Results vs Rigor: Balancing Tensions for Productive Evaluation Partnerships

In the demand for evidence, there exist long-standing tensions around the timing, content, and certainty of research and the results it yields. Such challenges faced by researchers and practitioners are often made more complex as they intertwine with the needs of policymakers and funders. Disagreement around the purpose of the research (internal learning vs. accountability) further complicates decision-making. This rings true from research conception to execution to dissemination.

At the core of these tensions lies competing incentives, both perceived and real. Researchers enter into partnerships to test hypotheses. They aim to execute ground-breaking research to advance their field. They want to publish, but anchor to the "null hypothesis." They want their partners to succeed, but are also skeptical. On the other side are implementers and practitioners. They are deeply invested in the success of the program, and have to be comfortable working in the grey to decide what is reasonable when it comes to evidence and programming. Within organizations, there is constant tension between the desire to learn and be transparent, and the need to advance the organization. Research for the real world is rooted in mutual dependence. Constructing successful partnerships requires all actors to acknowledge and address these competing incentives.

Grounded in the belief of the power of rigorous evidence, our four panelists started as research analysts and diverged to becoming a professor, a government advisor, an M&E practitioner at a pre-primary and primary education social enterprise, and evaluation expert at an NGO working to make secondary school more effective. We explore three tensions through specific panelist experiences and their work in sub-Saharan Africa. Specifically, we explore tensions around the timing of research, the certainty required of rigorous research versus the certainty needed to make decisions, and the tensions around the research process through the lens of six specific projects:

1. A randomized evaluation on Educate!'s flagship program in Uganda, with end-of-course results and a 4-year follow-up study to shed light on long-term skill, labor market and other life outcomes;

2. A randomized evaluation of a high-profile pilot program in Liberia with multiple non-state actors managing public schools, with 1-year results disseminated in 2017 and 3-year results anticipated in 2019, all amidst a government transition;
3. A randomized evaluation in Rwanda, designed to test the impact of Educate!'s model when delivered through teachers and integrated into government infrastructure;
4. A potential randomized evaluation turned quasi-experimental baseline study of Bridge in Nigeria, examining differences both within and across school types, commissioned by DFID;
5. A qualitative evaluation with IDinsight, conducting a deep examination of the pathways through which Educate!'s program in Uganda operates, designed to answer specific internal questions; and
6. A randomized evaluation within Bridge to determine whether cross-age tutoring could work and if so, how Bridge could best use scheduled instructional time.

Through authentic conversation from people on opposite sides of the tensions, we offer guidance and a few cautionary tales to help the community interrogate their process for embarking on research and debunk the myth of independence.

ELEMENT 1: TIME (Stir Crazy)

- The model lifecycle of an evidence based program could be as follows: an innovation is piloted at a small scale, the operational kinks are ironed out, it is then scaled up in a way, and with a large enough sample, to be rigorously evaluated. But there tend to be fewer resources and smaller research incentives for early evaluations of small-scale programs. To get any benefits of research, or of interacting with academics, organizations may feel pressured to conduct an impact evaluation sooner than they'd like, even if the program is premature. Evaluation too soon may stifle innovation and progress or prematurely stunt a project with high potential. Not evaluating may leave decision-makers without evidence to act, or waste precious resources on something that is not working. Are impact evaluations the only path to evaluating effectiveness? Who decides what methodology to use?
- High-quality research often takes time, but often everyone wants the research yesterday. In a setting where we want to do the most good for the most people

without causing harm, what are the dangers of waiting to share insights versus sharing too soon? When is waiting a luxury we cannot afford; when is sharing the cause of unintended negative consequences? And if an organization knows it will have to make a decision before results are ready, should it bother pursuing the evaluation at all?

ELEMENT 2: CERTAINTY (Star Crazy)

- Decisions are easy when answers are definitive, but the existing body of research rarely provides definitive answers. Policymakers and practitioners have to rely on the research that exists; researchers are often wary of drawing definitive conclusions prior to a preponderance of the evidence. How can researchers help practitioners weigh the evidence? What if the evidence is from a different context or at a different scale? What if the parameters of the policy are different? It is unlikely that every iteration of a program can be tested.
- How do policymakers, funders, and practitioners weigh statistical significance? If impact cannot be detected at a 95% confidence level, what odds are reasonable odds? When impact is detected, is it always meaningful?
- Is the role of the researcher to describe the way things are or change the way things are? How can a researcher do the latter responsibly? What if such conditions cannot be met?

ELEMENT 3: INDEPENDENCE (Star Power: Is independence under-scrutinized, misunderstood, or a red herring?)

- Justin Sandefur writes "Independence in impact evaluations is woefully under-scrutinized. Even researchers acting entirely in good faith would benefit from clearer codes of conduct to protect them and the integrity of their results." At the same time, even though a researcher may be independent from the funder or the implementing organization, the mere act of choosing the researcher may introduce bias. With exactly the same data and context, different researchers will draw predictably different conclusions. Self-interested organizations will account for this in their selection - sometimes to avoid an unnecessary critique and sometimes to tip the balance in their favor. Is independence therefore under-scrutinized, misunderstood, or a red herring?
- Funders of interventions also fund evaluations of the intervention; what rules should guide this relationship?

- Conversations with practitioners are key: They have an in-depth understanding of the context and can be critical partners when designing data collection and fieldwork. They also can offer insights to help researchers interpret the findings, surface factual inaccuracies, and point out text that is unclear, misleading, or could be misunderstood. Why are such conversations not concerning?
- People don't like to hear evidence against their preconceived views. This is true for researchers and practitioners alike. But sometimes the dissonance is not driven by bias but rather a different set of facts. What is the difference between confirmation bias and having enough background to notice that something seems unreasonable or feels “off”?