To Flop is Human:

Can We Invent Better Scientific Approaches to Anticipating Failure to Meet Expectations in Randomized Controlled Trials and to Learn from the Failure?

Robert Boruch and Alan Ruby

University of Pennsylvania

12/9/11 6:30pm Draft

Preamble

Health professionals have a disciplined approach to post mortems and autopsies at the individual level, and sometimes at the hospital unit level so as to learn from medical failures. Engineers have "safety factors" in designing structures and systems so as to avert unpredictable failures and a visible, though not always orderly, tradition of learning from their failures. Books on the topic in each arena are in ample supply. Some are good.

The applied social sciences, including education sciences and criminology, enjoy less clarity in understanding failure for a variety of reasons. A bridge collapse, for instance, is often plain and spectacular. The failure of an intervention program in education or of a police intervention aimed at crime control is often quieter. Despite journalistic hyperbole and political theater, the outcome is not spectacular and many reasons for it rarely transparent.

More to the point, the applied social sciences lack disciplined , well developed, and transparent 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. We focus on randomized controlled trials in what follows simply because the cause effect relation is clearer than it is in other contexts.

Failure Aversion

Of course, one can always define failure out of existence, or define it in a way that reduces its ostensible frequency. Corrections colleagues in some Australian states have done so, for instance, by using nuanced definitions in the context of a performance indicator, the outcome variable, called "escape from prison." An escape is not an escape, for example, unless the body is missing for 24 hours. If, after 24 hours, the convict returns to the prison voluntarily, this is denominated as an "unexcused absence." If our convict is released temporarily to attend college courses outside the prison, a worthy enterprise, but does not come back, this is labeled at least temporarily as "failure to return," as opposed to "escape."

It is not very difficult to find similar examples/shenanigans in education. For instance, it has taken over twenty years for the US States to agree, more or less, on embracing only two hs taken two decades at least to rech more less transparent definitions of "school drop-out" in the states of the United States. In the medical sector, here and abroad, "cause of death" is similarly nuanced and oriented toward the benign or the grim depending on who is counting. Readers may need no reminders about issuances on output and performance indicators employed by Goldman Sachs, Lehman Brothers, and other financial institutions, and by Bulgaria and Greece and other governments, on their financial well-being. They employed evidential standards that are outside the ones used in our more transparent scientific efforts.

Circumlocutions are just that, attempts to avoid labeling an event as a failure or admitting error. There is a deep seated aversion to acknowledging shortcomings in performance. Some 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 (Valiknagas et al 2009). It reminds us of our vulnerability as humans. These traumas and the wish to avoid them lead us at times to learn less from our own errors than prefer to learn from the errors of others. Baum and Dahlin (2007) illustrate the idea in the context of unnatural experiments, notably train wrecks.

Despite common aversion to thinking seriously about failure in mny social sectors, colleagues with a taste for evidence from well controlled trials have urged the idea that "R and D strategies should be planned with failure in mind" (Besharov 2009 p.210). In the organization research arena, Levitt and March (1988), for instance, have spawned interesting work based on the theory that organizations can learn and may learn from their failures as well as successes. It is in the spirit of these sources of counsel and experience, and others including Boruch (2007), that we explore the topic here.

Scientific Vulnerability of Randomized Trials

Scientific evidence-based initiatives are also vulnerable on account of the outcome variables used and how they are measured. But, unlike governments and the for-profit sectors, science depends on a bit more transparency at least with respect to methods of producing the evidence.

Randomized controlled trials, for example, when done right, permit fair comparisons, on account of the randomization. They permit a legitimate scientific statement of one's confidence

in a relatively unequivocal causal inference. The declaration that "A" worked better than "B," under a particular statistical test of the null hypothesis is legitimate and satisfying, and gets lots of attention.

In particular, we members of the tribes called "statisticians," or "methodologists," are vulnerable on subtler grounds. For instance, we help design randomized trials on interventions in a variety of social sectors. And we analyze results that often do not permit a conclusion that "A" works any better than "B." That is, we declare that the mean difference in outcomes of two interventions are not significantly different. Technically, the failure to detect a statistically significant difference in outcomes for the interventions tested in a trial is itself a scientific success, provided that the trial is designed well and done right. Good on the tribes for this line of thinking.

For a good trialist, it is proper and conventional to declare that there is "no statistically significant difference" in outcomes between A (the new and expectedly better intervention) and B (the control) when indeed no difference that meets a standard in a formal statistical test of a null hypothesis is found. Saying that "A failed" in this context is wrong on statistical grounds. Recall papers on statistical hypothesis testing by Weisburd, Lum, and Yang (2003) and Boruch (2007) in criminology and Wainer (2007) in education research, among others.

This common scenario—discerning no remarkable difference between interventions tested--presents a nice opportunity to think about how to get beyond the conventional declarations in a randomized trial that compares A to B. In particular, 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, especially in randomized trials, and

to exploit and advance theory and practice in education, social services, policing, corrections, and other sectors.

The Questionary's Approach to Null Findings

We indulge here in an interrogatory approach. This is to provoke conversation and to educate ourselves. We use the word "questionary" advisedly. During the medieval inquisition, the questionary was the chap who asked questions of the souls being tortured. Rebarbativeness – unpleasantness- of this sort is not our intention. Neither the questions nor proposed answers to them are out of bounds for correction or criticism. The questions ut simply are:

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

Q2. What empirical approaches might we use to estimate rates of failure to meet expectations?

Q3. How might the interventions that are tested in randomized trials be designed so as to reduce the likelihood of failing to meet expectations?

Q4. 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?

Q5. How might we learn about plausible reasons for failure to meet expectations *ex post facto* in a scientifically and orderly way?

Q6. How might we build cumulative knowledge base on when, how, and why the failure occurred?

These questions of course constitute a research agenda. Herewith are some tentative answers to the questions. We leave it to abler readers to get beyond these answers or to ask different ones.

The Questions and Tentative Responses

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

This question begs a predecessor at least in the research policy arena. *Who* indeed defines "failure?" In this, we are exquisitely sensitive to Nelson Algren's aphorism: "The vocation of assessing the failures of better men can be turned into a comfortable livelihood, provided you back it up with a PhD. " Algren, a writer without benefit of the more advanced degree, appears to have got burnt frequently by academic critics. The temptations of Monday morning quarterbacking and *post facto* finger pointing (digital omni-directional causal inference), are esteemed in circles apart from the literary variety.

We have no easy answer to the "who?" part of the question except to aver this. It is the scientific communities, not legislatures or the courts, who are best equipped to identify failure to meet expectations in scientific contexts. To the credit of the courts, the judicial system has not been eager to engage such issues but does have recourse to standards of scientific evidence issued for instance by the Federal Judicial Center. This, at times, has not prevented one researcher from suing another in quarrels about what one may conclude from a failure to replicate an analysis.

This first question begs another antecedent. Why bother to define failure at all? Doing so complicates life. A reasonable answer is that unless we properly define failure of a tested program, we cannot define its success, much less declare when either occurs. And we cannot dodge the matter by talking about "mixed effects."

Further, we aver that it is "failure to meet expectations" about an intervention's value that is of primary interest in educational, criminological, and other randomized trials. The scientist hopes that "A" will be better than "B" in a fair trial, for instance, otherwise would not

bother to make a fair test. But the correlative aim is to test the more conservative expectation, the hypothesis, that there is no appreciable difference between the two intervention's effects. The trial itself is no failure, if done right, regardless of the outcome.

The failure to reach expectations might then 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 value probability as a threshold condition. It is essential in the sense of avoiding our deluding ourselves into thinking that an effect that is found is dependable rather than a matter of chance.

For applied statisticians and other scientists, the "effect size," defined in a standardized way, is at least as important nowadays as a probabilistic threshold and it is more important than the latter depending on the stage of research and the research design. The effect size is the mean difference in outcome adjusted for the inherent variability of the groups being compared. This statistic is perhaps more important when sample sizes are small and most important in the long run of small trials on the same intervention. Conscientious methodologists can concoct standardized effect size indicators based on acceleration rates or deceleration rates, retardency rates, decay or sustainability rates, and replication. See Borenstein (2009) on technical definitions and variations.

To keep things simple in what follows, we define "failure" narrowly and *as the failure to meet expectations about an "effect size."* The rationale is this. The expected effect size, or minimum detectable effect, is the scientifically accepted basis for designing randomized controlled trials that are sensitive to the expected effect of the tested intervention. In particular, this expected effect size is the basis for statistical power analysis. Borrowing from colleagues in the humor production trade, the matter can be put into less reverent form. If you buy the

premise, you buy the joke. If you buy the expectation of effect size in the design of the study, you have bought an expectation about possible results.

Q2. What empirical approaches might we use to estimate rates of failure to meet expectations?

Neurotically conscientious statisticians are entitled to question their own questions. Why indeed should we bother with estimating such rates of failure? By way of answer, we depend on one of John Graunt's (1662/1973) responses to a similar question about the reasons why he generated his mighty tome on statistics, the first ever, in the 17th Century. The collection of these statistics is in the interest of "good, certain, and easy government." In contemporary terms, think of such data as an evidence-based approach to science policy. One of Graunt's other responses to the question boils down to (we paraphrase) "because it is interesting and fun to do this." Graunt, incidentally did not get to failure rates except in the crude sense of counting deaths. But he did anticipate the invention of social indicators that are now used by the World Bank, among others, to judge a nation's health.

A more contemporary rationale for the question is normative. Developed countries, for instance, depend heavily on spontaneous reporting and surveillance systems in regard to accidents. Witness 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, e.g. Baum and Dahlin (2007) on train wrecks. Witness also the post marketing surveillance system run by the Food and Drug Administration for medical devices and procedures and augmentations of the system so as to understand rates, possible causes and

possible consequences, and changes to rates that are a function of learning about the failures. In the entrepreneurial sector, estimated rates of failure of innovation run from 40% to 90% to judge from work cited in Valikangas et al (2008).

Assuming the question is legitimate, more or less as framed, how do we do we address it? Gilbert, McPeek, and Mosteller (1977), for instance, tried to understand how often innovations in primary surgery worked to help patients rather than to kill them or do no better than conventional surgical methods. They focused on randomized trials, toting up the number of times that innovations appeared to work better, or worse, or had no discernible difference relative to a control condition. They learned that the rates were about 32%, 21% and 47% respectively. The rate of failure to meet expectations in this sector, at the time, and with the reports available then might be estimated as 68%. Roughly similar rates were produced in a paper 20 years ago by Glass and colleague in the social sector.

We started to wonder if our colleagues had stumbled on the social science equivalent of Mendelian ratios' but this earlier work can be criticized easily. These authors relied on articles published in refereed journals which, at the time, were more likely to contain papers reporting successes than papers reporting on failures. They did not have the repetitive empiricism of the pea experiments.

Nowadays, we are in a better position to characterize the odds on failure. For instance, the Institute for Education Sciences, a government entity in the US, embodies a brave approach to identifying all studies on particular interventions and to review those studies for quality and results so as to make judgments about the value of the study and of the particular innovation. The results are published by the IES's What Works Clearinghouse (2011). The Campbell Collaboration (2011) in the social sector and Campbell's older sibling, the Cochrane

Collaboration (2011) in the medical sector aim to be as thorough in their coverage of literature, published and otherwise, in reviewing particular innovations or classes of innovations. The Coalition for Evidence Based Policy does similar things but smaller resources and narrower aim. These organizations are independent as possible from political influence.

Joshi and Boruch's (2009) reconnaissance on the reviews issued by the IES's What Works Clearinghouse suggest failure rates in the 10-50% range. The range is large and depends on the particular education area that is the target of controlled trials or high end quasi-experiments and on how one construes failure (effect size small or lack of discernable effects beyond chance level, etc.). The US curriculum publishers' rate of failure to generate any trustworthy evidence at all is far higher, in the neighborhood of 80% or more (Whitehurst, 2009).

No efforts to empirically estimate "failure to meet expectation" rates based on reviews of controlled trials have been undertaken yet under the auspices of the Campbell or Cochrane Collaborations. The opportunity to do so looms large for these, however, and for related organizations such as the Coalition for Evidence Based Policy (2011). The challenges, as our British colleagues might say, are "not unformidable."

Q3. How might the interventions that are evaluated in RCTs be designed so as to reduce the likelihood of failing to meet expectations?

For a moment, forget our tentative answers to questions 1 and 2. Reckon that we would just like to get a fat effect size in a fair randomized controlled trial and that the only way we can get one is through design of the intervention. That is, assume that successful execution of the statistical design of the trial in manageable and assured. Those of us who merely design the

trial, Boruch for instance, are often cut out from the privilege of advising on the intervention's design. Intervention developers who test their own intervention in a trial can do this. We now bolt from this tribal constraint so as to anticipate happier findings.

To address the question at hand, let us first be attend to counsel of Al Reiss, statesman in criminology, who declared the following during our advisory duty for the Spouse Assault Replication Project (SARP). In paraphrase, Reiss said: "Design the new crime prevention intervention, A, so as to look at lot different in particular ways from conventional practice, B. Otherwise there's not much point to mounting a controlled trial on the difference in outcomes between A and B."

In the Spouse Assault Replication Program (SARP), for instance, Sherman and other colleagues (1992) had to assure that people got arrested for misdemeanor domestic violence rather than merely being lectured by a cop and let off the hook. The "A" intervention, arrest, looked a lot different from "B," no arrest, to many of us including theoreticians interested in the specific deterrent effect of arrest n recidivism.

Arrests were indeed carried out but, to the surprise of many, had no discernable effect on recidivism. This is *possibly* because A and B do not look different *to the perpetrator*. Arrest, in point of fact, is often not a remarkable deprivation of liberty. Rather, it can be construed as a transient event with little consequence to the perp. Its deterrent value, a day or less in the pokey, if one makes bail, is low. The refinement on Reiss's law is that A and B have to look a lot different "to the target population of the intervention". Doubtless there are counterexamples, with the big dose being more effective. Understanding when and why it is effective, or not, seems important.

Reiss's counsel is related to the engineering theme of safety factors. The latter engenders the basic idea of designing A so that you'd get what you'd like to see on theoretical grounds. Then, multiply the resources by 2 or 3, "a safety factor," so as to take into account uncertainties that are inevitable in the field. Such a safety factor in structural and other engineering can be most accurately construed as an ignorance factor. Calling the thing a safety factor is better. We, in the social and education sectors, lack a thematic emphasis on safety factors in design of interventions. In principle, at least, the notion of planning for our ignorance based on earlier failures and safety factors seems worth exploring. See Ruby and Boruch (2011) on Dewey versus Pharaohic thinking.

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, for example. Ruby (1995) implied the idea in the area of education. The recent 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. Heidi decided to boost business volume by permitting her customers to sign IOUs instead of paying directly. She borrowed money on these notes to pay suppliers and indeed banks bought the notes in batches expecting that payments and profits ensue. The bar flies did not pay their debts, and the bar went belly up, as did the banks that bought the notes.

Medical writer Atul Gawande (2009) considered related topics in his "Checklist Manifesto." Gawande's theme is that one ought to develop lists of things that are necessary to assure that, in effect, our expectations are met. He describes how such checklists are operationalized at length and how they are used in construction work, airplane safety, hospital

procedures, and other areas. Checklists in each of these arenas can be book length. For scientists in the social sector, equally thorough checklists are in short supply whether the program is an educational curriculum package or a crime prevention program. The checklist idea, of course, depends on foreknowledge of what is essential and what could go wrong. Any such checklist approach is itself subject to empirical testing, as in the case of prospective cohort studies in medicine such as Pronovost et al (2006) , so as to learn about their effectiveness..

In some of the applied social sciences, we have cultures of due diligence. Rap sheets or arrest records in are illustrative at the individual level, and exploited well by criminologists and cops. At the intervention level, however, there is not yet a clear and well articulated tradition of due diligence in characterizing planned innovations. Somehow, the delicious opportunity to invent ostensible improvements at the intervention level often suppresses or supplants the ordinary, even pedestrian, need for due diligence and checklists in planning improved interventions. We place a premium on innovation over integrity when it comes to process and routine, valuing the novel over the known.

As another framework, consider an approach that combines the ingredients that we have already covered in some respects: SWOT. Regular attention to "Strengths, Weaknesses, Opportunities, and Threats" (SWOT) is part of some corporate efforts in planning to reduce failures to reach expectations. We are indebted to colleagues at the American Institutes for Research for helping us understand the idea and how it might be actualized. Whether SWOT can be exploited well in designing educational or criminological or other social interventions remains to be seen. There are no empirical examples as yet. SWOT deserves some attention in this context partly because it invites transparency and order, and is scientific in the limited sense of taxonomy.

Though we distrust the hyperbole attached to the phrase "systems theory" in this context, we would be remiss if we did not mention the idea's importance as a vehicle for designing interventions and trials so as to anticipate failure. Consider the following example.

A cluster randomized trial in Philadelphia and Pittsburgh, mounted in 2007, depended on a nominally well designed study and nominally well designed intervention to test the intervention's effect on science knowledge of middle school students. The adverb "nominally" here means that all theory and evidence at hand was used to deploy the trial in roughly 30 schools. That is, information about local parameters such as number of schools, number of teachers within schools/classes, and number of kids within classes was exploited for statistical power calculations. Cognitive science principles, developed over the past two decades, were used to revise science curriculum modules so as to enhance kid's achievement.

What was not taken into account fully in design of the intervention, much less the trial, was systems related. In particular, Boruch and Merlino (2011) found that ambient positional instability (API), the "churn," among teachers in the school system is potentially critical and that this higher level systems factor ought to be taken into account. The API indicates, for instance, that 42% of teachers in Philadelphia had taken a position in September 2011 that is different from what they had in September 2010. About 46% did so in Pittsburgh. The reasons for such instability are complicated, vary, and are doubtless localized. They include medical, maternity, and sabbatical leaves, subject area reassignments from science to math for instance), grade band reassignments, and assignments to administrative duties, among others. This notion is a remarkable bridge between the local randomized trial experience and the work at a higher level of systems by Boe et al (xxxx), for instance, on teacher supply, demand, attrition, and transfer.

The point here is that intervention designs in education, as in other sectors, usually assume stability at the delivery system level. They do not, as in this case, assume instability at a higher levels of the system. More to the point, it is reasonable to assume that interventions will not achieve an effect of an expected size unless Ambient Positional Instability is taken into account. And further, this has to be taken into account in the design of the randomized trial. "Attrition," for instance, is taken seriously in the What Works Clearinghouse standards for judging the quality of the execution of the trial's design. But such standards and good practice does not focus on the underlying reasons for people's disappearance from a trial scenario nor make a bridge to the trialist's interest in the implementation of the intervention.

Q4. 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 methodologists who design trials, do not ask this question. We usually figure we've done our duty in designing a randomized trial well, so as to test a formal null hypothesis fairly and to come up with a dependable estimate of the intervention's effect and its variability. As the Royal Statistical Society's 19th century escutcheon says, *Aliis Exterendum.* In other words, let others thrash it out. Able statisticians nowadays get beyond the Society's nineteenth century slogan. Statistical analysis is now closely tied to the design of controlled trials. But we are only beginning to get to the point of encroaching on the intervention's design or the matter of designing studies so as to anticipate the intervention's failure and learn from it.

In recent years, for instance, some trialists in education and criminology have done well in getting beyond "black box" trials in ways that anticipate the possibility of failure to meet expectations about the intervention's effects. Measuring the extent to which cops participate in actualizing a new intervention in the trial is integral to good practice in criminological hot spots studies and others for instance (Boruch, Weisburd, Berk, 2010). In education, Garet and his colleagues (2008), for instance, have advanced understanding of how to measure implementation/deployment of the intervention in professional development programs in education during the course of the trial showing that teachers learned but students did not. More generally, Grubb emphasizes getting into the field, in notably the classroom, too understand, for "without such understanding it's impossible to know the reasons for failure (quoted in Besharov (2009) page 211). The ground level work seems desirable, but so is work at higher levels in te systems that contain and affect the intervention being tested. More about the systemic approach later.

Future implementation research is likely to focus on how media can be better exploited. Recall for instance Stigler, et al's (2000) pioneering work to understand from video recordings of Japanese classrooms how the Japanese manage to teach mathematics as opposed to how we teach it in the U.S among other comparisons. Newer efforts by the Educational Testing Service involve 360 degree real time video recording of everything in sight in a classroom. Learning how to dimensionalize the material and how to analyze it in multiple ways that produce ideas about missteps, poor pedagogy, and so on is a remarkable challenge.

Assume, however, that we methodologists will be profoundly ignorant in designing some, perhaps many, trials. Assume that we will not be able to plan for ignorance except in the sense of learning as we go.

In an engineering tradition, for example, this learning often requires a "run in period," dedicated to stabilizing the system to be tested and to working out the kinks in both the intervention and the study's design. 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 and may have to abstain from analysis of outcome data from it on account of all the missteps engendered by the run in period.. The second cohort is dedicated to the real comparison of the interventions and estimating the actual effect size. If things work well in both cohorts, that is to the good. One may then have doubled the sample size and consequently increased the study's ability to detect modest effects of the intervention.

Sherman and others involved in SARP seem to have first used the tactic in criminological trials. Lottery based "rollout" trials are a natural vehicle for gradually reducing ignorance in real time in the interest of reducing the likelihood of failure to meet expectations. In other contexts, the matter may come under the rubric of "mid-course corrections." Spybrook (2008) and others , for instance, are busily identifying how statistically related mid-course corrections in a trial are made so as to better detect the intervention's effects, i.e. better estimate the effect size and assay its dependability. We are aware of no analogous effort that examines the changes made in an educational or criminological *intervention* during the early stages of a trial that then lead to mid-course corrections in the design or operationalization of the intervention.

In all of these frameworks, "trouble shooting" in the early stages is essential. This involves getting into the street with serious attention to detail about who is doing what, with what incentives and resources, and who is not, with what disincentives and resources. In this stage of ignorance, we then depend on local/focal theory to guide observation and questions, and

on street level intelligence from any potentially dependable source. Anthropologists, innocent and otherwise, denominate such thinking as developing "grounded theory." Trialists and members of trial advisory boards (at their best) engage in "ride-alongs" to uncover issues cop experiments (after purchasing body armor of course). The best of education researchers who engage in randomized trials alsoengage in classroom observations and talks with principals, teachers, or parents, and so on. This is in the interest of better design and to trouble shoot the trial's execution, e.. Kim (2010).

Trouble shooting of this sort, regardless of the trial's specific design, is tradecraft. It is a marketable commodity, of commercial value in the contracting industry and in academia, based on experience that is very good at times But it is laden with potential embarrassment. The latter perhaps is why trouble shooting that reduces the likelihood of failure of the intervention is not often written up in peer reviewed research journals in education, criminology, or other sectors. The future lies with developing more transparent and orderly approaches to the activity and to learning how to describe the results so as to get beyond the tradecraft.

Suppose now that we do indeed have prior theory, as opposed to a grounded theory that is invented on the run. And further assume that we can plan to measure things in the trial based on the prior theory. That is, suppose we have a theory about how things are supposed to work andhow things might *not* work, and that we build measurement systems in trials to recognize both. Even if we cannot exploit the measures in real time, we might be able to exploit evidence so as to learn after the trial. This seems good, and is related to the next question.

Q5. How might we learn about plausible reasons for failure to meet expectations *ex post facto* in an orderly and scientific way?

This is the hard part. Unlike engineers, we in the applied social sciences, including education and criminology, cannot execute trials to failure so as to understand at what point the intervention (a structural support, for instance, for engineers) fails. Unlike colleagues in the pharmaceutical industry, we cannot do trials in which deliberate low or high doses in different animals can be tried out so as to determine what is too weak to assist 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 etical trials as yet so as to test directly the causes of failure.

Further, an intervention's developer may simply fail to recognize the failure to meet expectations relative to standards of evidence that are determined by people other than the developer. One such developer in the intervention arena declares in its advertizing that the study of the interventions effects was recognized by the What Works Clearinghouse as "fully meeting" the WWC's evidence standards. In fact, the WWC had found that the intervention package had indeterminate at best.

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 very similar to one invented by Robert St. Pierre and colleagues (1995) in their reports on the first randomized field tests of the Even Start Family Literacy Program. The effects of the intervention in the first large randomized trial were negligible. Ditto for the second, but the latter is beside the point here.

In regard to item (a), good standards exist for assessing the quality of a trial's design and its execution. Helpful checklists abound. See website for the IES's What Works Clearinghouse in education, Campbell Collaboration, and Cochrane Collaboration, among others. If the trial wasn't done right, or was sabotaged, we still know nothing about the benefits of A over B. So… we may take a shot at another trial. The early trials on enriched oxygen environments for babies born prematurely, which led to retrolental fibroplasia and no appreciable decrease in their mortality, are a case in point (Silverman, 1980) but we believe there are many others. The earlier remarks on two cohort trials, run in periods, trouble-shooting etc. are pertinent here. We spend no more time on this.

Let us also spend no additional attention on the second question, item (b), as to whether the interventions were delivered as advertised. It is fundamental and 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), about 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 much more challenging. *Ex post facto,* the trial in which the intervention failed to meet expectations usually results in some correlation data and some local focal knowledge that the trialist may have picked up along the way. This is unless we figured out how to generate measures that ex post facto would have informed our analyses. For instance, the local focal knowledge may be that a regional recession/business

close-downs occurred during the trial. Further the trialist may plausibly speculate that this is reason why the trial comparing exit and reentry programs for ex-cons failed to show any effect. Lots of people were out of jobs, including the trialist's target group.

The matter is rarely so obvious as in the foregoing example. There are usually many factors that are related the intervention's failure to meet expectations. For instance, Carol Weiss (2002), among others, has encouraged the practice of laying out a causal path diagram to represent how a community based training intervention can produce effects. Her diagram involved over 20 causal arrows going, thoughtfully laid out, but to the naïve eye appearing to go hither and thither. More important, she had the audacity to lay out a second diagram portraying how the intervention might go belly up. It displayed yet more causal arrows. Comrade Weiss may have got well beyond our cortical capacity in the matter of causal linkage arrows. One implication is that there are always far more ways to fail than to succeed.

Although Weiss employed causal path diagrams with a zeal that others may not admire, the notion is right. After the fact, as before the fact, "logic models," "causal path diagrams," etc. are cheap and inexpensive ways to portray ex ante what 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 failure or in post mortems. For a scary example from engineering, read MacDonald's (2010) report on how BP misestimated, and /or poorly measured, and misreported oil leakage volume in the Deepwater Horizon episode in the Gulf of Mexico. The range in reported volumes is stunning. The arguments about competing explanations for the failure were lengthy and remain less than settled.

Of course, it would be good to put the failure to meet expectations in a particular large scale trail in the context of earlier and smaller trials (bench science) in which expectations were met. Barghaus and McMacken (2010) reviewed the available literature to learn and to invent ideas about how to do so. They built on others work to invent and justify a scale up effect that ties the results of larger scale tests to the smaller ones. In brief, their *scale up index* is an ratio of the median effect sizes from large scale stuies to the median effect sizes of samller scale studies where the controls are much greater than in the scale up variety. They proposed a *superrealization bias index* by inverting the scale up index. This indicator reflects the degree to which effect sizes form small scale studies are elevated relative to one that one finds in the larger studies. Their arithmetic, for instance, would suggest that an effect size of an effect found in the bench studies would be 40% of what one would fid in the field studies. But they and we lack the right percentage multiplier as yet. And it is likely to change over time. The point is that people, like Barghaus and KcMacken can speculate well about failure to reach expectations in the context of a contemporary scientific history,

Any post mortem has to be 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.

Q6. How might we build a cumulative knowledge base on when, how, and why failure occurred?

Some good institutional vehicles are in place to cumulate and synthesize knowledge. Search vehicles such as Google, Bing, et al are not among them, though these engines may provide ingredients for study. To judge from experience at Microsoft Asia's Bejing Office, colleagues in the industry aim to assure that we can identify 6 million "relevant" sites in .005 seconds instead of merely hitting on 3 million sites in .05 seconds. This is a fine aim. But it is not entirely satisfying for those of us interested in the quality of evidence, or for scientific understanding what works and what does not.

The Campbell Collaboration in the social sector, the Cochrane Collaboration in the health sector, the Coalition for Evidence Based Policy, and the Institution for Education Sciences' What Works Clearinghouse and Slavin's Best Evidence initiative in the education sector are among the more interesting instruments for learning about what does work and what does not. . The URLs for the web sites are given in the reference list. Their standards of evidence are demanding and reasonably clear. The WWC's standards are the most clear and neurotically conscientious.

The missions of these organizations may have to be augmented, however, if the aim is to learn more from failures to meet expectations. The organizations are oriented toward aggregate statistics on particular interventions and to reaching a dependable scientific conclusion about whether "A" works better than "B." They allow a causal inference about what works. None have provisions for orderly analyses of plausible reasons why "A" failed to do as well as expected as opposed to "B." That is, they are not yet well equipped to handle flops and the accruing knowledge about them, partly because we do not yet have a sturdy intellectual scaffolding. Determining whether claims of evidence about purported success arfe actually warranted has been a difficult challenge over the last decades and the organizations deserve high credit for trying to meet that particular challenge.

Concluding Remarks

There are three main justifications for encouraging the lines of thinking that we have proposed 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 all of these sectors. This is unlike medicine, for instance, where despite medicine's imperfections and institutional missteps, post mortems are part of the science. It is unlike engineering, where thematic books such as Engineering through Failure are not uncommon.

The third justification concerns efficiency of effort. Educational initiatives, for example, and others that are tested in randomized controlled trials do not routinely investigate the initiative's 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.

Footnotes

Some of the ideas in this paper were broached at the annual meeting of the Center for Evidence Based Crime Policy at George Mason University in 2010. The Youtube version is given on CEBCP's web site at http://gunston.gmu.edu/cebcp. The ideas were also broached at a meeting organized by the Institute for Education Sciences in March 2011, and we are grateful also for that opportunity to vet ideas..

References

Barghaus, Katherine and McMacken, Jennifer (2010) Scaling Failure: A Review of Success Rates and the Scale Up Effects in Education Research. Paper Prepared for ED/CRIM 871 University of Pennsylvania.

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.

Best Evidence Encyclopedia (2011). http://bestevidence.org.

Boe, E (2009?!)

Borenstein, M. (2009) Effect Sizes for Continuous Data. In H. Cooper, L. Hedges, and J. Valentine (Eds) *Handbook of Research Synthesis and Meta-analysis*. New York: Russell Sage Foundation, pp.222-235.

Boruch, R. F. (2007) The Null Hypothesis is Not Called That for Nothing: Statistical Tests in Randomized Trials. *Journal of Experimental Criminology, 3*, 1-20.

Boruch, R. and Merlino, J. (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.

Boruch, R., Weisburd, D. and Berk, R. (2010) Place Randomized Trials. In A. Piquero and D. Weisburd (Eds) *Handbook of Quantitative Criminology*. New York: Springer, pp.481-502.

Campbell Collaboration (2011) http//:www.campbellcollaboration.org.

Cochrane Collaboration (2011) http://cochrane.org

Coalition for Evidence Based Policy (2011) http//: coalition4evidence.org

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.

Joshi, P. and Boruch, R. (2009) The Probability of Failure and Success in the Education Sector. A Descriptive Report Based on the What Works Clearinghouse Reviews of Evidence. Draft Report. Center for Research and Evaluation on Social Policy, Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania.

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

Petroski, Henry (1985) *To Engineer is Human: The Role of Failure in Successful Design*. New York: Barnes and Noble.

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. (1995) Benchmarking the Pursuit of Quality. *Unicorn, 12*(2), 43-47.

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

Ruby, A. and Boruch, R. (2011) Dewey, Meet the Pharaohs: What Educators Can Learn from Engineers. Report 10.2011. University of Pennsylvania, Graduate School of Education, Philadelphia, Pennsylvania.

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

Slavin, R. and Lake, C. (2008) Effective Programs in Elementary Mathematics: A Best Evidence Synthesis. *Review of Educational Research. 78*(3), 427-515.

Spybrook, Jessaca et al (2008) Are Power Analyses Reported with Adequate Detail? Evidence from the First Wave of Group Randomized Trials Funded by the Institute of Education Sciences. *Journal of Research on Educational Effectiveness. 1*, 215-235.

Stigler, James, Gallimore, R., and Hiebert, J. (2000) Using Video Surveys to compare Classrooms and Teaching across cultures: Examples from th TIMSS and the TIMSS-R Video Studies. *Educational Psychologist, 35*(2), 87-100.

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.

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.

Weisburd, D., Lum, C. M., and Yang, S. (2003) When Can we Conclude that Treatments or Programs "Don't Work?" *Annals of the American Academy of Political and Social Sciences. 587*, 31-48.

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.

Whitehurst, G. (2009) *Rigor Relevance Redux: Director's Biennial Report to Congress*. Washington DC: Institute for Education Sciences, US Department of Education.