Economics
The Open-Access, Open-Assessment E-Journal

Vol. 14, 2020-7 | February 18, 2020 | http://dx.doi.org/10.5018/economics-ejournal.ja.2020-7

Inference in economic experiments

Norbert Hirschauer, Sven Grüner, Oliver Mußhoff, and Claudia Becker

Abstract

Replication crisis and debates about *p*-values have raised doubts about what we can statistically infer from research findings, both in experimental and observational studies. With a view to the ongoing debate on inferential errors, this paper systematizes and discusses experimental designs with regard to the inferences that can and – perhaps more important – that cannot be made from particular designs.

JEL B41 C18 C90

Keywords Economic experiments; ceteris paribus; confounders; control; inference; internal/external validity; randomization; random sampling; superpopulation

Authors

Norbert Hirschauer, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany, norbert.hirschauer@landw.uni-halle.de
Sven Grüner, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
Oliver Mußhoff, Georg August University Göttingen, Göttingen, Germany
Claudia Becker, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

Citation Norbert Hirschauer, Sven Grüner, Oliver Mußhoff, and Claudia Becker (2020). Inference in economic experiments. *Economics: The Open-Access, Open-Assessment E-Journal*, 14 (2020-7): 1–14.
http://dx.doi.org/10.5018/economics-ejournal.ja.2020-7

Received October 15, 2019 **Published as Economics Discussion Paper** November 18, 2019
Revised February 6, 2020 **Accepted** February 10, 2020 **Published** February 18, 2020

1 Introduction

Starting with CHAMBERLIN (1948), SAUERMANN and SELTEN (1959), HOGGATT (1959), SIEGEL and FOURAKER (1960), and SMITH (1962), economists have increasingly adopted *experimental* designs over the last decades. More recently, Banerjee, Duflo, and Kremer received the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019 "for their experimental approach to alleviating global poverty" (cf., BANERJEE et al. 2016). The motivation of economists to adopt experiments was to obtain – compared to *observational* studies – more trustworthy information about the causalities that govern human behavior. Unfortunately, it seems that in the process of adopting the experimental method, no decidedly inference-focused systematization of economic experiments has emerged. Some scholars use randomization as the defining quality and equate "experiments" with "randomized controlled trials" (ATHEY and IMBENS 2017). Despite ensuing changes in the nature of feasible inferences, other researchers include non-randomized designs into the definition as long as behavioral data are generated through a deliberate treatment manipulation (HARRISON and LIST 2004). One might speculate that economists tend to conceptually stretch the term "experiment" because the seemingly attractive label suggests that they have adopted "trustworthy" research methods that are comparable to those in the natural sciences. Whatever the reason, confusion regarding the different types of research designs that are labeled as experiments entails the risk of inferential errors.

The causal inferences that can be made from controlled experiments where "everything else but the item under investigation is held constant" (SAMUELSON and NORDHAUS 1985: 8) are different from those that can be made from observational studies. The former rely on the research design to ex ante ensure *ceteris paribus* conditions that facilitate the identification of causal treatment effects. Observational studies, in contrast, rely on an ex post control of confounders through statistical modeling that, despite attempts to move from correlation to causation, does not provide a way of ascertaining causal relationships that is as reliable as a strong ex ante research design (ATHEY and IMBENS 2017). But even within experimental approaches, different designs facilitate different inferences.

Empirical researchers commonly use the term "validity" to refer to the fundamental inferential question of what we can reasonably claim to have learned from a particular set of data. GERBER et al. (2004) note that validity increases the smaller the "inferential leap" from the idiosyncrasies of a particular set of data to the social group and real-life setting of interest. The most basis approach is to distinguish *internal validity* from *external validity* between which there is an inherent tension as regards the research design (JIMENEZ-BUEDO and MILLER 2010). Internal validity refers to causal inference and can be best ensured in completely controlled artificial environments such as experimental laboratories. External validity, in contrast, is concerned with regularities in complex real-life situations where many influences interact. It can itself be divided into two subtypes: *population validity* is about generalizing from the idiosyncratic study sample towards the population of interest. *Ecological validity* is about generalizing from the idiosyncratic study setting towards the real-life setting of interest. Irrespective of the study type, population validity depends on what is known about the sample-population relationship (e.g., Were subjects selected at random or not?). With regard to ecological validity, one might argue that the inferential leap is smaller in studies based on direct

real-life observations as opposed to experimental studies where data are often extracted from contrived experimental environments. This might be too rash, however. Including survey data into the definition of observational data (as counterpart of experimental data) introduces problems of ecological validity into observational studies. Answers to survey questions run the risk of revealing little information about real-life phenomena – not least because they are simple statements but not revealed preferences. It is therefore not possible to make general propositions regarding the external validity of observational as opposed to experimental studies. Way beyond the generic study type, external validity depends on the specific research design.

In this paper, we focus on the question of statistical and scientific induction and, more particularly, the role of frequentist statistical tools, such as standard errors, confidence intervals or *p*-values, for making inferences beyond the confines of a single *experimental* study. We aim at an adequate differentiation of experimental designs that contributes to a better understanding of the inferences that can and – perhaps more important – that cannot be made from particular designs. While our reasoning generally addresses the applicability of frequentist probabilistic tools, we put a focus on *p*-values because of their high prevalence in empirical research and the recent scientific criticisms regarding *p*-value-based statistical practices (cf. e.g. AMRHEIN et al. 2019; WASSERSTEIN et al. 2019).

2 Experiments aimed at identifying causal treatment effects

The label "experiment" is first of all used for causal studies that, instead of using survey data or pre-existing observational data, are based on a deliberate intervention (treatment) *and* a design-based control over confounders. Identifying the effects of the treatment on the units (subjects) under study requires a comparison.[1] Two basic designs are used to ensure control and thus *ceteris paribus* conditions that facilitate the identification of causal relationships: (1) Randomized controlled trials, which are commonly based on between-subject designs where each subject experiences only one treatment, rely on randomization to generate equivalence between compared groups; i.e. we randomly assign subjects to treatments to ensure that known *and* unknown confounders are balanced across treatment groups in expectation[2] (statistical independence of treatments). (2) *Ceteris paribus* conditions in non-randomized controlled trials, in contrast, are often obtained in within-subject designs where each subject is exposed to both treatments and where everything but the treatment is held constant over time; i.e. instead of using the treatment as a between-subject variable and comparing groups of treated and untreated

[1] For the sake of simplicity, we limit the discussion of treatment comparison to binary treatments. Often, "no-treatment observations" are compared to "with-treatment observations."

[2] DUFLO et al. (2007) note that randomization only achieves that confounders are balanced across treatments "in expectation;" i.e. balance is only ensured if we have a "sufficiently" large experimental group to start with (law of large numbers). In the case of a very small experimental group, an observed difference between two randomized treatment groups could easily be contaminated by unbalanced confounders.

individuals, we now establish causality by using the treatment as a within-subject variable and comparing the before-and-after-treatment outcomes of *all* subjects in the experiment.[3]

The persuasiveness of causal claims depends on the credibility of the alleged control. Comparing randomized treatment groups is generally held to be a more convincing device to identify causal relationships than before-and-after treatment comparisons (CHARNESS et al. 2012). This is due to the fact that randomization balances known *and* unknown confounders across treatment groups and thus ensures statistical independence of treatments. In contrast, efforts to hold everything else but the treatment constant over time are limited by the researcher's capacity to identify and fix confounders. A particular threat to causal inference arises when subjects' properties change through treatment exposure. That is, holding "everything" but the treatment constant over time can be difficult because sequentially exposing subjects to multiple treatments may cause order effects that violate the *ceteris paribus* condition (CHARNESS et al. 2012). However, in such repeated measure designs, the threat of bias through order effects can be mitigated through counterbalancing (ALLEN 2017); and, as CZIBOR et al. (2019) emphasize, within-subject designs also have their advantages: besides the fact that they can more effectively make use of small experimental groups, they can control for variation across subjects (GELMAN 2019) and they facilitate the identification of higher moments of the distribution. Whereas between-subject designs are limited to estimating average treatment effects, within-subject designs enable researchers to look at quantiles and assess heterogeneous treatment effects among subjects. It therefore depends on the research context whether between-subject or within-subject designs are more adequate.

Due to the particular credibility of randomization as a means to establish control over confounders, the use of the term "experiment" – accompanied by the label "natural" – has even been extended to observational settings where, instead of a deliberate treatment manipulation by a researcher, the socio-economic or natural environment has "randomly assigned treatments" among some set of units. Regarding this terminology, DUNNING (2012: 16) notes "that the label 'natural experiment' is perhaps unfortunate. […], the social and political forces that give rise to as-if random assignment of interventions are not generally 'natural' in any ordinary sense of that term. [… and], natural experiments are observational studies, not true experiments, again, because they lack an experimental manipulation. In sum, natural experiments are neither natural nor experiments" but may be structurally close to randomization.[4]

3 Inferences in experiments based on treatment comparisons

Insofar as they succeed in providing an ex ante, design-based control over confounders through the introduction of a well-defined treatment into an otherwise controlled environment,

[3] The term "non-randomized controlled trial" is also used for between-subject designs when the technique of assigning subjects to treatments, e.g. alternate assignment, is not truly random but claimed to be as good as a random.

[4] HARRISON and LIST (2004) speak of "*doing* natural field experiments" to tag field experiments with subjects from the social group of interest and a *covert* manipulation of subjects' real-life environment. For terminological clearness it should be noted that such field experiments are not natural experiments but deliberate interventions.

randomized-treatment-group comparisons (with treatment being a between-subject variable) and before-and-after-treatment comparisons (with treatment being a within-subject variable) facilitate causal inferences. The meaning of statistical inference and the *p*-value are different in the two cases, however. In *randomized-treatment-group comparisons*, the *p*-value linked to the average treatment difference is usually based on the approximation of the *randomization distribution* (cf. RAMSEY and SCHAFER 2013), i.e. the distribution of the *difference between group averages* and the standard error used in a two-independent-sample *t*-test. Regardless of how participating subjects were recruited, the resulting *p*-value targets the following question: when there is no treatment-group difference, how likely is it that we would find a difference as large as (or larger than) the one observed when we repeatedly assigned the experimental subjects at random to the treatments under investigation (VOGT et al. 2014: 242). In randomized controlled experiments, the evaluation of internal validity and causal inference can be aided by statistical inference based on the *p*-value, which represents a continuous measure of the strength of evidence against the null hypothesis of there being no treatment effect in the group of experimental subjects. While scientific inferences beyond the confines of the experimental group under study are often desired, it must be recognized that randomization-based inference is no help for generalizing from experimental subjects to a broader population from which they have been recruited. Using statistical inference to help make such generalizations would require that, besides being randomized, the recruited experimental subjects had been randomly drawn from a defined parent population. If they are not, extending inference from the experimental subjects to any broader group must be based on scientific reasoning beyond statistical measures such as *p*-values. This implies accounting for contextual factors and the entirety of available knowledge including external evidence for the phenomenon under study.[5]

When we not only randomize a given group of experimental subjects but also recruit them from a defined parent population through random sampling, the question arises of how to link randomization-based inference, which is concerned with internal validity and causality, to sampling-based inference, which is concerned with external validity and generalization towards the broader parent population. The "true" standard error of the randomization distribution would reflect the idea of frequently re-randomizing a *given* group of, let's say, $n = 100$ subjects in hypothetical experimental replications. The standard error in a two-independent-*sample t*-test, in contrast, presumes that we repeatedly draw random samples of $n = 100$ subjects from a population before carrying out the randomized experiment. As stated above, two-sample *t*-tests are often used for causal inferences from randomized-treatment-group comparisons even though they are conceptually based on random sampling from populations. If we accept the sampling-based standard error as an approximation of the randomization-based standard error (ATHEY and IMBENS 2017) – it is an upwardly-biased approximation because it considers sampling error in addition to randomization error – the resulting *p*-value can be used as an aid for simultaneously assessing internal and external validity. One should be explicit about the fact,

[5] Even though the *p*-value is the cornerstone of the statistical methodology that is currently in dominant use, it is but a summary statistic of the data at hand from which inductive inferences do not flow automatically. Elaborating on the per se limited inferential content of the *p*-value is not within this paper's scope, however. For a discussion see for example, AMRHEIN et al. (2019), HIRSCHAUER et al. (2018, 2019), MCSHANE et al. (2017), TRAFIMOW et al. (2018), WASSERSTEIN and LAZAR (2016), WASSERSTEIN et al. (2019), or ZILIAK and McCloskey (2008).

however, that the interpretation of the *p*-value must be strictly limited to causal inferences *within* the *given* group of experimental subjects when the group of experimental subjects was not recruited through random sampling.

Contrary to randomization, a *p*-value associated with the treatment difference in *before-and-after-treatment comparisons* is conceptually per se based on random sampling and the *sampling distribution*, i.e. the distribution of the *average individual before-and-after difference* and thus the standard error in a paired *t*-test. This is just another label for a one-sample *t*-test on the variable "individual before-and-after differences." Statistical inference based on the one-sample *p*-value implies that we concern ourselves with the question of what we can learn about the population mean from a random sample. In other words, we are asking the following question: assuming there is no difference in the population, how likely is it that we would find an average before-and-after difference as large as (or larger than) the one observed if we carried out very many statistical replications and subjected repeatedly drawn random samples to the same treatment procedure. Therefore, our *p*-value is a continuous measure of the strength of evidence against the null of there being no treatment effect in the parent population. While being an inferential tool to help make generalizations from the sample of experimental subjects to a broader population (external validity), it must be recognized that a *p*-value in before-and-after comparisons is no help whatsoever for assessing causality. Instead, causality claims hinge on the credibility of the *ceteris paribus* claim and must be based on transparent experimental protocols that show what exactly researchers did to hold everything but the treatment constant over time. A *p*-value in a one-sample *t*-test informs us about the random sampling error, *irrespective* of whether our experimental procedure was successful in holding everything but the treatment constant over time or not. The only important assumption is that the treatment that leads to the observation of individual before-and-after differences presumably remains *unchanged* over all statistical replications. One should be clear that there is no role for a *p*-value when subjects in before-and-after-treatment comparisons are not randomly recruited.

Being a probabilistic concept based on a chance model (i.e. a hypothetical replication of a chance mechanism), frequentist tools such as *p*-values are not applicable if there is no random process of data generation (either randomization or random sampling). When there is no randomization, maintaining the *p*-value's probabilistic foundation therefore poses serious conceptual challenges when we already have the data of the whole target population (DENTON 1988: 166f.). An example is a non-randomized experimental within-subject design where experimental subjects are clearly a non-random convenience sample, or where we do not want to generalize beyond the confines of the particular sample to start with. In such cases, the sample already constitutes the finite population to which we are limited. Due to the lack of a chance mechanism that could hypothetically be replicated, there is no role for the frequentist *p*-value and statistical significance testing. The fact that there is no room for statistical inference when we already have data of the entire inferential target population is formally reflected in the finite population correction factor. Rather than assuming that a sample was drawn from an infinite population – or at least that a small sample of size n was drawn from a very large population of size N – the finite population correction factor $(1-n/N)^{0.5}$ accounts for the fact that, besides absolute sample size, sampling error decreases when the sample size becomes large relative to the whole population. The correction reduces the standard error and is commonly

used when sample share is more than 5% of the population (KNAUB 2008). Having the entire population corresponds to a correction factor of zero and thus a corrected standard error of zero. If frequentist inferential tools such as *p*-values are nonetheless calculated for entire populations or non-random samples for generalization purposes (cf. ABADIE et al. 2014 for a justification to do so in causal analysis of observational data), one would have to imagine an infinite "unseen parent population" (or "superpopulation"), i.e. an underlying stochastic mechanism that is hypothesized to have generated the observations. DENTON (1988) critically notes that this rhetorical device, which is also known as "great urn of nature," does not evoke wild enthusiasm from everybody. "However, some notion of an underlying [random] process – as distinct from merely a record of empirical observations – has to be accepted for the testing of hypotheses in econometrics to make any sense" (DENTON 1988: 167). We would add that researchers who resort to the *p*-value in such circumstances should transparently explain their inferential reasoning and emphasize that no generalization towards a numerically larger population is involved. Instead, inferences would be limited to the unseen superpopulation in terms of a random process that is supposed to "apply" to the subjects who happen to be in the sample.

4 Inferences in experiments without treatment comparisons

In experimental treatment comparisons, the term "control" means first of all generating *ceteris paribus* conditions (ex-ante control over confounders) with the objective of identifying causal treatment effects. We know that this ex-ante control comes in two forms: in randomized-treatment-group comparisons, control over confounders is achieved without exercising control over the environment; i.e. randomization, which balances confounders (including unknown ones) across treatment groups, replaces environmental control. In non-randomized before-and-after-treatment comparisons, in contrast, control over confounders requires that we exercise control over the environment and fix and maintain all factors that could influence subjects' behaviors besides the treatment under investigation.

Often, economic experiments do not settle for identifying causal treatment effects among experimental subjects in more or less artificial experimental environments. Instead, experimenters want to learn what governs the behaviors of certain social groups in relevant real-world contexts and, eventually, how policy interventions would work in these contexts. This requires not only going beyond internal validity and causality. It also requires moving external validity beyond statistical inference, which is solely concerned with random error in repeated random sampling from the same population and thus the *sample-population relationship* (population validity). That is, we cannot limit ourselves to the question of how we can generalize from the behavior of experimental subjects in a particular but potentially uninformative experiment to the would-be behavior of the parent population in this very experiment. Instead, we need to address the *experiment-real-world relationship* (ecological validity). Using a well-known expression coined by SMITH (1982), we need to exercise "control over subjects' preferences" and search for experimental designs which ensure that subjects' choices in the experiment reveal what we want to know, i.e. their "true" preferences. In the terminology of measurement theory, we would say that, besides the *uncertainty* of the

measurement due to sampling error (measurement precision), we are now concerned with the *accuracy* of the measurement (measurement validity) and the question of whether the measurement instrument "experiment" yields a manifest variable (observed experimental behavior) that is informative regarding the latent variable of interest, i.e. people's true preferences. It should be noted that an experiment's measurement accuracy cannot be evaluated by statistical tools. It can only be evaluated based on the logical consistency of the argument that is put forward in justification of the particular experimental design and/or in relation to a presumed standard of knowledge.

Ecological validity is a crucial part of external validity irrespective of whether economic experiments are based on treatment comparisons or not. However, it is often more salient in economic experiments that study only one treatment and do not aim for causal inferences through *ceteris paribus* comparisons. While still relying on an experimenter's intervention, such experiments are focused on measuring latent preferences such as individual risk or social preferences. Prominent examples are experimental games such as prisoner's dilemmas, trust games, or public goods games that are implemented to find out, for instance, whether the choices made by individuals are in line with conventional rational choice predictions.[6] For example, one might deliberate how large the real payments (incentives) that are linked to subjects' abstract earnings in a dictator "game" would have to be to achieve a valid measurement in that these incentives make subjects reveal their true prosocial preferences. Another example is the attempt to avoid "experimenter demand effects" that often threaten external validity because subjects are usually aware of participating in an experiment and often inclined to please experimenters (DE QUIDT et al. 2018). When assessing the quality of the experimental control over subjects' preferences, one should be clear that this aspect of external validity has nothing to do with statistics such as *p*-values. In other words, we may jointly have randomization *and* random sampling *and* control over subjects' preferences in an experiment. However, we may also have an experiment without randomized treatment assignment and without random recruitment of subjects, but with a credible control over subjects' preferences. Imagine a well-incentivized dictator game carried out with a non-random convenience sample of students who happen to be in an experimenter's classroom on a particular Friday. In this case, all inductive inferences – be they towards the experimental behavior of a broader population of students or other demographic groups, or towards the real-life behavior of the classroom students or broader populations – must be based on scientific arguments beyond *p*-values. It would therefore be a gross abuse to use the term "statistical significance" for a purported corroboration of such inferences.

Control over the environment, in terms of shaping, knowing, and describing all behaviorally relevant factors besides the treatment of interest, generally decreases from lab experiments to field experiments, irrespective of whether they are based on treatment comparisons or not. Any taxonomic proposal that takes account of the diminishing control over the environment from the lab to the field is open to debate – at least for non-randomized experiments. Attaching the label "experiment" to studies that rely on proper randomization to control for confounders is likely to cause little controversy even when they are carried out in the field where it is difficult to know,

[6] Of course, all these games could also be used within a randomized design. Simply imagine an experiment in which subjects are randomly assigned to two dictator experiments with differing initial endowments.

let alone fix all relevant factors besides the treatment. In non-randomized designs, in contrast, the classification is likely to become controversial at some point; i.e. an arguable minimum level of control over the relevant environment would seem to be a prerequisite for calling a non-randomized approach an experiment. Irrespective of the label, we must take account of the specific research design when making inferences: (1) Inference regarding *causal relationships* must be based on scientific arguments but cannot be supported by a *p*-value when an experiment is not based on randomization. An important example are experimental within-subject designs. When causal inferences are based on doubtful claims of control over confounders, one should consider alternative experimental designs (e.g. randomized instead of non-randomized designs) or even a regression-based statistical control of observable confounders.[7] (2) Inference dealing with the *sample-population relationship* must be based on scientific reasoning but cannot be supported by a *p*-value when there was no random sampling from a broader (numerically larger) population. This is the case, for example, when randomized experiments are carried out with subjects from non-random convenience sample. (3) Inference dealing with the *experiment-real-world relationship* and thus the question of whether experimental subjects reveal their "true" preferences in a particular experiment cannot be supported by a *p*-value at all. When the control over subjects' preferences is in question, one should avoid overhasty conclusions and check the robustness of results in replication studies with more valid experimental designs – preferably in field experiments carried out with subjects from the relevant parent population and a manipulation of subjects' real-life environments.

5 Inferences in quasi-experiments

Often, non-randomized study designs focus on the behavioral outcomes induced by an intervention in one social group as opposed to another. Such designs are examples of "quasi-experiments" (CAMPBELL and STANLEY 1966) in which the *ceteris paribus* condition is in question. For illustration, imagine a dictator "game" in which a mixed-sex group of experimental subjects are used as first players who can decide which share of their initial

[7] There is no need to resort to regression when proper randomization ensures ex ante that confounders are statistically independent of treatments. In some cases, for instance when only a small experimental group is available (cf. Footnote 2), switching to an ex-post control of confounders in a statistical model may be appropriate, however. It may therefore be useful to realize how, in the simplest case without confounders, a treatment-group comparison relates to a linear model where we regress the response to a binary treatment dummy and a constant. Generally speaking, the sampling distributions of estimated regression coefficients $\hat{\beta}_j$ that link predictors x_j to response y are the distributions of the point estimates derived from a hypothetically repeated random sampling of the response variable at the fixed values of the predictors (RAMSEY and SCHAFER 2013: 184). Using a dummy regression (and a *p*-value based on the *sampling distribution*) instead of comparing two group averages (and a *p*-value based on the *randomization distribution*) can therefore be questioned on the grounds that it implies switching to a chance model that is at odds with the actually applied chance mechanism. There are specific constellations (equal variance in both groups or, alternatively, heteroscedasticity-robust regression standard errors) that lead to identical standard errors. However, group comparison and dummy regression only coincide as long as the former is based on the sampling-based approximation of the standard error of the randomization distribution. If the group comparison were based on the correct standard error of the randomization distribution, we would obtain a lower standard error compared to which the standard error in the regression would be upwardly biased (ATHEY and IMBENS 2017).

endowment they give to a second player (one person acts as second player for the whole group). Additionally, assume that the experimental subjects are a convenience sample but not a random sample of a well-defined broader population. What kind of statistical inferences are possible? Neither one of the two chance mechanisms – neither random sampling nor randomization – applies. Consequently, there is no role for the *p*-value: (i) Statistical inference towards a wider population beyond our experimental subjects is not possible because we are limited to a non-random sample.[8] (ii) Statistical inference regarding causal relationships is not possible because there was no random assignment of subjects to different treatments. Instead, *one* treatment was used to obtain a behavioral measurement in *two* predefined social groups. We should therefore simply describe, without reference to a *p*-value, the observed difference and the experimental conditions – or carry out a regression analysis to control for confounders if necessary; for example, the male subjects may be more or less wealthy than the female subjects which could be another explanation for the differences between the two groups.

Due to engrained disciplinary habits, researchers might be tempted to implement "statistical significance testing" routines in our dictator game example even though there is no chance model upon which to base statistical inference. While there is no random process, implementing a two-sample *t*-test might be the spontaneous reflex to find out whether there is a "statistically significant" difference between the two sexes. One should recognize, however, that doing so would require that some notion of a random mechanism is accepted. In our case, this would require imagining a *randomization distribution* that would result if money amounts were repeatedly assigned to sexes ("treatments") at random. Our question would be whether the money amounts transferred to the second player differed more between the sexes than what would be expected in the case of such a random assignment. We must realize, however, that there was no random assignment of subjects to treatments, i.e. the sexes might not be independent of covariates. Therefore, the *p*-value based on a two-sample *t*-test for a difference in mean does not address the question of whether the difference in the average transferred money is *caused* by the subjects' being male or female. That could be the case, but the difference could also be due to other reasons such as female subjects being less or more wealthy than male subjects. As stated above, it would therefore make sense to control for known confounders in a regression analysis ex post – again, without reference to a *p*-value as long as the experimental subjects have not been recruited through random sampling.

Another justification for using *p*-values in research based on convenience samples is to assume that the "great urn of nature" provides a process of data generation that equals random sampling even though the researcher has not complied with the "empirical commitment" of random sampling (BERK and FREEDMAN 2003) when collecting the data. We know that convenience samples might be contaminated with selection bias that precludes a conventional use of inferential statistics unless corrected for in a sample selection model. Applying

[8] Convenience sample are potentially contaminated with selection bias. Due to the violation of independent random sampling they may be systematically different from the broader target population. Unless we have enough information from the population to apply sample selection models (e.g. propensity score models) that correct for such bias, we cannot generalize from the measured sample to the broader population. That is, we will misestimate population quantities and standard errors in unknown ways. The correction of selection bias in non-random samples is a highly complex methodological issue whose representation – even in a very basic manner – is beyond this paper's scope. For an introduction, see, for example, ROSENTHAL and ROSNOW (2009: book 3).

frequentist statistics to (uncorrected) convenience sample and assuming that there is a random process of data generation therefore requires that the observed sample becomes in imagination a sample of a (parent) superpopulation. While such a conceptual view introduces an imaginary *sampling distribution*, it does not facilitate statistical inference in the conventional sense of generalizing from the measured sample towards a well-defined and numerically larger population. Instead, inferences would be limited to the unseen superpopulation in terms of a random process from which one has presumably observed one realization, and which is assumed to be valid for exactly the sample under study. In many research contexts, it seems doubtful whether such a perspective is useful for answering the fundamental inferential question of what we can claim to have learned from a particular set of data. Irrespective of an individual researcher's opinion, transparency is an indispensable requirement for successful scientific communication and progress. That is, researchers who use frequentist inferential statistics in the analysis of non-random samples should explicate and justify why and how they base their inferential reasoning on the notion of a superpopulation. This is but a specification of the general desideratum that researchers explicitly describe the sampling process as well as the population of interest from which their sample comes and to which they want to generalize.

6 Conclusion

Systematizations of economic experiments have not predominantly addressed the inferences that can be made in different types of experimental designs. Usages of the term "experiment" range from a narrow view of "applying randomization" to identify causal effects, to a broad definition of "deliberate treatment manipulation." Our paper has shown that an adequate differentiation of experimental designs advances the understanding of what we can infer from different types of experimental studies. Several points should be kept in mind: *first*, a random process of data generation – either random assignment or random sampling – is required for frequentist tools such as *p*-values to make sense, however little it may be. *Second*, the informational content of frequentist tools such as *p*-values is different in randomization-based inference as opposed to sampling-based inference. Randomization-based inference is concerned with internal validity and causality, whereas sampling-based inference is concerned with external validity in terms of generalizing from a sample to its parent population. *Third*, while being conceptually different, the sampling-based standard error used in a two-sample *t*-test is often used as an approximation in randomization-based inference. If one accepts the approximation, and if experimental subjects are recruited through random sampling, the resulting *p*-value can be used as an aid both for assessing internal validity and for generalizing to the parent population. However, if experimental subjects are not randomly recruited, statistical inferences must be limited to assess the causalities *within* the studied sample. *Fourth*, in the context of economic experiments, there are two different meanings of the term "control." In experiments aimed at identifying causal treatment effects, control means first of all ensuring *ceteris paribus* conditions (statistical independence of treatments). Besides internal validity, the term "control" is also associated with ecological validity. The expression "control over preferences" is used to indicate designs in which a valid measurement is achieved in that experimental subjects can be believed to reveal

their true preferences. This design quality, which is crucial for making externally valid inferences, is part of scientific reasoning but cannot be aided by *p*-values.

Acknowledgments The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding.

References

Abadie, A., Aghey, S., Imbens, G., and Wooldridge, J.M. (2014): Finite Population Causal Standard Errrors. NBER Working Paper No. 20325. https://www.nber.org/papers/w20325

Allen, M. (2017): Counterbalancing. The SAGE Encyclopedia of Communication Research Methods. https://dx.doi.org/10.4135/9781483381411.n103.

Athey, S., and Imbens, G.W. (2017): The Econometrics of Randomized Experiments. In: Banerjee, A.V., and Duflo, E. (eds.): *Handbook of Field Experiments*. Volume 1. Amsterdam: Elsevier.

Amrhein, V., Greenland, S., and McShane, B. (2019): Retire Statistical Significance. *Nature* 567: 305–307. http://www.igienistionline.it/docs/2019/10nature.pdf

Banerjee, A., Duflo, E., and Kremer, M. (2016): The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. Paper prepared for "The State of Economics, The State of the World". Conference proceedings volume. https://scholar.harvard.edu/files/kremer/files/the-influence-of-rcts-on-developmental-economics-research-and-development-policy.pdf

Berk, R.A., and Freedman, D.A. (2003): Statistical Assumptions as Empirical Commitments. In: Blomberg, T.G., and Cohen, S. (eds.): *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger* (2nd ed.). New York: de Gruyter.

Campbell, D.T., and Stanley, J.C. (1966): *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Chamberlin, E.H. (1948): An Experimental Imperfect Market. *Journal of Political Economy* 56(2): 95–108. https://www.journals.uchicago.edu/doi/10.1086/256654?mobileUi=0

Charness, G., Gneezy, U., and Kuhn, M.A. (2012): Experimental Methods: Between-Subject and Within-Subject Design. *Journal of Economic Behavior & Organization* 81(1): 1–8. https://www.sciencedirect.com/science/article/abs/pii/S0167268111002289

Czibor, E., Jimenez-Gomez, D., and List, J.A. (2019): The Dozen Things Experimental Economists Should Do (More of). *Southern Economic Journal* 86(2): 371–432. https://doi.org/10.1002/soej.12392

Denton, F.T. (1988): The Significance of Significance: Rhetorical Aspects of Statistical Hypothesis Testing in Economics. In: Klamer, A., McCloskey, D.N., and Solow, R.M. (eds.): *The Consequences of Economic Rhetoric*. Cambridge: Cambridge University Press.

de Quidt, J., Haushofer, J., and Roth, C. (2018): Measuring and Bounding Experimenter Demand. *American Economic Review* 108(11): 3266–3302. https://www.aeaweb.org/articles?id=10.1257/aer.20171330&&from=f

Duflo, E., Glennerster, R., and Kremer, M. (2007): Using Randomization in Development Economics Research: A Toolkit. In: Schultz, T., and Strauss, J. (eds.): *Handbook of Development Economics*, Volume 4. Amsterdam: Elsevier.

Dunning, T. (2012): *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge: Cambridge University Press.

Gelman, A. (2019): Why Do a Within-Person rather than a Between-Person Experiment? https://statmodeling.stat.columbia.edu/2019/11/16/why-do-a-within-person-rather-than-a-between-person-experiment/

Gerber, A.S., Green, D.P., and Kaplan, E.H. (2004): The Illusion of Learning from Observational Research. In: Shapiro, I., Smith, R., and Massoud, T. (eds.): *Problems and Methods in the Study of Politics.* New York: Cambridge University Press.

Harrison, G.W., and List, J.A. (2004): Field Experiments. *Journal of Economic Literature* 42(4): 1009–1055. https://www.aeaweb.org/articles?id=10.1257/0022051043004577

Hirschauer, N., Grüner, S., Mußhoff, O., and Becker, C. (2018): Pitfalls of Significance Testing and *p*-Value Variability: An Econometrics Perspective. *Statistics Surveys* 12: 136–172. https://projecteuclid.org/download/pdfview_1/euclid.ssu/1538618436

Hirschauer, N., Grüner, S., Mußhoff, O., and Becker, C. (2019): Twenty Steps towards an Adequate Inferential Interpretation of *p*-Values in Econometrics. *Journal of Economics and Statistics* 239(4): 703–721. https://ideas.repec.org/a/jns/jbstat/v239y2019i4p703-721n8.html

Hoggatt, A.C. (1959): An Experimental Business Game. *Behavioral Science* 4(3): 192–203. https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830040303

Jimenez-Buedo, M., and Miller, L.M. (2010): Why a Trade-off? The Relationship between the External and Internal Validity of Experiments. *THEORIA. An International Journal for Theory, History and Foundations of Science* 25(3): 301–321. https://www.ehu.eus/ojs/index.php/THEORIA/article/view/779

Knaub, J. (2008): Finite Population Correction (fcp) Factor. In: Lavrakas, P. (ed.): *Encyclopedia of Survey Research Methods*. Thousand Oaks: Sage Publications.

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2019): Abandon Statistical Significance. *The American Statistician* 73(s1): 235–245. https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1527253

Ramsey, F.L., Schafer, D.W. (2013): *The Statistical Sleuth: A Course in the Methods of Data Analysis*. Belmont: Brooks/Cole.

Rosenthal, R., and Rosnow, R.L. (2009): *Artifacts in Behavioral Research*. Oxford: Oxford University Press.

Samuelson, P.A., and Nordhaus, W.D. (1985): *Economics. 12th ed*. New York: McGraw-Hill.

Sauermann, H., and Selten, R. (1959): Ein Oligopolexperiment. *Zeitschrift für die gesamte Staatswissenschaft* 115(3): 427–471. https://www.jstor.org/stable/40748152

Siegel, S., and Fouraker, L.E. (1960): *Bargaining and Group Decision Making*. New York: McGraw-Hill.

Smith, V.L. (1962): An Experimental Study of Market Behavior. *Journal of Political Economy* 70(2): 111–137. http://econ.ucsb.edu/~oprea/176/Competitive.pdf

Smith, V.L. (1982): Microeconomic Systems as an Experimental Science. *The American Economic Review* 72(5): 923–955. https://www.jstor.org/stable/1812014

Trafimow et al. (2018): Manipulating the Alpha Level Cannot Cure Significance Testing. *Frontiers in Psychology*. 15 May 2018. https://doi.org/10.3389/fpsyg.2018.00699

Vogt, W.P., Vogt, E.R., Gardner, D.C., and Haeffele, L.M. (2014): *Selecting the Right Analyses for your Data: Quantitative, Qualitative, and Mixed Methods*. New York: The Guilford Publishing.

Wasserstein, R.L., and Lazar N.A. (2016): The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70(2): 129–133. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Wasserstein, R.L., Schirm, A.L., and Lazar, N.A. (2019): Moving to a World Beyond "p < 0.05". *The American Statistician* 73(s1): 1–19. https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913

Ziliak, S.T., and McCloskey, D.N. (2008): *The Cult of Statistical Significance. How the Standard Error Costs us Jobs, Justice, and Lives*. Ann Arbor: The University of Michigan Press.