EXTERNAL VALIDITY CONFERENCE

April 3-4, 2023

Titles and Abstracts

Institute for Advanced Study in Toulouse · Investir l'Avenir · France 2030 · Toulouse School of Economics · Université Toulouse Capitole

Graeme Blair, UCLA

Community policing does not build trust or reduce crime: Evidence from six coordinated experiments

Abstract:

Is it possible to reduce crime without exacerbating adversarial relationships between police and citizens? Community policing is a celebrated reform with that aim, which is now adopted on six continents. However, the evidence base is limited, studying reform components in isolation in a limited set of countries, and remaining largely silent on citizen-police trust. We designed six field experiments with Global South police agencies to study locally designed models of community policing using coordinated measures of crime and the attitudes and behaviors of citizens and police. In a preregistered meta-analysis, we found that these interventions led to mixed implementation, largely failed to improve citizen-police relations, and did not reduce crime. Societies may need to implement structural changes first for incremental police reforms such as community policing to succeed. We reflect on the benefits and challenges of coordinating interventions and outcomes in field experiments.

Sylvain Chabé-Ferret, TSE

Treatment (Mis)Allocation under Publication Bias

Abstract:

Publication bias has emerged in the last decade as a major impediment for the accumulation of scientific knowledge. In this paper, I study the consequences of publication bias for the use of scientific evidence by policymakers. I delineate a model where a decision maker uses published results to decide on which treatments to allocate. I show that publication bias distorts the optimal allocation of treatments through several mechanisms. First, under publication bias, more ineffective treatments are implemented than what would be deemed optimal without publication bias. Second, under publication bias, the allocation of treatments does not converge to the optimal one as more studies are added to the evidence base. Third, policy-makers can undo some of this bias by ranking programs that they wish to implement. I show that publication bias makes the actual allocation less efficient that the unbiased one. I also show that there are cases in which this approach is severely biased because publication bias does not preserve the ranking of treatment effectiveness. I find evidence for all these sources of bias in an empirical application where I compare the treatment allocation obtained using the results of pre-registered replications to the treatment allocation that comes out of published studies.

Issa Dahabreh, Harvard University

Learning about a target population by combining data from multiple sources: causal assumptions, study design, sampling, and estimation

Abstract:

In recent years there has been increasing interest in analyses that combine data from multiple sources to estimate some causal, predictive, or descriptive parameter of interest. Examples of such work involve "transportability" analyses that estimate causal effects in a target population by combining data from a completed randomized trial and a separately obtained sample from the target population; tailoring of prediction models to a target population in which outcomes cannot be ascertained (and related work on covariate shift / domain adaptation); causally interpretable meta-analysis; and various other "data-fusion" activities. Using the example of causally interpretable meta-analysis, we examine the delicate interplay between causal assumptions, study design, and sampling properties when learning by combining data from multiple sources. We argue that, though often ignored, study design and sampling critically impact the identification and estimation of target parameters, and therefore need to be considered on par with causal assumptions and estimation methods, which have attracted most attention to date.

Ray Duch, University of Oxford

Cash for COVID-19 Vaccines in Africa: A Financial Incentives Trail in Rural Ghana

Abstract:

We implemented a clustered randomized controlled trial with 8,854 residents in six rural Ghana Districts to determine whether financial incentives produce substantial increases in COVID-19 vaccine uptake. Villages were randomly assigned to receive one of four video treatment arms: a placebo, a standard health message, a high cash incentive ($10) and a low cash incentive ($3). Non-vaccinated subjects, assigned to the Cash incentive treatments had an average COVID-19 vaccine intention rate of 81% compared to the 71% for those in the Placebo treatment arm. Two months after the initial intervention the average self-reported vaccination rates for subjects in the Cash treatment were about 3.6% higher than those for subjects in the Placebo treatment 40% versus 36.5%. The verified vaccination rates of subjects in the Cash treatment arms exhibited more modest treatment effects: 70.3% of verified subjects had at least one dose of the COVID-19 vaccine compared to the 67.3% for those in Placebo. The low cash incentive ($3.00) had a larger positive effect on COVID-19 vaccine uptake than the high cash incentive ($10.00). There is no evidence of spillover effects of the financial incentives to subjects in non-financial treatment arms nor to non-treated proximate residents.

Michael Denly, IAST, **and Michael Findley,** University of Texas at Austin

External Validity for Social Inquiry (Book)

Abstract:

Social science's current focus on "all else equal"--i.e., experiments or natural experiments in empirics, as well as highly abstract theoretical work--needs revision to account for causal structure, shielding, and other concerns that imperil broader learning. To help social inquiry achieve its goal of fostering more consistent learning, we argue that researchers, reviewers, and journal editors need to pay more attention to external validity. Part of obtaining better external validity involves better conceptualization, including by paying more attention to the distinction between populations and samples, generalizability and transportability, and the various dimensions of external validity: that is, mechanisms, settings, treatments, outcomes, units, and time (M-STOUT). Another part of obtaining better results on external validity is evaluating it, for which we proffer three new criteria: Model Utility, Scope Plausibility, and Specification Credibility. Thereafter, we show how to use the new evaluative criteria in various quantitative and qualitative methods. Finally, we argue that social scientists need to report on external validity accurately and provide relevant guidance to do so.

Julia Moeller, Leipzig University

Generalizability crisis meets heterogeneity revolution: Inductive and deductive approaches to shedding light on unknown boundary conditions

Abstract:

Whether research findings obtained in an empirical study generalize to other conditions is a crucial question that rarely receives the attention it deserves. The external validity can be limited for many reasons, some of which are widely known, whereas others are less often discussed. Among the typically better-known reasons are measurement issues, describing limitations to the external validity due to the problem that measures used in a study may fail to capture the exact phenomena in the real world to which a study's findings shall be generalized.

This presentation addresses less widely discussed limitations to the generalizability of research findings. It will present a taxonomy of possible boundary conditions that may limit the generalizability of a research finding. In particular, this presentation addresses in what ways research findings may be specific to the time points, contexts, and individuals in which they have been obtained.

I will then explain why the nomothetic and deductive research logic and the predominant between-person analytical methods uses in social sciences are only of limited use in our quest to understand time-, context-, and person-specific boundary conditions better.

As contribution to a solution, I will then present both inductive and deductive, idiographic and nomothetic, within- and between-person approaches to studying boundary conditions limiting the external validity of research findings.

The conclusion can be summed up as: The question not only if, but under which circumstances research findings are trustworthy should be asked more systematically. Following recent debates about generalizability in different disciplines, many different approaches of how to study unknown boundary conditions have been recently proposed. Some of them transcend the established deductive, nomothetic research logic and inter-individual comparisons typically used in social science studies. A next challenge will be to integrate the available research approaches within a new epistemological framework that enables us to ask and study systematically under which circumstances which research finding provides a meaningful description and/or prediction of the real world that we aim to describe with our research.

Jörg Peters, RWI - Leibniz Institute for Economic Research
(with Lise Masselus and Christina Petrik)

Lost in the Design Space? Construct Validity in the Microfinance Literature

Abstract:

Randomized controlled trials (RCTs) are at the center of the credibility revolution. While individual results often do not hold beyond the particular context under study, the accumulation of many RCTs can be used to guide policy. But how many studies are required to confidently generalize? Our paper focusses on construct validity, an important element affecting the generalizability of RCTs that is often neglected. Construct validity deals with how the operationalization of the treatment corresponds to the theoretical construct it intends to speak to. The universe of potential operationalizations is referred to as the design space. We use microfinance as an empirical example, a literature that is exceptionally rich in RCTs. We systematically review 38 microfinance RCTs to demonstrate that even this deep experimental literature only covers a tiny fraction of the design space and that small variations in the treatment operationalization matter for the observed effects. Construct validity is hence low if individual studies make general statements about the impact of microfinance – which most reviewed papers do. Construct validity could be high if a study semantically restricts its findings to the intervention under evaluation and abstains from generalizing to the construct microfinance – thereby trading relevance for rigor.

Cyrus Samii, New York University

Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference
(*joint with* **Michael Gechter, Rajeev Dehejia, and Cristian Pop-Eleches**)

Abstract:

We derive a formal, decision-based method for comparing the performance of counterfactual treatment regime predictions using the results from randomized experiments. Our approach allows us to quantify and assess the statistical significance of differential performance for optimal treatment regimes estimated from structural models, extrapolated treatment effects, expert opinion, and other methods. We apply our method to evaluate optimal treatment regimes for conditional cash transfer programs across countries where predictions are generated using data from experimental evaluations in other countries and pre- program data in the country of interest.

Beth Tipton, Northwestern Unviersity

Designing Randomized Experiments to Predict Site Specific Treatment Effects
(*joint with* **Michalis Mamakos)**

Abstract:

The evidence-based policy movement has long focused on providing estimates of the causal effects of interventions to decision makers. Increasingly, however, it is recognized that treatment effects likely vary across a variety of contextual and demographic factors. As a result, decision-makers are often less interested in the average effect and more interested in how well an intervention might work in their site or locale. To meet this need, one might provide subgroup average effect size estimates, or develop models that predict these site specific treatment effects. While answering the question posed by the decision-maker, one concern with these models is that if incorrect, they result in biased predictions. Another concern is that the predictions provided may not be very precise and may thus be less useful than the average effect estimate. In this paper, we begin by framing the problem as one in which the goal of an impact study is explicitly to predict site-specific treatment effects for a population of sites. We then consider how different estimators and sampling processes affect the average squared prediction error. The results indicate, for example, that the choice of average versus site-specific estimator has to do with degree to which variation in treatment effects can be explained, and that the concerns with sample selection bias found when estimating the average treatment effect are also found when predicting local effects

Eva Vivalt, University of Toronto

Weighing the Evidence: Which Studies Count?
(*joint with* Aidan Coville and Sampada KC)

Abstract:

We present results from two experiments run at World Bank and Inter-American Development Bank workshops on how policymakers, policy practitioners and researchers weigh evidence and seek information from impact evaluations. We find that policymakers and policy practitioners care more about attributes of studies associated with external validity than internal validity, while for researchers the reverse is true. We also find that researchers provided more accurate forecasts if they placed relatively less weight on factors associated with internal validity than their peers, while policymakers and policy practitioners provided more accurate forecasts of program impacts if they placed relatively less weight on factors associated with external validity. This result could reflect past learning under the bias-variance tradeoff.

Anna Wilke, Washington University in St. Louis

To Harmonize or Not? Research Design for Cross-Context Learning
(*joint with* Cyrus Samii)

Abstract:

To enhance generalizability, researchers increasingly study the same question in multiple contexts. Efforts at cross-context knowledge accumulation range from meta-analyses of studies that differ widely in design to coordinated initiatives that implement nearly identical experiments. We use a decision-theoretic framework to understand the conditions under which cross-study harmonization of research designs furthers learning about treatment effects. We model a decision-maker who conducts two simultaneous studies of the same treatment, each in a different context. Treatment effects consist of a common and a context-specific component. We first consider learning about treatment effects in-sample. By holding research-design related errors constant, harmonization makes it possible to attribute diverging estimates to cross-context effect heterogeneity. Hence, harmonization is optimal if treatment effects vary widely across contexts. Otherwise, research design diversity is preferable, because it reduces the correlation across estimates and thereby enhances learning about the common effect component. Predictions about unstudied contexts are based solely on the common effect component. A decision-maker who cares about out-of-sample predictions thus never wants to harmonize.