2020 Rossi Award Lecture*: The Evolving Art of Program Evaluation

Running header: Evolving art of program evaluation

Author information: Randall S. Brown; PhD Economics, University of Wisconsin. Senior Fellow Emeritus at Mathematica, and Consulting with Randy, LLC, Sandown, NH. Randybrown1119@gmail.com. Address correspondence to Randall Brown, 32 Holts Point Road, Sandown, NH 03873. Telephone: 603-347-6987.

Acknowledgements: This paper benefited greatly from valuable comments and insights from Douglas Besharov, Jacob Alex Klerman, Nancy McCall, Deborah Peikes, Greg Peterson, and Craig Thornton. Superb colleagues at Mathematica over the past 42 years made comments, questions, and critiques that continually improved my work and inspired many of the ideas in this paper. In addition, I thank Kieron Dey at The Nobi Group for introducing me many years ago to experimental designs and how to convince providers to participate in them, and to Mariel Finucane for her tutoring on Bayesian models. I also thank Renee Mentnech and Timothy Day at the Centers for Medicare & Medicaid Services (CMS) for their consistent openness to new ways of conducting analyses. Their support and encouragement made it possible to explore these approaches, and their thoughtful questions and comments on the many evaluations we conducted for CMS inspired me to think of new and better ways to address them. Any shortcomings in this paper are attributable solely to me.

* The Peter H. Rossi Award is given every other year to honor the lifetime achievements of Peter Rossi by recognizing important contributions to the theory or practice of program evaluation. This paper is an edited version of the acceptance remarks of the 2020 awardee, Dr. Randall S. Brown, delivered on November 12, 2020, at the virtual 2020 APPAM Fall Research Conference.

Abstract

Background. Evaluation of public programs has undergone many changes over the past four decades since Peter Rossi coined his “Iron Law” of program evaluation: “The expected value of any net impact assessment of any large-scale social program is zero”. While that assessment may be overstated, the essence still holds. The successes far outnumber the failures, and the estimated favorable effects are rarely sizeable.

Findings. Despite this grim assessment, much can be learned from “failed” experiments, and from ones that are successful in only some sites or subgroups. Advances in study design, statistical models, data, and how inferences are drawn from estimates have substantially improved our analyses and will continue to do so. However, the most actual learning about “what works” (and why, when, and where) is likely to come from gathering more detailed and comprehensive data on how the intervention was implemented and attempting to link that data to estimated impacts. Researchers need detailed data on the target population served, the content of the intervention, and the process by which it is delivered to participating service providers and individuals. Two examples presented here illustrate how researchers drew useful broader lessons from impact estimates for a set of related programs.

Conclusion. Rossi posited three reasons most interventions fail—wrong question, wrong intervention, poor implementation. Speeding the accumulation of wisdom about how social programs can best help vulnerable populations will require that researchers work closely with program funders, developers, operators, and participants to gather and interpret these detailed data about program implementation.

Key words: Evaluation, implementation, research design

2020 Rossi Award Lecture: The Evolving Art of Program Evaluation

INTRODUCTION

Program evaluation has improved in many ways over the past 4 decades, since Peter Rossi laid out his Iron Law of Program evaluation in 1983 (eventually published in Rossi 1987). The Iron Law states that "The expected value of any net impact assessment of any large-scale social program is zero." He follows that with three other metallic laws of evaluation:

- 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."

Rossi's paper also made 2 other key points; first, he notes that designing effective programs is difficult, and exacerbated by how poorly we accumulate knowledge. Second, Rossi identified the main reasons for a program's failure to have effects as misunderstanding the problem, designing the wrong intervention, or poorly implementing a good intervention.

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. This paper focuses on how evaluators can add the most value to public policy decisions, despite the grim prognosis of the Iron Law. It first briefly touches on some of the major changes in program evaluation occurring since the mid-1970s. Attention then shifts to changes in intervention design, context, and the environment in which programs operate, and how they increase the need for stronger implementation analysis.

The paper stresses the importance of understanding the details of a program and how it is implemented, regardless of whether the estimates suggest that it has favorable impacts overall. I believe this area has the largest payoff in terms of learning, which is the true purpose of evaluation research. The task is difficult—much harder than producing valid impact estimates—but essential if we are to maximize the value of large public investments in social programs. Without it, we will continue to fail to accumulate the knowledge needed to improve interventions designed to help the poorest, most vulnerable members of our society. Throughout the paper, I link the discussion to Rossi's Metallic Laws and the changes that can help address the problems he raised.

SOME MAJOR CHANGES IN PROGRAM EVALUATION

Program evaluation has changed substantially in four categories since the 1970s: evaluation design, statistical methods, data sources, and interpretation of findings. Below I briefly touch on some of the major changes; this discussion is in no way intended to be comprehensive or to resolve any disputes within the evaluation community about them. The point is just to acknowledge some of the major changes in these aspects of program evaluation before turning to the changes on which this paper focuses.

Evaluation design. While debate will continue about the relative importance of the pros and cons of randomized trials versus comparison group designs, both have improved considerably. Randomized trials are now simpler to implement and can be tweaked during the study to increase the proportion randomized to the treatment group if accumulating data show the treatment group has significantly better outcomes (Luce et al. 2009). When feasible, implemented carefully, and with adequate power, a practical randomized trial has the major advantage of producing unbiased estimates of program impacts for those in the study sample, with a known degree of statistical precision. However, as Metcalf and Thornton (1992), Deaton (2020), Deaton and Cartwright (2018) and many others have pointed out, if people are required to consent to participate in the trial, the generalizability of the results is questionable. Comparison group designs can yield more generalizable results in some cases but produce biased and misleading estimates if selection bias cannot be accounted for adequately in the design or estimation models (consistent with Rossi's Stainless-Steel Law that the stronger the design, the less likely one is to find statistically significant effects[1]). Nonetheless, it has become clear that very good evaluations can be produced from comparison group designs in many cases. Among the most compelling methods are regression discontinuity designs, when this is possible, and the many new advances in propensity score models. Even when randomized designs are feasible, which of these designs predominates differs across federal agencies sponsoring research and varies over time within agencies.

Statistical methods to estimate effects. Early evaluations used conventional regressions to estimate program effects. These early models often used fixed effects or random effects models to account for non-independent observations (observations on the same individual in multiple time periods, or clustered designs in which the intervention is implemented for groups of individuals, such as schools, geographic regions, or physician practices). In the 1980s,

[1] If the treatment group includes self-selected and/or program selected individuals who are most likely to benefit from the program (or to have better outcomes regardless of the intervention), and similar individuals cannot be readily identified for a comparison group, impact estimates will be biased toward a favorable result.

researchers began to use instrumental variable models and models to address selection bias developed by Heckman (1979) and others to estimate program impacts for studies that were not randomized trials. These methods attempt to account for potential unobserved differences between those who chose (or were selected) to receive the intervention and the comparison group.

In recent years, as computing capabilities have increased exponentially, researchers have begun to use Bayesian models to estimate program effects (e.g., Vollmer et al. 2018). This is an exciting development, for three reasons. First, these models can substantially improve the precision of our estimates by allowing us to "borrow strength" from related external studies and internal estimates (such as across subgroups or time periods). By including as the "prior" the assumption that programs have an expected impact of zero (Rossi's Iron Law) and informed estimates of the variance of true impacts, Bayesian models will shrink new estimates that are implausibly large, while reducing the variance of the estimate. Second, Bayesian models allow us to draw the nuanced inferences that policymakers need to be able to draw that are not possible with our conventional "frequentist" methods. Frequentist tests of the null hypothesis of no program effects simply indicates whether the treatment group's mean outcomes differed from the comparison group's by more than one might reasonably expect to observe by chance. Bayesian models, in contrast, allow us to estimate the probability that the true program impact exceeds a specified policy-relevant value (such as the cost of the intervention). Third, Bayesian models allow the researcher to reduce the probability that some estimated impacts among the many that may be produced by an evaluation will be statistically significant simply by chance. This building-in of penalties for multiple tests is important (Gelman, Hill, and Yajima 2012) given how many hypotheses are typically conducted in evaluations (due to the multiple outcome measures, subgroups, and time periods examined).

More recently, researchers are exploring new data mining techniques developed for big data bases and nonparametric models that offer promise for estimating program impacts that vary with characteristics of the participants (see Hahn, Murray, and Carvalho 2020). This new methodology, Bayesian Causal Forest analysis, combines regression trees with Bayesian prior distributions to estimate the relationships between covariates and the outcome variable in a highly flexible way, without overfitting. Whereas "data mining" once was criticized by researchers, it now may be a virtue. As in all evaluation methods, what matters is *how* the data mining is done and interpreted.

Data sources and availability. The changes in how we analyze the data have been accompanied by major changes in the data sources that we analyze. Early program evaluations relied heavily on in-person survey data to gather much of the information needed. These data were complemented by administrative data on measures such as program participation, employment and earnings, test scores, and health care utilization. While both survey and administrative data are still used, the expense of in-person surveys led most studies to shift to telephone surveys in the 1980s and 90s. But in the last decade, telephone survey response rates have dropped precipitously, as marketers and advocacy groups have inundated the population with phone calls, creating concerns about the representativeness of our surveys. The polling errors in the 2020 election provide a good illustration of the potential biases in estimates drawn from telephone surveys. On the other hand, administrative data have become more widely available, with electronic health records and new databases being made available (e.g., cellphone data for tracking compliance with stay-at-home orders or wastewater to detect opioid use or coronavirus).

What has remained the same is the need to assess the validity, completeness, and reliability of an evaluation's data sources. As the apt saying goes: Garbage in, garbage out.

How results are interpreted. Finally, one of the most important changes in recent years is in the understanding of how researchers should interpret the findings and hypothesis tests produced by a study. In the early years of program evaluation, hypothesis tests were rigidly conducted. If an estimate was not statistically significant at the predetermined significance level (typically p < .05), one essentially had to dismiss it, effectively saying "we don't know" whether the program was effective or not, since one can't prove the null hypothesis. That type of conclusion is not helpful to most policymakers. Even worse, however, is concluding that the program was ineffective. Unless a study has a high level of statistical power to detect small impacts, we simply cannot conclude from a frequentist estimate and hypothesis test that a program was ineffective.

Journal articles with this misuse of p-values abound. But many researchers, with the American Statistical Association leading the way, now realize that this is not the way results should be interpreted. All that we can conclude from a p-value larger than our .05 threshold is that there is a non-negligible likelihood that the observed treatment-comparison difference in mean outcomes is due solely to chance. A p-value of .10 does NOT mean that there is only a 10 percent chance that the observed difference is due to chance rather than to the intervention. It means only that if a policymaker randomly selected two groups from the same population many times and compared their mean outcomes, one time in 10 on average the observed difference would exceed the value you obtained. It does not imply that the program's effect is zero.

The focus on p-values is justly intended to make us be conservative in our decisions about whether a program works but can come at a high price. Programs that do in fact improve outcomes but are evaluated in underpowered studies may be rejected because the evidence does not meet the high bar for concluding that they worked. This type of mistaken inference, Type II errors, can be extremely costly, given how difficult it is to design effective interventions (as Rossi noted).

The best way to reduce the likelihood of Type II errors while being conservative in our decisions is to increase the sample size for the study, and program funders should strive to do so. However, since such increases, even if feasible, may still be insufficient for many worthwhile studies, we need to draw on three factors: the theory of change behind the intervention, the patterns of estimated effects across *all* of the statistical evidence that we generate, and the findings from a strong implementation analysis. Only by using all these key pieces of information are we likely to make accurate inferences about a program's effectiveness.[2]

Intervention design and context. The changes in program evaluation described above are often under the evaluator's control to some degree, but two other factors that greatly affect evaluations are not—changes in the nature of interventions, and changes in the environment into which interventions are introduced. Two of the most prominent of the early social experiments (described below) changed financial parameters and incentives for individuals and measured their responses. Today, however, most public policy interventions either dictate specific changes in service providers' behavior (sometimes in addition to providing funds) or use a "market-based" approach in which providers can choose their own approach and are rewarded based on

[2] The theory of change can provide valuable insights about whether impact estimates are credible. For example, consider an intervention designed to reduce hospitalizations and costs by increasing primary care visits. If we find a sizable but statistically insignificant effect on hospitalizations, but a significant reduction in total expenditures (after accounting for possible effects of outliers), the program may well have impacted hospitalizations as well as costs. However, if the estimated favorable effect on costs is solely through a large reduction in (say) expenditures on office visits, it's likely that the estimated impact on costs is due to a poor comparison group or statistical anomaly.

outcomes. The other big difference from early experiments is that the environment in which interventions are implemented changes constantly. For example, the number of overlapping alternative interventions at the local, state, and federal levels has grown exponentially over the past decades, to the point where it can be difficult to find an "uncontaminated" area from which to draw a comparison group. That is, these comparison areas in which tests of other experimental interventions are simultaneously underway do not represent the "business as usual" counterfactual that we seek. These and other environmental factors can affect outcomes, for both the treatment and comparison groups.

Correctly interpreting evaluation findings thus requires extensive knowledge of these contextual factors. This need underscores the importance of having a strong implementation analysis. Throughout this discussion, I will cite examples from health care research, since that is the area where I have conducted most of my research over the past 4 decades. Evaluators of education, employment, and other public programs will readily identify parallels in those areas.

THE IMPORTANCE OF UNDERSTANDING AND DOCUMENTING PROGRAM OPERATIONS AND CONTEXTS

The two major innovative public policy programs in the 1970s were the Negative Income Tax Experiment (NIT) and the RAND Health Insurance Experiment. Both were large, well-designed randomized trials, and both offered new financial parameters and incentives intended to affect behavior. Under the Negative Income Tax, people in the treatment/intervention group received a guaranteed income with a reduction of 50 cents (or some other fraction) for each dollar earned (see Kershaw and Fair, 1976, for a description of the treatment variations used). The goal of the program was to determine how this change in these financial parameters affected participants' work effort and income.

The second major program was the RAND Health Insurance Experiment (Brook et al., 1984), under which participants were randomly assigned to health insurance plans with different levels of cost-sharing. Like the NIT, the goal was to see how participants' behavior (care-seeking and health care expenditures) varied with the proportion of health care costs that their insurance covered. Evaluations of these programs produced a wealth of important information. As Rossi pointed out, the NIT avoided his Brass Law because it was not attempting to *change* individuals' behavior, but rather simply to *observe* how their behavior differed with the strength of different financial incentives. The RAND experiment shared this same feature.

While those programs had challenges of their own, today's interventions are more complex to evaluate. A typical intervention involves specifying changes in how certain services are delivered, not simply changes in financial parameters. For example, does educating students in charter schools rather than public schools yield better educational outcomes? Do new job training approaches for young workers increase their employment and earnings? Does increasing care management staff in primary care practices reduce patients' need for expensive and invasive health care services? Changes in service delivery such as these can be implemented in many ways and intensity levels that can greatly affect the impacts the program achieves.[3] An additional concern is that innovative programs being tested may attract mostly participants (such

[3] In practice (as pointed out by Jacob Klerman in his comments on an earlier draft), if the program were expanded, such variation in implementation would exist, and fidelity to the intervention is likely to be even lower than in the experiment. While that is certainly true, the point I am making here is that by identifying what leads to the variation observed in fidelity and intensity, we can look for ways to adapt the intervention to reduce these barriers to effectiveness.

as school districts or physician practices) that are already practicing reasonable facsimiles to the prescribed intervention. How much new "intervention" is actually occurring?

Non-replicability is a major problem. These variations in how interventions are delivered often lead to non-replicability of favorable effects found in an initial evaluation of a program (Ioannidis and Doucouliagos 2017, Mahoney 2010, Camerer et al. 2016). Given how difficult it is to find successful interventions, it is particularly frustrating for policy makers and researchers alike when the results from an expanded version of the tested program do not produce the expected improvements.

Mahoney (2010) provides an excellent discussion of the factors that can inhibit replicability of favorable results from randomized trials. Focusing on interventions designed to reduce falls among elderly individuals, she notes that although some programs may appear to be like one that was successful, they may produce quite different results for three reasons. First, the *study participants* may differ in nonobvious ways from those enrolled in the initial program. Second, the specific *content* of the intervention may differ—what were the detailed features and components of the intervention, such as which staff did what to which patients at what intensity level and for how long.[4] Third, the *process* used to fully engage providers and patients and family members to help them adhere to the behavior prescribed by the intervention may differ. For example, does the intervention build on models of readiness to change, and how is the program presented or "sold" to individuals (e.g., facilitating independence rather than preventing

[4] One could argue that these details should be part of the program design. While true in general, the variation in content would still exist for two reasons. First, many interventions (e.g., Accountable Care Organizations, charter schools, etc.) do not dictate how a program provider's behavior should change, but rather set out financial incentives and general structural parameters, intentionally leaving the details to the provider. Second, even if one does specify in detail the intervention content, substantial variation in intensity and quality of the intervention still will likely occur.

falls)? How aggressive is the follow-through after the initial intervention is delivered? This assessment includes but isn't limited to fidelity to the planned intervention. Mahoney also points out that one must assess how the care that the study group received differs from what the control group got. Experienced researchers know these things are important; what Mahoney stresses is that we need to know the *details* in order to learn anything useful.

Factorial designs could help. One approach to accounting for the variation in how interventions are delivered is to build key variations into the design of the study. Collins, Murphy and Strecher (2007), Grannemann and Brown (2018) and many others (e.g., Dey 2014, Ledolter 2007, Collins 2018) have promoted the use of full or partial factorial designs to specify which study units will implement which combination of key intervention components. Using efficient orthogonal designs, one can test many different components of an intervention and how it is delivered without a huge number of observations.[5] These types of studies are done every day by marketing departments. However, federal and state agencies have been reluctant to incorporate such designs, because they seem too complicated to implement. In the hands of skilled designers and implementers (to avoid spillover), such concerns can be alleviated. Unplanned variation in intensity and consistency will always complicate the analysis and require measurement, and there are many possible program variants that could matter. Nonetheless, creative factorial designs can enable researchers to obtain more rigorous evidence about the relative importance of different program components.

[5] Full factorial designs do require much larger samples if one needs to measure all interactive effects. However, if one only needs to identify multiple first-order effects, the sample sizes required do not increase with the number of treatment cells. Second-order effects can be detected in designs with many components with relatively little increase in the sample size, if efficient design algorithms are used. See Dey (2014) for details; Zurovac and Brown (2013) provides an empirical example of using such a design to test 8 components of a care coordination program with the patients of only 28 care coordinators.

Evaluators need to take a broader perspective than thumbs-up, thumbs-down assessments. Assessing the reasons for non-replicability and moving beyond the simplistic hypothesis-testing approach requires detailed data on the intervention—who is it delivered to, what is delivered, and how it is delivered to participants and/or service providers. Furthermore, we need to measure how those factors vary across participants, program sites, and time periods. And to complicate things further, the effects of these factors often interact with each other. We then need to link these measures to impacts, to the extent possible. This could help us avoid costly Type II errors from an overall impact analysis by identifying circumstances (e.g., sites with better implementation or a different participant mix or greater engagement) under which a program with an under-powered evaluation and statistically insignificant estimates may have been effective. The opportunity cost (in lives, not just in dollars) to dismissing an effective program because the test did not have enough observations to produce a statistically significant estimate could be far greater than the dollars spent on an intervention that is no better than the status quo. The challenge, of course, is not to err too far in the opposite direction—concluding that a program is likely effective in certain situations when the favorable findings are simply chance differences.

Another way of expressing the need to broaden our perspective is to make sure we focus on the "so what" question. That is, we have our estimates, but what do they imply for policy? If the estimated overall impact for an intervention is near zero, as Rossi's Iron Law tells us is likely, is that because:

- It was just a poorly designed intervention that was too weak to have the desired results? Or was it designed well, but poorly or inconsistently implemented, so that we really didn't have a good test of the planned intervention? Of course, if it was poorly implemented, we need to know why, so that the barriers encountered to implementing the design as it was intended can be addressed.

- The program had real effects, but only for a small subgroup of participants or sites? But if only some sites are effective, what was special about how they implemented the program or who they served that led to them having better results? And when power to detect differences in impacts across subgroups is limited, how can we assess whether the estimated differences in effects are "real" rather than just noise?
- The testing period was too short for the program to mature? In nearly every program evaluation I've been involved with, at the end of the study period the program operators say something like "we learned *so much* during the first couple of years. We are just now figuring out how to make this program work for our participants." That is usually a hard hypothesis to test, unless some sites started earlier than others, or some intermediate outcomes suggest that the program eventually might produce better core outcomes.[6] However, we also need to report what program operators and service providers think they learned about improving outcomes. It also argues for more pilot testing before performing a rigorous test (see the results for the University of Illinois at Chicago program in Gilman et al. 2020 for an example of how approach this can succeed). However, even when that approach is tried, it is unlikely to be successful unless the piloting is done by the providers included in the test. See Epstein and Klerman (2012) for a more thorough discussion of this issue.

[6] Depending on the nature of the intervention, the maturation of the intervention required for success may be either site-specific or generic. My experience is that program maturation is usually site-specific—lessons learned by other organizations during a pilot often do not apply or are insufficient for other organizations. This suggests that even well-designed and pilot-tested interventions will require time to become effective in new organizations and settings.

- The comparison group received services that had effects on outcomes approximating the effects of the services received by program participants?[7] I discuss this issue in a section below on the rapidly changing policy environment.

Even if the program estimates suggest the program produced favorable impacts overall, we still have important questions to address that go beyond that basic thumbs up/thumbs down conclusion. How likely is it that the effects will be observed if we expand the program to other areas or participants? What are the conditions under which we are likely to replicate the effects? Why were estimated effects larger for some sites or subgroups than others? What changes might make it work better for other sites and subgroups?

TWO EXAMPLES OF ENHANCED LEARNING

Two studies that my colleagues and I have conducted over the past few years illustrate ways that we were able to address the points raised above and learn some important lessons about possible determinants of intervention success. In both studies we were evaluating a small number of independent programs and looking across the set of impact estimates for possible broader inferences. A similar strategy could be used if these were different sites for the same intervention. In neither study did we use meta regressions, in which estimated program impacts are regressed on program characteristics in search of associations. We had too few programs and they were too diverse to run meaningful meta regressions, especially because many factors (and

[7] Greg Peterson noted an excellent example of this in his comments on an earlier draft of this paper. A care coordination program that had been successful in reducing hospitalizations over the first 6 years of a randomized controlled trial was no longer effective when expanded to a wider area. The hospitalization rate for the treatment group was just as low as in the earlier cohort. However, the control group's hospitalization rate was substantially lower in the later cohort than the earlier one, and similar to that of the treatment group. After careful examination of many possible reasons for the loss in effectiveness, the conclusion was that "usual" care received by the control group had improved for the target population. See Peterson et al. (2016).

their interactions) could have affected program impacts. Rather than rely on the "black-box" of a regression, which assumes a linear relationship and would have few degrees of freedom, we examine the data for factors and combinations that distinguish successful and unsuccessful programs. Deaton (2020) gives a compelling argument for careful examination of cross-tabulations and graphs such as this to better understand program effects.

The Medicare Coordinated Care Demonstration

The Medicare Coordinated Care Demonstration included nine unique programs, each with its own intervention and target population of Medicare beneficiaries with chronic conditions. The demonstration, which ran from 2002 to 2008, sought to reduce patients' need for hospitalizations and thereby reduce Medicare expenditures (Peikes et al. 2009). Each program was a randomized controlled trial. The estimates suggested that the intervention was successful in only four of the nine programs, and even there, only for a high- risk subset of patients that we identified (Brown et al 2012).[8]

To determine if various program features were important for reducing hospitalizations, we simply compared the features of the four effective programs to those of the 5 ineffective programs. The program characteristics were drawn from 10 different domains that we and

[8] The four programs had impact estimates on hospitalizations that were markedly higher than the other five programs and were statistically significant. Estimates for the four programs did not differ from each other but did differ significantly from the other five programs as a group. This study was a follow-on to an earlier study (Peikes et al. 2009) that presented results for the full study populations of programs funded under the demonstration, which found only one program to have statistically significant overall results. We estimated program effects for high-risk subgroups because most care coordination studies find impacts to be concentrated in a high-risk subgroup. Using the same definitions of high-risk for all nine programs, we estimated impacts using several alternative definitions for high-risk. Each high-risk indicator was constructed solely from claims data on diagnoses and prior utilization known to be associated with future hospitalizations. The different definitions produced very similar results. We selected the one that captured the largest proportion of the Medicare population (about 18 percent) and Medicare costs nationally (37 percent). See Brown, et al. (2012) for details.

program operators thought could contribute to programs' success. The data were gathered from annual site visits and follow-up telephone calls.

We found 6 characteristics that distinguished the successful and unsuccessful programs. The successful programs were much more likely to have care coordinators that served as a *communication hub* for the many providers their patients saw, had *monthly in-person* contact with patient to establish trust, and had regular *personal contact with patients' primary care physicians*. The successful programs also used *proven behavioral change approaches* to patient education and had *aggressive medication management* with reliable sources of information on patients' medication. Finally, the successful programs included a *proven transitional care component* to minimize the likelihood that a patient who was discharged from a hospital did not have to return there in the next 90 days. We defined each of these distinguishing features in more detail than can be described here, to increase the likelihood that others could replicate their success.

The study illustrates how using good qualitative and quantitative data to learn as much as possible, evaluators can go beyond assessing whether a program worked—we can also help in the ongoing process of program improvement. We cannot provide specific levels of confidence about the relative effects of each of the 6 characteristics we identified. But program operators can consider using these findings to strengthen their own efforts to improve outcomes for patients and lower their costs. We need evaluators and operators to work together with patients to keep improving and keep learning.

The Health Care Innovation Awards, Round 2 (HCIA2)

The HCIA2 program was a study of innovative programs, with the goal of finding some that could be scaled up to improve the quality of care and reduce costs to the Medicare and Medicaid programs. Although the program funded 39 separate programs, only 14 focused on patients with chronic conditions and could be evaluated with confidence, given problems with severe selection bias, small sample sizes, and/or poor data on outcomes. Four of these unique programs had favorable estimated effects on hospitalizations, emergency room use, and/or costs (Gilman et al. 2020, Brown et al. 2020).

Examination of 23 features measured in our implementation analysis revealed that 7 features were more common among the successful programs, and programs with these features had larger median impacts on one or more of the three core outcomes. For 3 specific intervention components—behavioral health, enhanced information technology, and telemedicine—programs having that component were associated with larger median impacts. Larger impacts were also associated with 4 features of the program or its host organization—prior experience providing the intervention, having nonclinical staff as frontline providers of intervention services, targeting socially fragile individuals (those lacking adequate income/transportation/housing/social supports), and focusing on improving *patient* support and behavior rather than changing *provider* behavior. However, median impacts did not vary meaningfully with any of our 10 measures of how effectively the intervention was implemented. We also found that combinations of these factors mattered. Programs targeting socially fragile patients were not successful in improving core outcomes unless they either used nonclinical frontline staff or had a behavioral health component to address patients' nonmedical needs.

The findings from the HCIA2 study are a good illustration of Rossi's point that an intervention's failure to address the right question or provide the appropriate intervention is a common reason for programs' failure to produce measurable impacts. In the HCIA2 programs, patients were enrolled in the study because they had a chronic illness, such as congestive heart failure or diabetes, but the real problem for many patients was behavioral health issues or poverty that made it difficult to adhere to prescribed medication and self-care. Those programs that addressed these concerns with the right type of staff and intervention had much larger estimated reductions in hospitalizations and expenditures (see Gilman et al. 2020 for more details). While this correlation could be due to other factors, the strength and consistency of the association suggests that it is a potentially causal relationship that warrants further consideration. The finding is also supported by many studies showing the importance of integrating behavioral health care with medical treatment of patients with chronic conditions (e.g., Klein and Hostetter 2014).

EFFECTS OF THE RAPIDLY CHANGING POLICY ENVIRONMENT

Unlike the early days of social program experiments, today many competing initiatives at federal, state, or local levels are likely to be implemented at the same time. The proliferation of efforts to improve the health, education, and welfare of our most vulnerable citizens is laudable. However, it becomes harder for evaluators to determine whether and what is effective. That is, it is often far more difficult now to define the counterfactual for a specific evaluation. The program may or may not have had effects but compared to what alternative? In addition to concurrent alternative programs, there are often unanticipated shocks that complicate program evaluation. These shocks could be global events that affect all areas of the country, or local changes.

Alternative interventions or programs in comparison areas. There are many examples of this conundrum, especially in the health research area. To illustrate, consider the Comprehensive

Primary Care+ model (CPC+), a major effort to transform how primary care is delivered. The model seeks to improve the quality of care and reduce Medicare expenditures through reductions in patients' need for expensive services. CPC+ includes nearly 3000 participating physician practices, serving nearly 3 million Medicare patients. The model provides funding to cover the costs of transformation, gives feedback to practices on their costs and quality measures, requires practices to meet milestones for improved care quality, and provides financial incentives for reducing costs and/or improving the quality of care provided. After two years, however, the evaluation has found little or no impacts on costs or hospitalizations, despite the high power to detect small impacts (Anglin et al. 2020).

One potential reason for the lack of impacts to date is the large number of other interventions occurring in the areas from which comparison practices were drawn. The evaluation team identified at least 13 concurrent national and regional programs with the same goal of reducing health care expenditures.

The study team will use "triple difference" comparisons (Wing et al. 2018) and regression control variables to net out or account for these other program and regional effects. These alternative programs could be affecting the trend in outcomes in treatment and comparison areas in different ways, and possibly offsetting program impacts (Peikes et al. 2018). The triple difference approach nets out any area-wide difference in outcomes between treatment and comparison areas that are occurring. This is accomplished by including in the study all primary care practices in the two areas that are in neither the treatment nor comparison groups. The regression will include binary control variables for each of the alternative programs operating in the area, indicating whether the practice was participating in that program. The regression modeling will be exploratory and would have been infeasible had the study not had a large sample of participating providers (practices) with substantial variation in these confounding factors.

The effects of external shocks. In addition to the proliferation of other concurrent interventions that can affect outcomes, evaluators must account for the effects that unanticipated areawide or global events can have on outcomes and impact estimates. National or world events such as the Covid19 pandemic, economic recessions, wars, or major technological advances affect outcomes in both treatment and comparison areas, but not necessarily to the same degree or even in the same direction. Local changes, such as the opening of a major new manufacturing plant or hospital, or the closing of important existing facilities can also influence outcomes and can have similar or even larger effects on outcomes in comparison or treatment group areas. These shocks can affect both the ability of programs to implement the planned intervention and the evaluation.

The randomized control trial of the Transitional Care Model (TCM) that my colleagues and I are evaluating provides a good illustration of the effects of such shocks. TCM uses advance practice nurses to work closely with patients once they are discharged from a hospital to reduce the likelihood that they will need to be readmitted within 30 days after discharge. Several previous small randomized trials have found the model to be very effective, reducing readmissions and costs by about one-third (Naylor et al. 1999, Naylor et al. 2004). The current study is assessing whether TCM can be equally effective in 4 different hospital systems (each implementing in two or three hospitals), and with an expanded set of diagnoses included in the eligibility criteria. Funded in Fall of 2019, it was scheduled to begin enrolling patients in May 2020.

The onset of Covid19 affected the study in several critical ways and will continue to affect program operations and the evaluation for the next 4 years. These effects include changes in the number and characteristics of enrollees, the content of the intervention and the process of delivering it, the statistical analysis that will be conducted, and the interpretation of the findings. Note that each of these are factors identified by Mahoney as reasons for the inability to replicate the success of programs previously found to have favorable effects:

- Effects on enrollment and study participants
    - The start of enrollment was delayed
    - Participating hospitals are identifying fewer eligible patients
    - Eligible patients are less willing to participate than those in previous TCM studies (concerned about in-person visits)
    - Study participants are more severely ill on average than patients hospitalized with the target conditions in the past (fears about contracting the virus in the hospital, beds unavailable)
    - Consequently, intake will take longer than originally planned; eligibility criteria are being expanded to help reach target enrollment sizes
- Effects on the content and process of the intervention itself
    - Some patients are requiring that nurse coordinators contact them by telephone rather than in person.
    - Due to added burdens of Covid19, hospital administrators and program staff may have less time for implementing and monitoring the intervention, which could adversely affect its impact.
- Effects on the statistical analysis
    - Differences across hospitals and systems in the effects of Covid19 on the intervention heighten the importance of the system-specific impact estimates, for which statistical power is lower.

    - As Covid19 waxes and wanes, program impacts may change over time, and in different ways for different sites. Accounting for this may require estimating impacts for subgroups of enrollees defined by when they enrolled, and the timing may differ across the four systems.
- Possible effects on impact estimates and interpretation of the findings
    - Changes in the patients, the availability of hospital beds, and the intervention protocols could result in impacts that are either higher or lower than in "normal" times.
    - If statistical power is reduced due to failure to attain target sample sizes, it will be difficult to assess whether sizeable differences in estimated impacts across systems are due to real differences in effectiveness or simply to chance.
    - If program impact estimates do differ significantly across systems, it will be even more difficult than usual to assess the reasons.

While the Covid19 shock creates numerous challenges to the study, it also creates opportunities to learn about how adaptable the intervention is to real-world complexities. TCM may be more important, and more impactful, in the presence of a pandemic than it has been shown to be in simpler times. The differences across systems may show that the ability to adapt the intervention is more important than fidelity to the strict protocol for the program to remain effective.

Rapidly changing policy environments are likely to continue. Such changes, exacerbated by policymakers' desire for short-term success, make it essential that evaluators understand the specific environmental factors and counterfactuals. At the same time, programs and providers may be more open to change and to keeping up with the latest innovations. Furthermore, the variation induced by the changing environments can, if analyzed properly, greatly expand and accelerate learning.

CONCLUSION

The main upshot of this paper is that, despite Rossi’s Iron Law that few interventions will produce measurable effects, they can all be successful by increasing our understanding. We need to learn how to better address the needs of our most vulnerable citizens. Even when successful interventions are found, we often fail to replicate that success. It’s highly likely that there is no single “silver bullet” (another metallic reference) that will improve outcomes for everyone and every setting. This highlights the need to balance the search for consistent drivers of success across programs with learning what worked from in-depth case studies of individual programs.

At the end of the day, what matters is producing a good predictor of what would happen if some version of the program were implemented more broadly or in different settings. This work requires a thorough, detailed understanding of what was implemented and in what context, along with reliable impact estimates. Doing this well will require a multi-disciplinary team of researchers.

Journals can assist in this effort to advance learning by basing acceptance of articles more on whether something important was learned than on the statistical significance of the impact estimates. Such a change would also lessen publication bias, which can lead policymakers and others to assume Rossi’s Iron Law no longer holds.

We will continue to make important advances in statistical methods and evaluation designs to improve our evaluations. I’m enthusiastic about the growing use of Bayesian and some big-data methods to produce more reliable and useful estimates. Agencies and other funders should consider incorporating factorial designs into their programs and models being tested, whether using RCT or not; they may find that they learn much faster how to improve their programs. These and other innovations will provide better and more insightful estimates of program impacts.

But to address Rossi's points about the reasons most interventions fail—wrong question, wrong intervention, poor implementation—we need to devote more attention and resources to addressing the "why", "how", and "where" questions. This will require better implementation data and more creative analysis of it than we typically see. We know that the process, content, intensity and effectiveness of implementation are critical to a program's success, and they are hard to measure well. That will require strong inter-disciplinary teams of researchers, and ongoing dialog with operators, funders and intervention recipients.

Making the changes suggested to program design and analyses, both quantitative and qualitative, suggested in this paper will require a concerted joint effort by funders, program operators, and evaluators. Funders can learn about factorial designs from skilled practitioners and work with them both to design studies that incorporate these designs and explain to program operators how they can be successfully implemented. Funders can also insist that evaluators provide Bayesian estimates of program effects in addition to traditional estimates, to see how evaluation inferences can be improved. Evaluators can provide hands-on workshops and tutorials to funders about these methodological developments. Program operators can do their part by being amenable to trying new approaches, and by helping researchers narrow the nearly infinite list of implementation details to measure, identifying those likely to be the most important drivers of change. Most importantly, funders, operators and evaluators should work together to ensure that evaluations gather the detailed implementation data needed to fully understand variations in how the program was implemented over time and across sites and use it effectively. Hopefully, this paper has illustrated how enhanced implementation data and analysis can help us expand and expedite learning about how social programs can improve lives.

REFERENCES

Anglin, G., Peikes, D., Peterson, D., et al. (2020). Evaluation of the Comprehensive Primary Care Plus Initiative: Second annual report. Princeton, NJ: Mathematica.

Brook, R., Ware, J., Rogers, W., Keeler, E., Davies, A., Sherbourne, C., Goldberg, G., Lohr, K., Camp, P., &. Newhouse, J. (1984). *The Effect of Coinsurance on the Health of Adults: Results from the RAND Health Insurance Experiment*. Santa Monica, Calif.: RAND Corporation, R-3055-HHS.

Brown, R., Derzon, J., Whicher, D., Gilman, B., & Dale, S. (2020). Features of innovative health care interventions associated with reduced service use and expenditures, manuscript under review.

Brown, R., Peikes, D., Peterson, G., Schore, J. & Razafindrakoto, C. (2012). Six features of Medicare Coordinated Care Demonstration programs that cut hospital admissions of high-risk patients. *Health Affairs, 31(6)*, 1156-1166.

Camerer, C.; Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., & Altmejd, A. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351(6280)*, 1433–1436. doi:*10.1126/science.aaf0918*

Collins, L., Murphy, S., & Strecher, V. (2007). The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New methods for more potent eHealth interventions. *American Journal of Preventive Medicine, 32(5),* S112–S118.

Collins, L. (2018). *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*, Springer.

Deaton, A. (2020). Randomization in the tropics revisited: A theme and eleven variations, NBER Working Paper 27600, Cambridge, MA.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Dey, K. (2014). *Competitive Improvement and Innovation: Statistical Design and Control*, CRC Press.

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, *36*(5), 375-401.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 185-211.

Gilman, B., Whicher, D., Brown, R., McCall, N., et al. (2020). Evaluation of the Round 2 Health Care Innovation Awards: Final report. Mathematica. Available at: https://innovation.cms.gov/data-and-reports/2020/hcia2-round-2-final-eval-report-sept-2020-0.

Grannemann, T., & Brown, R. (2018). Adapting evaluations of alternative payment models to a changing environment: Research designs, methods, and learning strategies. *Health Services Research, 53(2)*, 991-1007. Initially released online March 2017 at *https://doi.org/10.1111/1475-6773.12689*

Hahn, P.R., Murray, J., & Carvalho, C. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15(3)*, 965-1056.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47(1)*, 153–61. *doi:10.2307/1912352. JSTOR 1912352. MR 051883, 1979*

Ioannidis, J., Stanley, T., & Doucouliagos, H. (2017). The power of bias in economics research". *The Economic Journal*, *127(605)*, F236–F265. *doi:10.1111/ecoj.12461. ISSN 1468-0297. S2CID 158829482*.

Kershaw, D. & Fair, J. (1976). *The New Jersey Income-Maintenance Experiment: Volume 1: Operations, Surveys, and Administration*, Academic Press.

Klein, S. & Hostetter, M. (2014). In focus: Integrating behavioral health and primary care. Issue Brief. The Commonwealth Fund.

Ledolter, J. & Swersey, A. (2007). *Testing 1 - 2 - 3: Experimental Design with Applications in Marketing and Service Operations.* Stanford Business Books 1st Edition.

Luce, B., Kramer, J., Goodman, S., Connor, J., Tunis, S., Whicher, D., & Schwartz, J. (2009). Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Annals of Internal Medicine, 151(3)*, 206-9. *Doi: 10.7326/0003-4819-151-3-200908040-00126*

Mahoney, J. (2010). Why multifactorial fall-prevention programs may not work. *Archives of Internal Medicine, 170(13)*, 1117-1119. *Doi: 10.1001/archinternmed.2010.193*

Metcalf, C., & Thornton, C. (1992). Random assignment and child welfare reform evaluations. *Children and Youth Services Review, 14(1)*, 145-156.

Naylor, M., Brooten, D., Campbell, R., Jacobsen, B. S., Mezey, M. D., Pauly, M. V., & Schwartz, J. S. (1999). Comprehensive discharge planning and home follow-up of hospitalized elders: A randomized clinical trial. *JAMA, 281(7)*, 613-620. *Doi: 10.1001/jama.281.7.613*

Naylor, M., Brooten, D., Campbell, R., Maislin, G., McCauley, K., & Schwartz, J. (2004). Transitional care of older adults hospitalized with heart failure: A randomized, controlled trial. *Journal of the American Geriatrics Society, 52(5)*, 675-684. *Doi: 10.1111/j.1532-5415.2004.52202.x*

Peikes, D., Anglin, G., Dale, S., Taylor, E., O’Malley, A., Brown, R., Duda, N., et al. (2018). Independent evaluation of Comprehensive Primary Care Plus (CPC+): Design report. Princeton, NJ: Mathematica Policy Research.

Peikes, D., Chen, A., Schore, J., & Brown, R. (2009). Effects of care coordination on hospitalization, quality of care, and health care expenditures among Medicare beneficiaries: 15 randomized trials. *Journal of the American Medical Association, 301(6),* 601-616.

Peterson, G., Zurovac, J., Brown, R., Coburn, K., Markovich, P., Marcantonio, S., Clark, W., Mutti, A., & Stepanczuk, C. (2016). Testing the replicability of a successful care management program: Results from a randomized trial and likely explanations for why impacts did not replicate. *Health Services Research, 51(6),* 2115-2139.

Rossi, P. (1987). The Iron Law of evaluation and other metallic rules, *Research in Social Problems and Public Policy* (ISBN: 0-89232-560-7), *4*, 3–20.

Vollmer, L., Finucane, M., & Brown, R. (2018). Revolutionizing estimation and inference for program evaluation using Bayesian methods. *Evaluation Review, doi: 10.1177/0193841X18815817*. Epub ahead of print. PMID: 30537865

Wing, C., Simon, K. & Bello-Gomez, R. (2018). Designing difference in difference studies: best practices for public health policy research. *Annual Review of Public Health, 39,* 453-469.

Zurovac, J., Brown, R., Schmitz, B., & Chapman, R. (2013). The effectiveness of alternative ways of implementing care management components in Medicare D-SNPs: The Brand New Day Study. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.