Research Design Challenges and Moving Results to Action: An Interview with Thomas Cook

Ellicott Matthay, our E4A Postdoctoral Scholar, sat down with E4A National Advisory Committee member Dr. Thomas Cook to discuss some of the methods challenges facing population health researchers, possible approaches to overcoming those challenges, and how these issues impact real-world decision-making. Dr. Cook is widely considered an expert on causal inference design approaches, having written extensively on the topic and been awarded numerous accolades for his work. He is a Research Professor at George Washington University and a Professor Emeritus of Sociology, Psychology and Education at Northwestern University.

Matthay: What do you think are the most important method challenges in population health research?

Cook: One of the most important challenges is how to justify for population health a set of causal methods other than randomized assignment that empirical evidence shows generate causal estimates like those from an experiment. I think the only observational study mechanisms we can trust are those that stubbornly reproduce causal results similar to those from a randomized experiment with the same intervention, same outcome measurement, and same causal estimate as the observational study of interest. For me, is the key question is: How to warrant causal practices other than randomized assignments? And the best answer to this question is by showing similar causal results to a randomized experiment with largely similar operational details other than the way the intervention is assigned.

Matthay: This makes me think of the four main branches of study validity - internal, external, statistical, and construct validity. It sounds like the emphasis for you is really on internal validity because that is where the randomization plays out most saliently. Is that right?

Cook: All four kinds of validity are important, though I would say internal validity is the most important for cutting edge causal issues. That internal validity is not enough is exemplified by fields that are characterized by very frequent randomized experiments. In social psychology, for example, almost everybody does randomized experiments, but that does not silence scholarly debate. There are still big fights about the interpretation of those studies, fights about statistical conclusion validity, about external validity, and mostly about the construct validity of the cause, but sometimes of the effect too. So even if we solve the issue of internal validity in population health research, that does not mean there aren't a lot of issues that would still need to be resolved, especially around:

- What is the intervention and how can it be labelled in abstract, general terms?
- What are the components of the intervention that make the difference?
- What is the outcome when expressed in general and theory-relevant language?
- What other outcomes might be affected by the intervention?
- Given the outcomes the intervention did affect, what is the effect of those outcomes on later sequelae in a longer causal chain?
- How can we give "meaning" to the size of any one result and to the pattern of effects?

So while we might say internal validity is the most important, it's not the same thing as saying it's the only thing that's important.

Matthay: That makes sense. That was really well articulated. Do you think that the most pervasive method challenges in population health have changed over time? Or do you think that it's really been the same of key issues, over and over again?

Cook: I have the sense that it's quite the same for most of population health. But there are some obvious and important changes. One is the recent growth of concern in medicine about translating recent micro-biological knowledge into more macro-biological actions that affect clinical practice or pharmacology. This change is in medicine, of course, but population health has a role to play in tracing out its consequences for public health, including the control of health costs. At a more traditional population health level, I am struck by the better and broader measurement of biological processes via, say, dried blood or spit. We can now extend traditional survey methods to include analyses of physical processes and endocrinology by virtue of the ability to measure lead, iron, blood pressure, C-reactive protein, IL6, and all those other things. And finally, I am struck at the broadest policy levels by the political debates about system change in health care. It cannot stay the same, reflecting the same political forces emanating from the insurance and hospital industries. There will be serious change at some policy levels as yet to be determined that will be larger than even Obamacare. Population health has to be at the forefront of monitoring and reporting on what happens on a broad front of policy, implementation, and health and cost implications. I think these are three main emerged and emergent novelties, but understandably there is much of the same old.

Matthay: So the three novelties you point out are: (1) the measurement piece that we've dramatically improved on; (2) the link population health needs to develop to translational processes in medicine that seek to take the products of bench science and do more things with them faster for medical practice; And (3) the concern with larger system-wide issues of health policy formulation, funding and provision. Is that right?

Cook: Yes. The first two are somewhat related. It's the measurement that gets you into assessing biological processes at a population level, and it's the discoveries in medical science that point us to identifying what methods of broad-scale assessment need to be developed for the population level. The interface between molecular biology and the more macro processes that population health deals have to have a firm foundation in molecular biology - you can't do without that. And that's changing very rapidly.

Matthay: That makes sense. And that's affecting all the types of validity?

Cook: Yes.

Matthay: Do you think there are particularly promising solutions or areas of inquiry that are being pursued or could be pursued to address some of the most fundamental methods challenges?

Well if the most serious issue methodologically is how to justify causal interference from observational studies, then I would say there are no new designs being evolved. There are a lot of new data analytics being evolved, but most of them I would categorize as dancing on a head of a pin. Most of them are not fundamental – they produce minor, more precise estimates. I would actually include informatics and data science in that category. Data scientists and informaticists

are doing a great job of prediction, but they're not going to do a much better job of causal inference. In my view, the best thing going on in causal inference is the conduct of design experiments in which you try and find out which non-experiments consistently yield similar results to randomized experiments. If the findings are consistently similar, then the observational study results are empirically valid.

There are now about 80 studies using this design experimental method to test which observational study designs and analysis procedures produce the same results as randomized experiments. We've shown, for example, that regression discontinuity does provide the same results as randomized experiments at the cut off, though this is not theoretically surprising. We have also shown that comparative interrupted time series studies largely produce the same results as randomized experiments when it comes to the change in means from pretest to posttest, although there are at present only eleven studies of this. As for non-equivalent control group designs – what some call difference-in-difference designs – the combination of a local comparison group, a pretest measure of the study outcome, and a rich set of other covariates also seems to generate causal estimates similar to those from a randomized experiment testing the same causal hypothesis using the same intervention and outcome. Only when we know which observational study designs we can dependably trust will we know which are so poor we should not trust them at all.

Matthay: Of course, this only applies to the subset of questions can be answered with a randomized trial or with a regression discontinuity design. Not all questions lend themselves to those designs.

Cook: Yes, that is the fundamental limitation of this method today, but sometimes observational studies are compared against both experiments and regression discontinuity studies and in the area of comparative interrupted time series at least it makes no difference which is used. Moreover, many might come to believe in the future that comparative interrupted time series studies work often enough to be dependable as causal benchmarks. If that were to be the case, then we can compare observational studies against comparative interrupted time series. In the future, it might be possible to get around that limitation more often than today. Moreover, if a certain kind of observational study is generally unbiased or minimally biased against an experimental benchmark, does this not change the odds it will be unbiased in some domain where an experiment is absolutely impossible?

We're doing randomized experiments now on topics that would never before have been thought conceivable. I was at a population health meeting about a month ago with researchers who study pediatric screening. Two kinds of causal issues came up. The first and least important was "What are the determinants of successful quality screening– factors that get a lot of people into screening and results in very few false positives or false negatives?" That's good screening. That kind of causal question can and should be answered with randomized experiments varying possible individual determinants of good screening, including multivariate packages with many such individual determinants within them? When I asked this group how many randomized experiments were there with pediatric populations on what constitutes high quality screening, they all shook their heads. I was disappointed since such experiments are do-able. The more

important question to the group was: "What are the consequences of screening children?" Now, that question does not lend itself so easily to randomized experiments because it is ethically inappropriate to withhold screening from kids. So it's very hard to do randomized experiments on the consequences of pediatric screening versus non-screening, though it's easy to do experiments on the consequences of different screening regimens or on the determinants of quality screening.

I am never sure that researchers have a good grasp of the best causal design and analysis in a given application. Rarely do I get the sense they know the relevant methods, those that are technically better, and what the trade-offs are. Nor do I see people running through their head about which methods they could use given constraints in the setting. Alas, I rarely see a lot of flexibility in thinking through a large range of a causal method options.

Matthay: That's fascinating, though. That seems like one of the most fundamental method issues in and of itself, as well.

Cook: I think so. I think we're being trained by too many narrow statisticians. They think more in terms of analysis than design, more in terms of precision than bias. I often hear researchers idolizing "propensity scores" due to their current standing in formal statistics. I don't mind using propensity scores, but they are not a causal method. They create a single dimension out of many potential control measures. That alone makes them useful. But they reduce all bias if, and only if, you have been able to select the right covariates – those that fully describe the process of selection into treatment. But that is unknowable in almost all quasi-experimental contexts outside of regression discontinuity. Data analysis via propensity scores can work. But it is impossible in any application to know that it has worked. It stresses analysis and glosses over the operational complexity of arriving at well measured covariates that measure the true but unobservable true selection process. Researchers who are not methodologists are on the lookout for novel tools, and these are much more often in analysis than design.

Matthay: I listened to an interesting talk earlier this week from a prominent social epidemiologist and one of the things he said was he thought that epidemiology was not having as much impact on policy as it should, in part because they were spending too much time on methodological research and methodological debates and not enough time on applied, policy-relevant research. I am curious whether you agree, whether the same is true in your field, or in population health more broadly.

We all go into the field we are because we want to do good and want to influence policy. But we want to influence policy with the closest version of the truth. The issue is how complete does the knowledge have to be in order to act. And that varies depending on the urgency of the problem you're confronting or doing research on. There are people who are say "methodology first" because they want to be able to improve the truth content for the next generation. And there are those people who say we have to do more applied research today because they want to be relevant today. This is a tension. There should be a tension. I hope there always will be a tension. When some people come out and say we should do X, I hope people in the background will say, "That would be reasonable. But let's consider the conditions under which X is

likely to be effective today, and let's also think about whether Y might be even better." This is a creative tension. While public health exists in less of a politicized arena than other kinds of research – like climate change, for example – it is still in a political arena, not a vacuum. As such, advocates have to be loud and sometimes over-claim or gloss over difficulties. This little part of the political context alone reminds us to be careful about charging in with well-meaning remedies that aren't but might be. There's a tension here we should all live with and appreciate. Some people will be actors in the policy space, while others will be commentators at various levels of remove. We need all.

The E4A Methods Lab was developed to address common methods questions or challenges in Culture of Health research. Our goals are to strengthen the research of E4A grantees and the larger community of population health researchers, to help prospective grantees recognize compelling research opportunities, and to stimulate cross-disciplinary conversation and appreciation across the community of population health researchers. We welcome suggestions for new topics for briefs or training areas.

Support for this note was provided by the Robert Wood Johnson Foundation. The views expressed here do not necessarily reflect the views of the Foundation.