on the internet—without the traditional protections of peer review, etc.

As time goes on, I hope that we will develop more tools for assessing the trade-off between the slower and more precise results of randomized experiments and quicker but perhaps less precise results of nonexperimental methods. In this regard, and most recently, I commend to your attention a recent JPAM article by Cook, Shadish, and Wong: “Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons.”

No stirring call to action

I wish I could close this talk with the usual call for action by President Obama and congress. But the odds against major changes are simply too great.

Most programs would benefit from substantial increases in their R&D budgets, especially since many of these budgets have failed to keep up with inflation. Given the current economic situation, however, the budget battles are going to be fierce, and we in APPAM will be only bit players. And it’s not as if the program advocates will be pounding on the administration’s doors urging more evaluations. After all, they already know that their favorite programs work. Why risk a damaging evaluation?

On the other hand, we live in unprecedented times. Barack Obama will enter office with gargantuan deficits. But the financial emergency we also face will likely be used to justify substantial increases in social spending. If major expansions are in the offing, I hope that we will pound on the administration’s door to have the expansion take place in a way that systematically tests possible program reforms. For example, an expansion of job training programs could randomize different program elements in a way that allows them to be tested against the existing program.

Finally, based on Obama’s campaign pronouncements, I expect the new administration to make changes in OMB’s Program Assessment Rating Tool (PART). Continuous improvement government requires the development of learning organizations, in which managers are rewarded for figuring out what isn’t working and for trying something else. Such a process must be driven by a force outside individual programs.

PART has been controversial in many ways, but it does seem to provide the framework for encouraging rigorous R&D efforts—government wide. I hope that the new administration will build on its strengths and correct its weaknesses.

In closing, I console myself—and, I hope, you— with three upbeat thoughts.

- One, although slow, progress is being made in building knowledge about numerous domestic programs, as I have tried to highlight in this talk;

- Two, we in APPAM are major players in this evolving and expanding mosaic of program evaluation;

- And, three, we have learned enough since Rossi first propounded his Iron Law of Evaluation to offer a friendly update.

As I suggested earlier, Pete was always ambivalent about his Iron Law. He was human, so he enjoyed the attention it (and he) received, and he considered the objective, scientific method that it reflected to be one of the highest forms of social science. Nevertheless, as an unrepentant and proud liberal, Pete was deeply saddened by the disappointing results of so many evaluations, and the fact that evaluation had become a profoundly conservative force.

So, with what I hope is due modesty, but also based on my years of friendship and discussions with that great bear of a man, I would like to close by taking the liberty of offering a friendly amendment to Pete's Iron Law of Evaluation:

> "The expected value of any net impact assessment of any large scale social program is zero…
>
> …unless it systematically assesses the impact of services provided by innovators and outliers in comparison to those of other providers."

I think he would have approved.

Thank you.