Impact Evaluation and Welfare Evaluation

Glenn W. Harrison

August 2013

C.V. Starr Chair of Risk Management and Insurance,
Department of Risk Management & Insurance, and
Director, Center for the Economic Analysis of Risk,
Robinson College of Business, Georgia State University, USA.

Forthcoming, *European Journal of Development Research*, 2014.

The expression "impact evaluation" means different things to different people, but to most economists now it means the use of randomized controlled trials (RCTs) or "quasi-experiments." I want to focus on that side of the research and argue that, even when statistical circumstances are ideal, it constitutes at best an intermediate input into the welfare evaluation of interventions. That intermediate input is valuable, but should not be confused with the final product, a proper cost-benefit analysis. I summarize arguments here that are developed at greater length, with literature citations, in Harrison [2011a][2011b][2013].

There are several simple reasons that such impact evaluations (IE) fall short. The first is that they limit themselves to evaluations of observables: this price change in delivering that product leads to what revealed change in demand? The second is that they limit themselves to average effects: what is the average change in demand? The third is that their focus has been partial equilibrium: what is the effect of this price change for the commodity I am able to randomize (or study the randomization of)?

Looking for Keys Where the Light Is Better

The problem with just looking at observables is that they tell us nothing about the latent variables that are of interest in welfare evaluation. For that we need to make inferences about consumer surplus, and for that we need to know a lot more about latent preferences that people bring to their choices, such as risk preferences and time preferences. We also need to know a lot more about the subjective beliefs that people bring to their choices. The reason that there is this dogmatic focus on observables is easy to discern and openly discussed: a desire to avoid having to take a stand on theoretical constructs as maintained assumptions, since maintained assumptions might be wrong. The same methodological precept guides the choice of statistical methods, but that is another story about modeling costs and benefits. One can fill in these blanks in our knowledge about latent preferences and beliefs with theories and guessed-at numbers, or with theories and estimated numbers. But one has to use theory to make conceptually coherent statements about preferences and beliefs, and then undertake welfare evaluations. That is the rub: an agnosticism towards theory.

Advocates of these IE often portray the tradeoff here in overly dramatic fashion. Either one uses the methods that avoids these theoretical constructs, or one dives head in to the shoals of full structural modeling of behavior. This is a false dichotomy, raised as a cheap rhetorical device to still debate over the role of theory. The missing middle ground becomes apparent when empirical puzzles emerge, leading to casual theorizing and even more casual behaviorism, sadly illustrated in Banerjee and Duflo [2011] and Karlan and Appel [2011].

Gaussianity

Why the fascination with the average? On a good statistical day, it is one measure of central tendency, that is true. But there are many reasons why we are directly interested in knowing the full distribution, not just one *in*sufficient statistic.

First, we might simply care about winners and losers. Assume that there is a modest change in the average, in some direction that the researcher deems a welfare improvement. What if this comes about with large gross changes at the individual level: lots of people do wonderfully from the intervention, and lots of people do terribly from the intervention? One does not have to swerve too far from the strict Utilitarian social welfare concept to wonder if this is indeed a welfare improvement.

Tradeoffs between efficiency and equity aside, we might also be interested in identifying winners and losers in order to design a better intervention, in the spirit of the compensation criteria of welfare economics. Or to design a more robust intervention that could survive rent-seeking attacks from losers, and hence be more politically sustainable.

Second, we might care about the distribution when evaluating the "policy lottery" that any intervention affords a decision-maker. One reason is to extend consumer sovereignty in welfare evaluation to consider the risk attitudes of those affected, if there is some statistical risk that any given individual is a winner or a loser. Another reason is to reflect uncertainty aversion or ambiguity aversion, arising from imprecision in estimated effects (e.g., due to "intent to treat" slippage twixt lip and cup in RCTs). Either of those two, which are often confused terminologically, require that one undertake welfare evaluation over the *distribution* of impacts, whatever specific modeling church one attends. These specific modeling alters vary in how they weight the distribution, but they all agree that the essence is to take it into account in some way: that is, in fact, what differentiates uncertainty and ambiguity aversion from "familiar" risk aversion.

Third, and assuming away the statistical identifiability of whether any given individual is a winner or a loser, and even assuming away equity concerns, we might care deeply about the distribution shape of things to come from an intervention if it makes more people vulnerable to certain thresholds. In development, the most important threshold is the absolute poverty level, defined here as that level of resources below which the individual unit experiences some asymmetric physiological effects. For now, equate resources with income. Imagine an intervention that keeps the fraction below that poverty line the same, bunches a lot of people ε above the poverty line when they were well above it prior to the intervention, and somehow allows the "rich and famous" to enjoy gains such that the average income of the population increases. Surely the tsunami of vulnerable individuals hovering ε above the poverty line, compared to the baseline, should matter for our welfare evaluation? Again, to see the fundamental point, rule out equity effects, and rule out risk aversion (or even uncertainty aversion) with respect to the estimated impact of *this* intervention in *this* domain. What if there are "background risks" that might nudge this tsunami below the poverty line, even after we have ascertained the impact of the foreground intervention under study? It is a commonplace in developing countries, magnificently documented by the *Portfolios of the Poor* of Collins, Morduch, Rutherford and Ruthven [2009], that the poor face myriad risks at any given time, and value flexibility in risk management options.[1]

Worms, Teachers, Fertilizer and Savings

One can certainly be interested in worms and whatever they do, absentee teachers and whatever they do not do, the optimal use of fertilizer, wherever it comes from, savings rates, and so on. But these are not substitutes for the rigorous measures of welfare from a policy, given by the equivalent variation in income. We need these measures of welfare for the application of cost-benefit analysis familiar to older generations: comparing a *menu* of disparate policies potentially

[1] An important technical point must be made here: these issues arise naturally and conventionally if one assumes non-additive utility defined over multivariate risk. However, it is an unfortunate commonplace to assume additive utility in many applications, such that the risk over *final* wealth positions is all that matters.

spanning all of these interventions. How do I decide if it is better to reduce worms, increase teacher presence, use fertilizer better, or increase savings rates, if I do not know the welfare impact of these policies in a way that allows comparability? Of course, the best intervention might be "costless" to implement, but that is rare.

A related concern is the sample selection effect that comes from only doing IE on things that one is allowed to randomize, or that serendipity randomizes for us. What if we care about an intervention in some area that does not permit randomization, such as tariff policy? How do we then trade off interventions in the areas we can study, with those in the areas that we cannot study (or cannot study for the foreseeable decision-making future)?

It is often difficult to design a careful RCT quickly, not because of any flaws in the method, but because of the logistical constraints of coordinating multiple sites and obtaining necessary approvals. Worrall [2007; p.455-459] presents a detailed case study of a surgical procedure which was identified as being "clearly beneficial" on the basis of observational studies, but where it took years to undertake the requisite RCT needed for the procedure to become widely recommended and used. Lives were lost because of the stubborn insistence on RCT evidence before the procedure could be widely adopted. Of course, counter-examples probably exist, but the costs and benefits of having complementary evaluation methodologies are often lost in the push to advocate one over the other.

This Debate

Lensink [this issue] correctly points to the problem of the missing counterfactual as the core issue facing the type of IE I consider, and the essential role of theory in addressing how one fills that void. Whether one buries it in arcane statistical acronyms or not, theoretical assumptions are being made, and not always attractive ones, a point made forcefully many years ago by Rosenzweig and Wolpin [2000]. Even without the assumptions in arcane acronyms, we have to know the preintervention baseline, or have some other convincing control group. And we have to know that the act of randomization is not itself generating some sample selection process. I disagree with two points he makes:

- The RCT *approach* is hardly the "brainchild of Esther Duflo," and it is not hard to find antecedents in the statistical literature, and even among economists doing social experiments many, many decades ago.
- I also question if an RCT always improves the internal validity of an evaluation: what if I am interested in making causal statements about non-observable, latent variables? For instance, statements about the effect of some intervention on consumer surplus or welfare? As conventionally practiced and applied, this class of IE is silent on issues like these.

Picciotto [this issue] stresses that there is "old school" IE and "new school" IE, and that distinction is an important one to remind youngsters of, just as all Australians need to know the difference between fresh-water crocodiles and salt-water crocodiles. He also makes the same point about IE not being complete if it does not have the welfare evaluation component, although he uses terms such as "merit," "worth" and "value," which amount to the same thing. He makes several incorrect claims about the new school, however:

- That they do not explicitly attend to whether impacts are short-term, medium-term or long-term. Their stress on short-term impacts just reflects their passion, and astonishing skill, at getting published in the best places as soon as possible.
- That the limitations of their method derive from "the experimental method." Here the mistake is to equate field experiments with randomization, when the latter just happens to be one of the tools that all experimenters use at one time or another.

Gujit and Roche [this issue] write in a disciplinary code I do not quite understand, although I keep getting the feeling that I am supposed to agree with them. Take the three core purposes of impact evaluation they commend, and see if they can be translated for econmists. First, we must learn to "improve" as well as "prove" what works. Even if we forgive the implied acceptance of the metric "what works," is this just saying that we need to address the spillage 'twixt lip and cup when we go from positive economic insights to normative economic recommendations? Agreed, but the way we do that in economics is by knowing or guessing more about the structure of the decision-making process than is usually provided in an RCT, so that we can make some informed guesses about how the sentient "seeds" we randomized on one day will react on another day or place. Second, we have to be concerned with accountability for the funds being used. Here we get a matrix of forms of accountability (their box 2), none of which manage to mention, even in code, the one form of accountability I care about: the expected net welfare of the folks we are studying. Everything here is process oriented, and riddled with a jargon that is normally reserved for UN documents. The process matters, as any economist studying principal-agent theory knows, but it is not the first, or second, thing I would worry about when assessing the state of impact evaluation. Third, "influencing for empowerment." Silly me, I read this to mean that the folks we are studying should find something of interest to empower themselves. But all it means is "make your data and code available." Whatever failings I have identified in modern impact evaluation, that is not one of them (e.g., http://microdata.worldbank.org/index.php/catalog/impact_evaluation/about, http://dvn.iq.harvard.edu/dvn/dv/jpal, or the data links for published papers in major journals in economics).

My confusions get compounded when we then move to "what matters" for impact evaluation (§3), and are already on alert since we are moving from the "what works" slogan to an equally meaningless "what matters" slogan. We are asked to consider a "forward-looking debate about IE that transcends epistemological squabbles and methodological fights." Hold on, Beavis. Epistemology is about what we claim as knowledge: one of the core failings of modern impact evaluation, in my view, is the fear of making knowledge claims about latent concepts, such as welfare, risk and time preferences, and subjective beliefs. Not a squabbling matter at all. And methodology is driving these omissions: if you use a method that is only looking, by design, at a restricted set of observables, you simply cannot make knowledge claims about those latent concepts.[2] Three things have to be done:

- We are told that "standards matter," and wander off in a foggy trail of "seven clusters of tasks [that] need careful consideration." But the problem is that the modern impact evaluations are just poor economics: they do not deliver the well-known, but well-forgotten, basics of cost-benefit analysis.[3] Apparently I critique impact evaluations for something called their "precise inaccuracy," and I have to admit to having no clue what that refers to.
- We are told that "rigour and relevance" matter. The valid point here is that alleged rigor with respect to causal statements between selected observables is touted as critical, and other methods applied less rigorously (e.g., eliciting risk preferences with hypothetical surveys, despite decades of evidence of the biases of doing that). But when the best we can do is stress the importance of "relevant rigor" *and* "rigorous relevance," I have to shake my head and wonder if this is Marketing Slogans for $100 on *Jeapardy!*
- We are told that "power and politics" matter. No doubt these "power relations" affect the choice of evaluation methods: try applying for a grant from the standard places without an acknowledged "randomista" on board, or with a proposal critiquing those methods. So that does make it hard to do studies that show the limitations of the RCT method as it is

[2] To expand, one has to look at observables in order to infer these latent concepts, but at a much wider set of observables than is standard in impact evaluation. Examples include observed choices in controlled experiments designed to elicit risk and time preference, or subjective beliefs. I suspect the next generation of impact evaluations will do this; it is not difficult.

[3] Cost effectiveness is one component, and an important one, of cost-benefit analysis. They are not the same.

> currently applied, and the best way to improve things is, of course, just to do them better. But envy at the impressive way that one group has captured the hearts and blocked the minds of funding agencies should not be confused with the need to just sort out the correct application of economics to conduct cost-benefit analysis. Nor does that need much attention to the "politics of evaluation," just some straightforward scholarship.

White [this issue] argues cogently for more studies, and strikingly for more studies using traditional observational methods as well. Some attribute the saying that "quantity has a quality all of its on" to Stalin, but the myriad issues of development surely benefit from the additional attention, from some of the smartest places, being heaped on them. He correctly notes that there are ways to mitigate sample selection bias, and indeed ways to infer unobservables more generally. Frankly, the academics doing RCTs are sharp enough to do these things, and do them well, when they want to, and will just call them "augmented RCTs" and keep marching on. That is actually a fine outcome.

I have one significant qualification, and it is only from something implied by White [this issue] rather than something he says:

> Random assignment is a method. It has no inherent ideology. One could as easily randomly assign incentive structures under a central planning regime as in a market economy. But in practice proponents of randomization are part of the current atheoretical approach to economics dominant in much of the United States in which the behavioural assumptions required for modelling are abandoned in favour of empiricism (Harrison [this issue]). This point of view necessarily supports development through nudges rather than big pushes, as the latter requires more behavioural assumptions, assumptions which are embodied in theory. This debate is reflected clearly in Easterly's [2007] critique of planners – the big push of Sach's Millennium Villages being one of his main targets – compared to seekers who favour small scale innovation and experimentation.
>
> I think the best we can say is that the jury is still out on this one. I am sympathetic to the idea that structural transformation requires deeper seated changes. However, challenged that the Indian public health system is broken so what is need is systemic reform rather than giving away plates as an incentive to parents to bring their children to be immunized, Esther Duflo replied that it would take years to achieve such reform, so what is the harm in giving out some plates to get children immunized now.

> I do see her point, but in the end, when policy makers and programme makers need to know if is particular programmes work.

This is all very well put. But we need to be very careful to not presume that the opposite is true: that "nudges" can be supported by atheoretical insights. White [this issue] does not say this, but it might easily be inferred from what he writes. Again, I believe the core problem is the use of the slogan "what works," as suggested at the end of the above extract. That one can indeed just focus on "what works" is far from evident, although many people take it for granted. Take the question of take-up of insurance: it is far from obvious, without knowing why people take up a product, that this is a good or bad thing for them.

This is one reason I take sharp, personal aim at many specific claims of this "I only care about what works" ilk in Harrison [2011b]. If one is passionate about development goals, even goals in the small, we cannot be causal and agnostic about the proper use of theory and econometrics because we are looking at local nudges. This simply does not follow from the excellent point that White [this issue] makes, that global, non-nudges do require some theism, and indeed huge swaths of it.

Concluding Maxims

"No gold-digging for me... I take diamonds! We may be off the gold standard someday." Mae West is not often a source of methodological insight, but we often hear that an RCT is the Gold Standard in medicine, and that this should be what we unwashed social scientists should aspire to. Such claims get repeated without comment, but, to quote a popular political refrain in the United States, advocates of RCTs are entitled to their own opinions but not their own facts. It is far from obvious that RCTs and observational studies dominate each other in the medical domain, when one does careful comparisons in modern times. Let's just decide on these things for ourselves, using our own

costs and benefits of alternative research methodologies for different inferential objectives.

Just don't drink the Kool Aid! These are exciting times in development economics, with more attention, more dollars, and some of the brightest minds being devoted to real problems of poverty and public policy than we have seen in a long time. But the gratuitous methodological precepts that seem to come along with the use of randomized evaluations should be a source of discomfort to scientists. The marketing claims of advocates of randomized control are easy to identify, but they come cloaked in seductive phrases such as "evidence-based economics" or only being interested in "what works." How can one stand against such things? [4] Well, rather easily if you see what they entail in terms of avoiding many of the tough problems of development economics.

[4] As Groucho Marx once said, "The secret of life is honesty and fair dealing. If you can fake that, you've got it made."

References

Banerjee, Abhijit V., and Duflo, Esther, *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty* (New York: Public Affairs, 2011).

Collins, Daryl; Morduch, Jonathan; Rutherford, Stuart, and Ruthven, Orlanda, *Portfolios of the Poor: How the World's Poor Live on $2 a Day* (Princeton: Princeton University Press, 2009)

Easterly, William, *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So little Good* (New York: Oxford University Press, 2007).

Gujit, Irene, and Roche, Chris, "Does Impact Evaluation in Development Matter? Well It Depends What It's For!" *European Journal of Development Research*, this issue.

Harrison, Glenn W., "Experimental Methods and the Welfare Evaluation of Policy Lotteries," *European Review of Agricultural Economics*, 38(3), 2011a, 335-360.

Harrison, Glenn W., "Randomisation and Its Discontents," *Journal of African Economies*, 20(4), 2011b, 626-652.

Harrison, Glenn W., "Field Experiments and Methodological Intolerance," *Journal of Economic Methodology*, 20(2), 2013, 103-117.

Harrison, Glenn W.; Jensen, Jesper; Lau, Morten Igel, and Rutherford, Thomas F., "Policy Reform Without Tears," in A. Fossati and W. Weigard (eds.), *Policy Evaluation With Computable General Equilibrium Models* (New York: Routledge, 2002).

Karlan, Dean, and Appel, Jacob, *More Than Good Intentions: How a New Economics is Helping to Solve Global Poverty* (New York: Dutton, 2011).

Lensink, Robert, "What Can We Learn from Impact Evaluations?" *European Journal of Development Research*, this issue.

Picciotto, Robert, "Is Impact Evaluation Evaluation?" *European Journal of Development Research*, this issue.

Rosenzweig, Mark R., and Wolpin, Kenneth I., "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, 38, December 2000, 827-874.

White, Howard, "Current Challenges in Impact Evaluation," *European Journal of Development Research*, this issue.

Worrall, John, "Why There's No Cause to Randomize," *British Journal of the Philosophy of Science*, 58, 2007, 451-488.