2013

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure

Robert Boruch
University of Pennsylvania, robertb@gse.upenn.edu

Alan Ruby
University of Pennsylvania, alanruby@gse.upenn.edu

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure

Abstract
Postmortems and autopsies, at the individual and hospital unit levels, are disciplined approaches to learning from medical failures. "Safety factors" that engineers use in designing structures and systems are based on past failures or trials and experiments to find points of failure.

The applied social sciences, including education sciences, labor economics, and criminology, have less clarity about failure. While a bridge collapse is usually plain and spectacular, failures of education innovations or attempts at crime control are often quieter, not spectacular, and often occur for no transparent reasons.

The applied social sciences lack disciplined, well-developed, and explicit approaches to anticipating the failure to meet expectations in testing the effectiveness of programs, analyzing the failures, and building a cumulative knowledge base on the phenomenon. Our fields can, for instance, identify "what works" pretty well from randomized controlled trials. *However, little serious attention has been dedicated to understanding "why" and "how" a particular intervention failed to meet expectations in well-executed randomized controlled trials*. This essay discusses a variety of research initiatives that are designed to better understand failure, especially in controlled trials.

Keywords
failure, postmortems, social sciences, labor economics, criminology, research

Disciplines
Criminology | Educational Assessment, Evaluation, and Research | Educational Sociology | Labor Economics | Social and Philosophical Foundations of Education

Comments
This version of the paper is titled "To Flop is Human: Can We Invent Better Scientific Approaches to Anticipating and Learning from Failure to Meet Expectations?"; it was published in its final form as "To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure"

To Flop is Human:

Can We Invent Better Scientific Approaches to Anticipating and Learning from Failure to Meet Expectations?

Robert Boruch and Alan Ruby University of Pennsylvania

Revised December 12 2013/ To be Published in *Emerging Trends*

Preamble

Post mortems and autopsies, at the individual and hospital unit levels are disciplined approaches to learning from medical failures. "Safety factors" that engineers use in designing structures and systems are based on past failures or on trials and experiments to find points of failure. Books on the topic in each arena are in ample supply. Some are good.

The applied social sciences, including education sciences, labor economics and criminology have less clarity about failure. While a bridge collapse is usually plain and spectacular, failures of education innovations or attempts at crime control are often quieter, not spectacular, and often occur for no transparent reasons.

More to the point, the applied social sciences lack disciplined, well developed, and explicit approaches to anticipating the failure to meet expectations in testing the effectiveness of programs, in analyzing the failures, and in building a cumulative knowledge base on the phenomenon. We can, for instance, identify "what works" pretty well from randomized controlled trials. *But little serious attention has been dedicated to understanding "why" and "how" a particular intervention failed to meet expectations in well executed randomized controlled trials.* Here, we consider a variety of research initiatives that are designed to better understand failure, especially in controlled trials (Footnote 1).

Failure Aversion

One can always define failure out of existence, or define it in a way that reduces its ostensible frequency. For example, you can change the definitions of a performance indicator, like number of escapes from prison per thousand inmates by determining that an escape is not an escape unless the body is missing for 24 hours. If, after 24 hours, the convict returns to the prison voluntarily, this can be denominated as an "unexcused absence." If our convict is released temporarily to attend college courses outside the prison, but does not come back, this is labeled at least temporarily as "failure to return," as opposed to "escape."

It is not difficult to find similar examples in education. For instance, it has taken over twenty years for the US to agree, more or less, on more or less transparent definitions of "school drop-out." How much time does the child have to be gone to declare the child a drop out? In the medical sector, "cause of death" can also be nuanced and oriented toward the benign or the grim, depending on who is counting and why.

Circumlocutions are just that, attempts to avoid labeling an event as a failure or admitting error. There is an aversion to acknowledging shortcomings in performance. Some scholars go so far as to do serious research on the traumas caused by the "strong accumulation of emotions stemming from… failure" that discourage people from learning from or admitting failure (Valikangas et al. 2009). These traumas and the wish to avoid them lead us away from opportunities to learn from our own errors. Indeed, some empirical studies suggest that it is easier to learn in other ways. Baum and Dahlin (2007) illustrate this in studies of train wrecks where companies learn from the errors of other operators more readily than they learn from their own.

Part of the lack of interest in failure comes from a perception that "trial and error" and incremental improvement underpin good public policy making. This has been a prevailing paradigm since Lindblom (1959) advocated "muddling through" as an alternative basic rationality. Part comes from the reluctance to make generalizations because failure is a social construct and on account of the notion that every failure is unique. For example, Bovens and t' Hart (1996) argue that the act of defining a policy as a failure or a success is inherently subjective and shaped by space and time. Others, such as Peters (1997), criticize those who take refuge in variations in time and place so to avoid drawing lessons that might avoid error and eliminate "more general patterns of policy that have the tendency to produce pathological results" (page 259).

Despite common aversion to thinking seriously about failure in many social sectors, Besharov (2009) boldly and properly declared that "R and D strategies should be planned with failure in mind" (p.210). The idea is not new. Levitt and March (1988), among others , fostered interesting work based on the theory that organizations can learn from their failures as well as successes and that some fail and survive nonetheless.

Our basic proposition is that failure has not been given the respect it deserves and the study of failure is compatible with the basic tenets of social science. There is a need to define a field of "failure analysis" in the applied social sciences.

Why Study Failure?

John Dewey (1933) claimed that "failure is not mere failure. It is instructive." (pp 114-115). Studying the causes and effects of failure may reveal ways to avoid repeating errors and prevent failure. At best, such studies can increase the chances of success by identifying how to do things better. This, of course, is a naive view of how social policies and programs are

designed and improved, because it overlooks the frailties and shortcomings of public policy making. Nonetheless, it is grounded in the tenets of rationality and evidence-based decision making. More information about what does not work and why it does not work is likely to lead to more effective policies and better designed programs. And as Sabatier (2005, p20), one of the originators of implementation analysis observes, the best designed institutions (the British Open University for example) are the most likely to be successful. This will save money or at least see that it is expended more efficiently, and improve the services delivered to those in need.

A compelling reason for attending seriously to the study of failure is that there are plenty of instances to examine. We know lots about failure "events," such as children dropping out of high school and college, passing grades that are not reached, and methods of instruction that do not always deliver the intended curriculum. But we do not know a lot about the system failures, their rates, and why these events occur.

There is plenty of knowledge in other fields and disciplines about the study of failure and which might inform the social sciences' examination of the topic. For instance, upwards of a dozen peer reviewed journals in medicine, pharmaceuticals, and engineering focus solely on the downside events. They carry such titles as *Journal of Null Results*: www.journalnullresults.com, *Journal in Support of the Null Hypothesis*: www.jasnh.com, Journal of Negative Results in Biomedicine: www.jnrbm.com, and the *Journal of Pharmaceutical Negative Results*: www.pnrjournal.com. While these journals are relatively recent, and their own life span may be short, the idea of learning from failure has been around for a long time. In the construction field, for instance, learning from failure in large-scale publicly funded endeavors dates, at least, to the construction of the pyramids. Dahshur's "Bent pyramid" suggests that its builders, working for Pharaoh Snefru, learned from the snafu (not Snefru) of their earlier Meidum (aka Maidum)

pyramid. They adjusted the design while construction was in progress (www.bbc.co.uk/.../**pyramid**_gallery_05.shtml). While it may lack the clarity of line that the Pharaoh desired, the Dahshur monument still stands.

The scale of infamy, the costs and visibility of failure and a long string of repeated instances of design and structural shortcomings have made identifying potential causes of failure and learning how to do better an important part of contemporary engineering. Conventional texts usually handle cases like the British Comet jet aircraft, the Tacoma Narrows Bridge, smoke stacks, hotel walkways, and others. See Petroski (2012) for net illustrations and Florman's (1996) *Existential Pleasures of Engineering*.

It is not that there haven't been any studies of failure in the social sciences. In public policy Pressman and Wildavsky (1984) explored the challenges of successfully designing and delivering economic development programs. Bovens and 't Hart (1996) reviewed literature on public policy fiascoes in the United Kingdom. Gornitzka, Kogan, and Amaral (2005) developed case studies from a number of countries on successes and failures of public policy reforms in higher education. More recently, Berman and Fox (2010) studied failures in a variety of criminal justice interventions including "drug courts" in Denver and Minneapolis, Boston's Operation Ceasefire which people tried to replicate in Minneapolis Chicago and Philadelphia, prison reform in California, and "Three Strikes" legislation in Connecticut. Berman and Fox give no instruction about how to go about examining failure systematically, though they and most other scholars seem to be confident about how to identify it.

In the adolescent health sector, sociologist Carol Weiss (2002) posited a theory of change for who a community initiative might or might not work, to enhance adolescent well- being. Her

diagrammatic representation of how the program *could* work had about 40 casual arrows connecting assorted actions and events to putative intermediate and long term consequences. For us, her more important insight lay in constructing a parallel diagram, also with lots of arrows, to portray how the program *could* fail, and have negative effects on adolescents. She did not pursue the idea deeply, but it merits attention.

In the social sciences, no one has developed a systematic approach to the topic. There is no protocol like that which guides the interpretation of autopsy results (Rutty, 2001), to guide assessments of why, say, a school has "failed." The use of measures like Adequate Yearly Progress (AYP) or graduation rates is not typically tied to a systematic analysis of causes of failure. The measures do not sum to a code of practice or set of processes to guide a coherent and structured inquiry into an institution. Yet there is a lot at stake.

The failure of successive cohorts of students costs more, we aver, than a bridge collapse. In aggregate the costs of college dropouts and failures to learn to read, in terms of lost productivity and opportunity, are arguably greater than the costs of an engineering error. The scale of this lost value justifies our search for lessons from failure in other disciplines that might help improve the quality of education.

In sum, we are arguing that studying failure in education sciences and other social sciences is needed. If the argument holds, what might such study look like?

Building a Field of Failure Analysis in the Social Sciences

Rather than declare answers, we take an interrogatory approach here. Our aim is to offer some questions to guide the development of a systematic approach to the study of failure and its anticipation.

In developing these questions and illustrating them, we draw on the lessons from work in an important but nicely bounded scientific and policy arena: randomized controlled trials. This is because these trials, when done right, permit fair comparisons and legitimate scientific statements of confidence in one's results for a relatively unequivocal causal inference. The declaration that "A worked better than B," under a particular statistical test of the null hypothesis (or a set a confidence interval for the mean difference) is legitimate, satisfying, and gets attention.

Frequently, the trial's results do not permit a conclusion that "A" works any better than "B" because the mean difference in outcomes of two interventions is not statistically significant when chance (noise) is taken into account. For good scientists, the failure to detect a statistically significant difference in outcomes for the interventions that are tested in a well-designed and executed trial is itself a scientific success. This common scenario—discerning no remarkable difference between interventions — presents an opportunity to think about how to get beyond the conventional statistical declarations in comparisons of A to B. Merely declaring that the bridge fell down is not enough. It is an opportunity to advance scientific practice in understanding "null findings" on the effects of interventions and to exploit and advance theory and practice in education, social services, policing, corrections, and other sectors.

We offer five questions to begin an investigation of how to study failure in the applied social sciences and in the context of randomized trials:

Q1. How can we define failure to meet expectations in such controlled tests?

Q2. How can the interventions that are tested in randomized trials be designed so as to reduce the likelihood of failing to meet expectations?

Q3. How do we design randomized controlled trials *a priori* so as to better learn from the inevitable failures to meet expectations about the effectiveness of the interventions?

Q4. How can we learn about plausible reasons for failure to meet expectations *ex post facto* in a scientifically and orderly way?

Q5. How can we build cumulative knowledge base on when, how, and why the failure occurred?

We offer some tentative responses in what follows, and acknowledge that there are other answers and other questions that may be better.

Q1. How can we define failure to meet expectations in randomized controlled trials?

Putting aside debates about "who" should define failure we suggest that unless we properly define failure of a tested program, we cannot define its success, nor declare when either occurs. And we cannot dodge the matter by talking about "mixed effects," taking refuge in the curate's description of the rotten egg, that "parts of it are excellent" (du Maurier, 1895). It is "failure to meet expectations" about an intervention's value that is of primary interest in educational, criminological, and other randomized trials in the social sciences. The scientist hopes that "A" will be better than "B" in a fair trial, otherwise would not bother to make a fair test.

The failure to reach expectations might be defined solely in terms of the tested intervention's failure to get beyond a pre-specified level of chance, i.e. statistical significance. This is fine for people who want to avoid deluding themselves into thinking that an effect is dependable rather than a matter of chance. But the statistically computed "effect size" is at least as important as a probabilistic threshold, and is more important in some respects. This effect size

is the mean difference in outcome between two groups that have gotten different interventions, adjusted for the inherent variability of the groups being compared. We then define "failure" narrowly here *as the failure to meet expectations about an "effect size."* The rationale is that the expected effect size is the scientifically accepted basis for designing randomized controlled trials that have sufficient statistical power, i.e. are sensitive to the expected effect of the tested intervention.

The bottom line is this. If you buy the expectation of effect size in designing the study, you have bought an expectation about possible results.

Q2. How might the interventions that are evaluated in RCTs be designed so as to reduce the likelihood of failing to meet expectations?

Al Reiss's Law (Personal Communication, circa 1987) avers that a new policy, program, or intervention must be sufficiently different from the status quo (the control condition or the alternative intervention) to justify an investment in a field trial so as to generate a fair estimate of the new approach's effect. In the Spouse Assault Replication Program (SARP), for instance, Sherman and other colleagues (1992) had to ensure that perpetrators would be randomly arrested, or not, in their trial on preventing misdemeanor domestic violence. To some theoreticians interested in the effect of arrests on recidivism "A," an arrest, looked a lot different from "B,", such as being told to calm down by a cop.

Arrests were indeed carried out. But, to the surprise of many, arrests had no discernible effect on recidivism. This is possibly because A and B do not look different *to the perpetrator*. Arrest is often not a significant deprivation of liberty. Rather, it can be a transient event of little consequence to the perpetrator. This suggests a refinement of Reiss's law: A has to look

different from B to the intervention's target. Our second proposition is that the designers of social interventions borrow the idea of "safety factors" from the engineers. The latter tend to design using norms and principles grounded in theory and practice about such things as the load bearing strength of concrete and steel. To be conservative, however, they multiply the calculated resources by 2 or 3 or more, "a safety factor," so as to take into account inevitable uncertainties of the field. Such a multiplier can also be accurately labeled as an "ignorance factor." But calling the thing a safety factor is more comforting.

We, in the social and education sectors, lack a thematic emphasis on safety factors in designing interventions and their field tests. In principle, at least, the notion of planning for our ignorance based on earlier failures and safety factors seems worth exploring. For example, there are estimates of the "guided learning time" likely to be required to progress from one level to another on the Common European Framework of Reference (CEFR) for English language teachers. The estimates are used when designing programs. The CEFR is based on a significant corpus of research and practice and exemplifies how lessons from research can lead to better program design (Cambridge University Press, 2013). The rough rule of thumb is adults can reach basic proficiency in four weeks in oral, aural, and written communication, but many program providers plan on six weeks.

A third line of thinking in the design of interventions that are tested in controlled trials lies in exercising "due diligence" in the intervention's planning and execution. Thoughtful CEOs and attorneys normally actualize the idea of due diligence in the context of mergers and acquisitions. Ruby (2010) applied the idea in the area of a recent spate of international branch campuses. Recent and far reaching lapses in due diligence are nicely exemplified by the financial industry. These lapses are complicated but illustrated well by apocryphal Heidi's Bar

and the notion of derivatives in the financial sector. Heidi decided to boost business volume by permitting her customers to sign IOUs instead of paying cash. She borrowed money from banks on these notes to pay her suppliers. The banks bought the notes from Heidi in batches. expecting that payments and profits would ensue. The bar flies "forgot" their debts. Heidi's Bar folded. So too did the banks that bought the notes.

Medical writer Atul Gawande (2009) considered related matters in his *Checklist Manifesto.* His theme is that one ought to develop lists of things that are necessary to assure that, in effect, *some* of our expectations are met. He describes how such checklists are operationalized and used in construction work, airplane safety, hospital procedures, and other areas. For scientists in the social sector, thorough checklists are in short supply whether the program is an educational curriculum package or a crime prevention program. A good checklist depends on understanding what could go wrong.

Strang's (2012) special issue of *Journal of Experimental Criminology,* on managing field experiments invites checklists beyond contemporary ones. Exogenous factors such as the job market, the endogenous stability of the prison system or the police department, the time frame for the experiment, and so on demand more attention. Checklists are subject to empirical testing, as in prospective studies in medicine, e.g. Pronovost et al. (2006).

Despite the ambiguity attached to the phrase "systems theory," the idea is important to designing interventions and trials so as to anticipate failure. Consider, for example, that a cluster randomized trial, mounted in 2007 in four cities and 180 schools, depended on a technically well-designed study and a seemingly well-designed intervention to test the intervention's effect on science knowledge of middle school students. Abundant theory and all evidence at hand was used to deploy the work. Information about local parameters such as

number of schools, number of teachers within schools/classes, and so on was exploited for statistical power calculations in designing the trial. Contemporary cognitive science principles were used to revise science curriculum modules so as to enhance student achievement.

What was not taken into account fully in design of the intervention, or the trial, was systems related. In particular, Boruch, Merlino, and Porter (2011) found that ambient positional instability (API), the "churn," among teachers in the school system is potentially critical. About 42% of teachers in one city had taken a position in September 2011 that was different from what they had in September 2010. About 46% did so in a second city. The reasons for such instability are complicated, varied, and localized, e.g., sabbatical leaves, teacher's subject area reassignments (from science to math), grade reassignments, and assignments to administrative duties.

Intervention designs and the experiments designed to evaluate effects, in education, as in other sectors, usually assume system stability. They do not assume instability at various system levels. It is reasonable nonetheless to assume that interventions that depend on teacher continuity will not achieve an effect of an expected size unless Ambient Positional Instability is taken into account.

Q3. How do we design randomized controlled trials *a priori* so as to better learn from the inevitable failures to meet expectations about the effectiveness of the interventions?

Applied statisticians and social scientists who design field trials of any kind, do not usually ask this question. They focus on designing the trial to test a formal null hypothesis fairly and to produce a dependable estimate of the intervention's effect and its variability. They often do not concern themselves with instability inside or outside the black box (the context) or with

what impact it has on individuals. We are only beginning to get to the point of designing studies so as to anticipate the intervention's failure and learn from it.

In recent years, some trialists in education and criminology have done well in getting beyond "black box" trials by anticipating the possibility of failing to meet expectations. Measuring the extent to which police actually apply a new practice is integral to good practice in criminology studies (Boruch, Weisburd, Berk, 2010). In education, Garet and his colleagues (2008) have advanced understanding of how to measure implementation of a professional development program in education during the course of the trial by showing that teachers learned, but students did not score any better than before the teachers engaged in the program. More generally, Grubb emphasizes getting into the field to look at fidelity of implementation and alignment of behavior with reports for "without such understanding it's impossible to know the reasons for failure" (quoted in Besharov (2009, p 211). Ground level work is desirable, but so too is work at higher levels in the systems that contain and affect the intervention being tested, such as instability.

Designing interventions and testing them in uncertain or volatile contexts is hard. But some design lessons can be drawn from other fields. One example is the engineering tradition of the "run in period," dedicated to stabilizing the system to be tested and to working out the kinks the design and construction. It is a good idea to make the elevator go up and down fifty times before allowing it to carry passengers and before building another one the same way. In some educational and criminological studies, this is equivalent to a two cohort design. We make all the mistakes we can, and learn from them, in the first cohort. We may have to abstain from analysis of outcome data from this cohort because of missteps during the run in period. The

second cohort is dedicated to the real comparison of the interventions and estimating the actual effect size.

People engaged in field experiments can deepen contextual knowledge by direct observation, paying attention to who is doing what, with what incentives and resources, and who is not, with what disincentives and resources. Anthropologists call this developing "grounded theory." In criminology, some quantitative researchers emulate their qualitative colleagues by engaging in "ride-alongs" to uncover issues in policing experiments and to understand the streets. The best of education researchers who engage in randomized trials also engage in classroom observations and talks with principals, teachers, or parents, and so on. This is in the interest of better design and to troubleshoot the trial's execution.

Regardless of the trial's specific design, the local knowledge is essentially tradecraft. It is a marketable commodity, of commercial value and intellectual value. But it is laden with potential embarrassment because the overlooked piece of human behavior that causes the intervention to fail it is often so "obvious". Perhaps this why it is not often written up in peer reviewed research journals. The future of failure analysis lies with getting beyond tradecraft by developing transparent and orderly approaches to the design and evaluation of interventions and learning how to report the results.

Q4. How might we learn about plausible reasons for failure to meet expectations *ex post facto* in an orderly and scientific way?

This is really hard. Unlike engineers, the applied social scientist, in education, criminology, and so on cannot execute "trials to failure" so as to understand at what point the intervention (a structural support) fails. Unlike colleagues in the pharmaceutical industry, we cannot do trials in which low or high doses in different animals are tried out to determine what is

too weak to help and what kills rather than cures. In the social sciences, we cannot usually make unequivocal declarations about causes of an intervention's failure based on randomized trials because we cannot design ethical trials to test directly the causes of failure.

Nonetheless, when the randomized trial is over and we've uncovered no discernible difference between "A" and "B," it is reasonable to do an orderly post mortem by asking a few obvious questions:

(a) Was the trial designed and executed well? And how do we know?

(b) Were the two interventions, "A" and "B" delivered as expected? And how do we know?

(c) Was the theory underlying the design of the expectedly better intervention "A" wrong? And how might we speculate well or know better?

This list is similar to one invented earlier by St. Pierre et al (1995) in their reports on the first randomized tests of the Even Start Family Literacy Program. The discernible effects of that program were negligible.

In regard to item (a), good standards exist for assessing the quality of a trial's design and its execution. If the trial wasn't done right, or was sabotaged, we still know nothing about the benefits of A over B. So we may try again like the early trials on enriched oxygen environments for premature babies (Silverman, 1980).

Item (b), assessing the fidelity of delivery is part of due diligence in any randomized trial and in any *ex post facto* analysis. But the evidence generated as a consequence of addressing earlier questions would make the post mortem in this context easy.

Item (c), concerning the theory underlying how A is supposed to work better than B, or how B is supposed to be inferior in effectiveness than A, is more challenging. *Ex post facto*, the

trial in which the intervention failed to meet expectations usually results in some correlation data and some local knowledge. In the latter case, for instance, a regional recession may have occurred during the trial, thus limiting the usefulness of employment measures in a trial comparing exit and job reentry programs for ex-offenders. Lots of people were out of jobs, including the ex-offenders in the experiment.

One way of uncovering these unforeseen variables and helping others foresee them when deigning or evaluating an intervention could be as basic as "after the fact, as before the fact" speculative scenarios and "logic models" like Weiss's "causal" diagrams, referred to earlier. They are inexpensive ways to portray *ex ante* what is likely to happen and *post facto* what might have happened. They are useful only to the extent that they embody counterfactuals. They are vulnerable to the extent that we cannot measure everything well as part of the trial in anticipation of the intervention's failure.

Any post mortem is perforce speculative as to what caused the failure to meet expectations. But explicit *ex-post facto* theory as well as empirical evidence can help to make the process of understanding more transparent. At least, laying out an explicit logic or theory can help to establish whether the theory is disprovable. And if the data are at hand, we might then test the data's fit to the model even if causal inference must remain equivocal.

Q5. How might we build a cumulative knowledge base on when, how, and why failure occurred?

Good institutional vehicles are in place to cumulate and synthesize some kinds of knowledge. The Campbell Collaboration (http://www/cammpbellcollaboration.org) in the social sector, the Cochrane Collaboration (http://cochrane.org) in the health sector, the Coalition for Evidence Based Policy (http://coalition4evidence.org), the Institution for Education Sciences'

What Works Clearinghouse (http//:whatworks.gov), and Slavin's (http://bestevidence.org) initiative in the education sector are for learning about what does work *and* what does not. Their standards of evidence are demanding and reasonably clear.

But these organizations' missions have to be augmented, if the aim is to learn more from failures to meet expectations. They do not pursue reasons *why* "A" failed to do as well as expected as opposed to "B," partly because we lack a sturdy intellectual scaffolding for doing so.

For research policy people, this begs the question "Should we bother with estimating rates of failure?" A response might be drawn from one of John Graunt's (1662/1973) rationales for his statistical tome, notably "good, certain, and easy government." These days, his phrase is dressed up as evidence-based policy. Another of Graunt's responses to the question can be paraphrased as "because it is interesting and fun to do this"

There is a normative response. Developed countries depend heavily on spontaneous reporting and surveillance systems in regard to accidents, like the rate monitoring for railway accidents and airplane crashes and resultant reports produced by the National Traffic Safety Board. The data and the associated analyses are a basis for understanding our society's progress in identifying the reasons for failures and averting them. They lead us to create backup systems and to build redundancy into circuits in anticipation of failure.

The Food and Drug Administration's post marketing surveillance system for medical devices and procedures looks at the empirical rates of failure. Analysts then try to deduce possible causes and possible consequences, and changes to rates of failure that come from modifying the interventions. Gilbert, McPeek, and Mosteller (1977), for instance, tried to assess innovations in primary surgery. Focusing on randomized trials, they toted up the frequency with

which innovations appeared to work better, or worse, or had no discernible difference relative to ordinary practice. They learned that the rates were about 32%, 21% and 47% respectively.

There is plenty of opportunity for honest and valuable work along the same lines, and to drill deeper.

Concluding Remarks

There are three main justifications for the lines of thinking we propose and for envisioning a larger research agenda on the topic: the inevitability of failure in all human endeavors, the paucity of scholarly research on the topic, and efficiency.

First, failure in innovative human enterprise is inevitable and abundant. Failure's incidence and character, and learning how to learn from it, need to be better understood in education, crime prevention, social services and welfare, and other sectors.

Second, there is an obvious absence of orderly and transparent approaches to studying failure in these sectors. This is unlike medicine, where despite imperfections and institutional missteps, post mortems are part of the science. It is unlike engineering, where thematic books on better engineering through failure are not uncommon.

The third justification concerns efficiency of effort. Social initiatives that are tested in randomized controlled trials do not routinely investigate failure despite the commonness of reports of "no progress," "no discernible effects," and "null statistical findings." On the rare occasions in which failure to meet the expectations about the intervention's effect are taken seriously, results of failure analysis are not published. This failure to examine failure and to capitalize on what can be learned from failure to meet expectations, in an orderly and transparent way, is inefficient in many senses. We can do better.

Footnote

The work reported here has been supported partly by the National Science Foundation's PRIME Program. We are grateful for that support. The opinions expressed here are the authors' and the NSF ought not to be blamed for any of them.

References

Baum, J. and Dahlin, K. (2007) Aspiration Performance and Railroads' Patterns of Learning from Train Wrecks and Crashes. *Organization Science, 18*,(3), 363-385.

Besharov, D. J. (2009) From the Great Society to Continuous Improvement Government: Shifting from Does It Work? To What Works Better? *Journal of Policy Analysis and Management, 28*(2), 199-220.

Boruch, R., Merlino, J. and Porter, A. (2011) Report on the Center Research, Cognitive Sciences and Science Education. Presented at a Conference Arranged by the Institute for Education Sciences, march 21-22 2011, Washington DC.

Bovens, M. and 't Hart, P., (1996). *Understanding Policy Fiascos.* Transactions, New Brunswick.

Cambridge University Press (2013). *Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers.* www.englishprofile.org

Dewey, J., (1933). *How We Think.* D.C. Heath & Co., Lexington.

Du Maurier, G. (1895). True Humility, *Punch, 9* November.

Florman, Samuel (1996) *The Existential Pleasures of Engineering*. New York: St. Martin's Press (Second Edition).

Garet, M., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., (AIR), Bloom, H., Doolittle, F., Zhu, P., Sztejnberg, L. (MDRC), Silverberg, M. (IES) (2008) *The Impact of Two Professional Development Interventions on Reading Instruction and Achievement*. Washington DC: Institute or Education Sciences, US Department of Education.

Gawande, Atul (2009) *The Check List Manifesto: How to Get Things Right.* New York: Metropolitan Books

Gilbert, J. P., McPeek, B., and Mosteller, F. (1977) Progress in Surgery and Anesthesia: Benefits and Risks of Innovative Therapy. In J.P. Bunker, B.A. Barnes, and F. Mosteller

(Eds). *Costs, Risks, and Benefits of Surgery*. New York: Oxford University Press, pages 124 – 169.

Graunt, J. (1662/1973) *Natural and Political Observations Made Upon the Bills of Mortality*. In P. Laslett (Compiler) The Earliest Classics: Pioneers of Demography Gregg International.

Gray, P., (1996). Reviews. *Public Administration, 74*(3), pages 552-553.

Gornitzka, A., Kogan, M. & Amaral, A. (Eds) (2005). *Reform and Change in Higher Education: Analyzing Policy Implementation*. Springer, Dordrecht.

Levitt, B. and March, J. (1988) Organizational Learning. *Annual Review of Sociology. 14,* 319-340.

Lindblom, C. E. (1959). The Science of "Muddling Through." *Public Administration Review, 19*, 79–88

Peters, G.B., (1997). Reviews, *Journal of Political Science, 59*(1), pages 259-261.

Petroski, H., (2012). *To Forgive Design: Understanding Failure.* Harvard University Press, Cambridge.

Pressman, J.L., & Wildavsky, A.B., (1984) *Implementation: How Great Expectations in Washington are Dashed in Oakland: Or Why it's Amazing that Federal Programs Work at All.* (3rd Edition), University of California Press,

Pronovost, P. and others (2006) An Intervention to Decrease Catheter-related Bloodstream Infections in the ICU. *New England Journal of Medicine, 355,* 2725-2732.

Ruby, A. (2010) Thinking of Opening a Branch Campus? Think Twice. *The Chronicle of Higher Education.* (March 21, 2010)

Rutty, G. N.(Ed.), (2001) *Essentials of Autopsy Practice*. London: Springer.

Sabatier, P. (2005) From Policy Implementation to Policy Change: A Personal Odyssey. In Gornitzka, A., Kogan, M. & Amaral, A. (Eds) (2005). *Reform and Change in Higher Education: Analyzing Policy Implementation.* Springer, Dordrecht, pp17-34.

Secrist, H. (1938) *National Bank Failures and Non-Failure. An Autopsy and Diagnosis*. Bloomington Indiana: Principia Press.

Sherman, L. (1992) *Policing Domestic Violence: Experiments and Dilemmas.* New York: Free Press.

St. Pierre, Robert, Swartz, J., Gamse, B., Murray, S., Deck, D., and Nickel, P. (1995) *National Evaluation of the Even Start Family Literacy Program: Final Report.* Washington DC: U. S. Department of Education.

Strang, Heather (Issue Editor) (2012) Managing Field Experiments. *Journal of Experimental Criminology,* Volume 8, Number 3.

Valiknagasn, L., Hoegl, M., and Gibbet, M. (2009) Why Learning from Failure Isn't Easy (and What to Do about It): Innovation Trauma at Sun Microsystems. *European Management Journal, 27,* 225-233.

Weiss, C. (2002) What to Do until the Random Assigner Comes. In F. Mosteller and R. Boruch (Eds) *Evidence Matters.* Washington, DC: Brookings Institution, pages 198-224.