Die externe Validität des soziologischen Laborexperiments

Projektzeitraum: 01.10.2015-30.09.2017

Institut für Soziologie, Ludwig-Maximilians-Universität München
Konradstraße 6
80801 München

Dr. Marc Keuschnigg
Institute for Analytical Sociology, Linköping University
Norra Grytsgatan 10
SE-601 74 Norrköping
Schweden

marc.keuschnigg@liu.se
Tel: +46 11 36 33 39

This two-year project, funded by the German Research Foundation (DFG), tested the minimal requirements for external validity of social experiments. The “transportability” of experimental findings to other than the particular implementation entails that rates of observed behaviors and estimated treatment effects are robust to changes in the specific research setting and the sample under study. We evaluated (1) the sensitivity of laboratory results to locally recruited student-subject pools (*parallel-test reliability*), (2) the comparability of behavioral data collected online and, under varying anonymity conditions, in the laboratory (*mode reliability* and *reactivity*), (3) potential differences in elicited behavior between student subjects and the general population (*sample generalizability*), and (4), with a replication at Amazon Mechanical Turk (MTurk), the stability of experimental results across contexts (*international context generalizability*). Different samples, modes, and settings may violate transportability in that they produce different rates of observed behaviors and—more worryingly for experimental research—heterogeneous treatment effects.

Altogether, we collected behavioral data from almost 6 000 experimental subjects using the same decision interface. To identify those parts of experimental designs which undermine their generalizability, we focus on decision-making situations frequently used in social cooperation research: the Dictator Game, the Ultimatum Game, and the Trust Game. These games differ in

complexity, carry the potential for socially desirable responses, and permit direct comparison with an extant literature. From a sociological perspective, these games measure prosocial behavior and thus reveal expectations about valid social norms in a particular population and setting.

To evaluate *parallel-test reliability*, we conducted a multi-location laboratory experiment at two German universities in Leipzig and Munich. In parallel sessions we drew on two newly recruited student-subject pools whose members had little prior experience with experimental studies. This study investigates parallel-test reliability of social experiments.

Targeting *mode reliability* and *reactivity*, we tested for comparability of behavioral data collected online and, under varying anonymity conditions, in the laboratory. First, we addressed reactivity with a variation of anonymity conditions in our two physical laboratories. Second, we conducted parallel sessions online. Comparison of online and laboratory results isolates mode effects of experimental data collection.

To establish *sample generalizability* in a nationwide online experiment, we tested our baseline results' generalizability to the broader population. We sampled participants from a high-quality offline-recruited Internet panel. Our sample is representative of the German-born population with regard to gender, age, and administrative district.

Concerning the *context* of experimentation, we scrutinized participants at MTurk. The crowdworking platform is considered a real online labor market in which workers seek profit-maximizing allocation of time and qualification in a relatively natural context limiting bias from unfamiliar testing conditions. We recruited participants from the U.S. and India as well as from 111 additional countries represented on the platform.

We found that rates of behavior and point estimates of treatment effects do not transport beyond specific experimental implementations. Most clearly, data obtained from standard participant pools (students, participants at MTurk) differ significantly from those of the broader population. This undermines the use of empirically-motivated social experiments to elicit descriptive parameters of human behavior. Qualitative results, in contrast, are remarkably robust to changes in samples and settings. Moreover, we find no evidence for participant reactivity and mode effects potentially biasing experimental measurement. These results underscore experiments' capacity to establish generalizable causal effects in theory-driven designs.