Experimental practices and objectivity in the social sciences: re-embedding construct validity in the internal–external validity distinction

Jiménez-Buedo, M.; Russo, F.

DOI
10.1007/s11229-021-03215-3

Publication date
2021

Document Version
Final published version

Published in
Synthese

Synthese (2021) 199:9549–9579
https://doi.org/10.1007/s11229-021-03215-3

ORIGINAL RESEARCH

Experimental practices and objectivity in the social sciences: re-embedding construct validity in the internal–external validity distinction

María Jiménez-Buedo[1] · **Federica Russo**[2]

Received: 17 January 2020 / Accepted: 12 May 2021 / Published online: 18 June 2021

Abstract

The experimental revolution in the social sciences is one of the most significant methodological shifts undergone by the field since the 'quantitative revolution' in the nineteenth century. One of the often valued features of social science experimentation is precisely the fact that there are (alleged) clear methodological rules regarding hypothesis testing that come from the methods of the natural sciences and from the methodology of RCTs in the biomedical sciences, and that allow for the adjudication among contentious causal claims. We examine critically this claim and argue that some current understandings of the practices that surround social science experimentation overestimate the degree to which experiments can actually fulfil this role as "objective" adjudicators, by neglecting the importance of shared background knowledge or assumptions and of consensus regarding the validity of the constructs involved in an experiment. We take issue with the way the distinction between internal and external validity is often used to comment on the inferential import of experiments, used both among practitioners and among philosophers of science. We describe the ways in which the more common (dichotomous) use of the internal/external distinction differs from Cook and Campbell's original methodological project, in which construct validity and the four-fold validity typology were all important in assessing the inferential import of experiments. We argue that the current uses of the labels internal and external, as applied to experimental validity, help to encroach a simplistic view on the inferential import of experiments that, in turn, misrepresents their capacity to provide objective knowledge about the causal relations between variables.

This article belongs to the topical collection "Objectivity in Social Research", edited by Julie Zahle and Petri Ylikoski.

✉ María Jiménez-Buedo
mjbuedo@fsof.uned.es

Extended author information available on the last page of the article

Keywords External validity · Internal validity · Objectivity · Construct validity · Experiments in the social science · Experiments · Background knowledge

1 Experiments and the revolution in causal identification in the social sciences

The (causal) identification revolution in the Social Sciences began at the turn of the twenty-first century as a way to alleviate the increasingly apparent limitations of regression analyses in regard to asserting causal relations in the face of the "often non-random distribution of observations, problems of endogeneity and omitted variable bias, and the complex causal relationships that abound in the social, political and economic world." (Morgan, 2016). These problems became central to empirical researchers by explicitly addressing them with new statistical methods (many based on experimental reasoning such as instrumental variables), and also, and by the (re)introduction of experimentation as a viable research strategy. Although the existing literature has nuanced and toned down the potential of experiments by showing their limitations (e.g. Guala (2005)), the idea that experiments are the gold standard is quite pervasive in social science methodology—the enthusiasm for the possible use of the potential outcome model in social science contexts is a case in point here.

The identification revolution and, in particular, its experimental dimension had, in turn, a policy derivative in the form of the Evidence Based Policy (EBP) movement, a series of institutional initiatives set out to improve the evidential standards of policy makers that in practice has translated most notably to the promotion of randomized controlled trials (RCTs) as the best means to both design and evaluate policy innovation.

Underlying some of the newly found enthusiasm in social scientific experimentation is often an idea (stated or unstated) that experiments are the best way to adjudicate contentious causal claims; this is mirrored in evidence hierarchies of various forms and in debates in (social science) methodology that invariably consider experiments as the most reliable methods for causal inference (Parkhurst & Abeysinghe, 2016; Petticrew & Roberts, 2006; Tellings, 2017). This is so, the argument goes, because unlike other methods, experiments are objective, or ensure the objectivity of the inferences made. The experimental impulse in this light would thus be part of EBPs broader goal "to replace subjective (biased, error-prone, idiosyncratic) judgments by mechanically objective methods" (Reiss & Sprenger, 2017).

As the EBP movement has become mainstream and its associated RCT hype has increased, more commentators have tried to caution against some of its bolder ambitions. Deaton and Cartwright (2018), for example, have discussed the many ways in which the value of RCTs' evidence is often overstated by defending the idea that RCTs are a useful method among many, but that they should not be thought of as the research 'gold standard'. First, because RCTs are, too, susceptible of bias. For example, in cases of differential experimental attrition among the experimental groups, or, in cases in which treatment blinding fails and impacts results, RCTs

will provide biased estimates of the causal impact of treatment on the output variables. Second, Deaton and Cartwright (2018) also call for caution by emphasizing the fact that RCT results cannot often translate into immediate "applicability" for policy purposes: causal processes often require highly specialized economic, cultural or social structures to enable them to work, and we often lack a theory about which 'supporting factors' are important (i.e., additional factors that function along with the treatment to create the observed outcome). According to these authors, and using their terminology, RCTs cannot tell us if what works 'there' (in the trial), will also work for us 'here' (in other situations of interest).

While endorsing these cautionary warnings regarding the capacities of RCTs, our emphasis here is different. First, we argue that even in the absence of biases, there are important limitations to the possibility that experiments can provide an objective adjudication among contentious causal claims; yet these limitations, we argue, are often obscured in the terminology we use to describe the inferential import of experiment. In particular, we argue that the way in which the distinction between *internal and external validity* (a key terminological tenet of the 'new experimentalism' in social sciences practices) is often used, helps to crystallize the assumption that for every experiment there is a clear-cut, univocal correspondence to an inference (or a given set thereof). In what follows we call this assumption regarding the collapse of an experiment with its ensuing inferences, the "one experiment, one inference" view of experimentation (1e:1i, for short). Second, the arguments put forward by Cartwright and collaborators focus on the limitations of RCTs for the extrapolation of successful interventions, while we also pay attention to the fact that even if extrapolation is not an immediate concern, different extract different lessons from the same experiment, depending on their different background knowledge or assumptions.

As we will argue, the concepts of internal and external validity started off as part of a typology of common threats to experimental design (a list of frequent potential confounds to field experimentation), mostly meant as a practical guide for experimentalists. Yet, internal and external validity gradually became something *else*: a *dichotomy* that ended up being used, both in philosopher and experimentalists' circles, as a way to map the inferential import of experiment (i.e. 'where' the experiment holds), rather than a *guide* through the process of model-validation. The problem is that this particular interpretation, in which internal and external validity are used as if they jointly could describe exhaustively the inferential import of an experiment, is riddled with assumptions that can obscure, rather than clarify, the relation between experiments and the inferences they allow. In particular, we argue that a dichotomous interpretation of the distinction between internal and external validity, by *ignoring construct validity*, implicitly assumes an easy correspondence between an experiment and the inferences that it licences (what we have called 1e:1i).

We contend that by using the concepts of internal and external validity as a dyad, we risk underplaying the crucial role of background knowledge or assumptions in the inferential process of experiments,[1] thus obscuring the fact that different scientists, depending on their different theoretical and empirical knowledge, may very well make different causal inferences from the same experimental intervention. This does not have to do so much with the important Cartwrightian qualm, i.e., insisting that we need to identify the right background conditions of a given intervention before we can extrapolate its results. We instead want to emphasise that the same intervention can legitimately give rise to several competing different inferences, and that these will depend on the type of constructs that we think are involved in the intervention, which are in turn dependent on our background assumptions. We think that ignoring this aspect of experimentation, centrally related to the notion of construct validity, helps to embed the conclusion that experiments are crucial or even sufficient to objectively mediate long-standing debates regarding causal claims. We also think that some uses of the internal/external validity distinction contribute to this overestimation of the extent to which experiments provide an objective way to adjudicate among contentious causal claims.

The paper is organised as follows. In Sect. 2 we offer a brief historical overview of the internal–external validity distinction and describe how the distinction entered the philosophical debates and jargon. Section 3 explains how some of the common uses of the internal–external validity distinction embed the 1e:1i assumption; we first make a general case against this uses and then we discuss in detail what construct validity is and in what sense it can help to establish the more limited sense in which objectivity is linked to experiments. Sections 4 and 5 build on the previous, general argument and focus on specific experimental settings, showing the role of construct validity in the process of experimental design; we also show that some common interpretations of the internal and external validity distinction are not adequate to account for experimental practices more generally, as they obscure the role of background assumptions in experimentation. We conclude with a general reflection on how abandoning this assumption critically affects the status of RCTs (or experiments in general) as objective adjudicators between contentious causal claims, and open a path of research in which background knowledge receives more systematic attention to establish the objectivity of experimental practices in social research.

[1] In what follows we use the notion of "background knowledge" as equivalent the whole body of assumptions, studies, obtained results (positive / negative) that scientists may use in their reasoning or inferences when designing and analysing an experiment, rather than as synonym for true or established, knowledge. Because it "background knowledge is nevertheless a loaded notion, we use it interchangeably with the notion of "background assumptions" throughout.

2 Internal and external validity: from methodology to philosophy of science

2.1 The Campbellian methodological project

The Campbellian methodological project is about *design analysis*. Its aim is, precisely, to systematize the way in which we convert a research question into a particular, concrete, design that we can test on the ground. Within this framework, the distinction between internal and external validity was conceived by Campbell (1957) as part of an admirable and sustained methodological effort directed at understanding the pitfalls of research analysis and causal inference in the social sciences. The project that developed gradually through the second half of the twentieth century was of both a theoretical nature and an applied one. Theoretically, it consisted of the systematic study of causal inference in relation to different research designs and it included discussions of the limits and advantages of randomized designs versus their alternatives. In its pragmatic dimension, it served to pave the way to what we now know as the Evidence Based Policy movement, as it evaluated, consulted and inspired myriad field try-outs of ameliorative programs (many of which were related to remedial education), and eventually gave rise to the Campbell Collaboration, which was born as a sister organization in the Social Sciences to the Cochrane project in Medicine (Davies & Boruch, 2001).

Within the massive methodological legacy left by the Campbellian project, a particularly important bit was the conceptual work on validity, which had its roots in Campbell's previous solo work on psychological test validation (Heukelom, 2011). Campbell's first distinction between internal and external validity came in the form of two questions that a researcher ought to ask herself in assessing the success of an experiment: Internal validity would be a response to the question "did in fact the experimental stimulus make some significant difference in this specific instance?" (Campbell, 1957, p. 297), and external validity was defined as "to what populations, settings, and variables can this effect be generalized?" (p. 297). After a posterior reformulation by Campbell and Stanley (1963), it is Cook and Campbell's definitions, embedded within a typology that also included statistical validity and construct validity, and that became the standard in the methodology of applied research settings.

In Cook and Campbell's work (1979), for years the research design manual of reference in many social scientific disciplines, *internal validity* "refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause", whereas *external validity* "refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times". Less known or considered are the other two types of validity. *Statistical conclusion validity* is defined as the "validity of inferences about the correlation (covariation between treatment and outcome", and *construct validity* is defined as "the validity of inferences about the higher order constructs that represent sampling particulars" (p.

38). We shall return to the concept of 'construct' later in Sect. 3, for now it suffices to mention that, in social science methodology, 'construct' has a specific meaning: borrowed from psychology, 'construct' roughly refers to a theoretical concept that represents or corresponds to an observed pattern (of behavior, normally). In other words, a construct refers to the theoretical concepts that aim at capturing a certain (recurring) phenomenon.

Cook and Campbell's validity typology, and in particular the internal–external types, did spur some controversy among contemporary methodologists, who thought of it as ambiguous (see, for example, (Hammersley, 1993)), and it competed against alternative validity typologies, notably, by Cronbach's (1982). Cook and Campbell finally integrated Cronbach's idea that experimental results needed to be tested systematically for robustness across units, treatments, observations and settings (Cronbach, 1982; Shadish et al., 2002). Despite the controversy, though, their fourfold typology got through to the practitioners and their classification and analysis of common "threats to validity", (or the list of common confounding factors that cannot always be entirely offset by randomization) are still regularly used by social scientists.

2.2 Experiments, inferences and validity

In 2002, and coinciding with a newly found interest in experimentation in other social sciences, notably, economics, Campbell signed his last collaboration (actually, posthumously) in the again, impressive volume by Shadishet al. (2002). There, internal validity is defined as "refer[ring] to inferences about whether the observed covariation between X (the presumed treatment) and Y (the presumed outcome) reflects a causal relation from X to Y. (p.38), and external validity is defined as the "degree to which a causal relationship found in a given study generalizes across various persons, settings, treatments, measures, and so forth".

The new definitions, while very close to the original ones, actually make a more direct reference to the inferences that would (not) be allowed from a given experiment. The difference is subtle, but perhaps reflects one of the ambiguities embedded in the definitions from the very start: the validity categories were and are often (improperly) used to refer to the experiments, rather than to the inferences from those experiments.

Campbell and co-authors were not without responsibility in this regard, as they have often spoken indistinctly about the validity of experiments and the validity of inferences. The issue is discussed in detail in Jiménez-Buedo (2011), and it had been previously noticed by authors such as Mark (1986) and Hammersley (1991, 1993). Shadish et al. (2002) reflect on this point, and agree that validity, if properly used should be reserved to inferences. They say:

> Validity is a property of inferences. It is not a property of designs or methods, *for the same design may contribute to more or less valid inferences under different circumstances*. For example, using a randomized experiment does not

> guarantee that one will make a valid inference about the existence of a descriptive causal relationship. (p. 34, emphasis added)

However, and though they admit that "it is wrong to say that a randomized experiment is internally valid or has internal 'validity'", they give themselves and others the licence to 'occasionally speak that way for convenience', and so they do refer to the internal or external validity of a given experiment or experimental design or, for example state that 'the decision to use a randomized experiment (…) often helps internal validity but hampers external validity' (p. 34).

Our contention here is that this terminological relaxation, which has rather become the norm, is not as benign as the authors seem to suggest, for to make sense of it we must suppose that there is a rather *direct correspondence between an experiment and the inferences it licences*, which often translate in misleading or confusing claims, such as, to name just one example, the widely accepted claim that RCTs have high internal validity and low in external validity, rather than discussing the types of inferences that they allow (or not), which will in turn depend on many other factors (such as the thickness of our background knowledge or how far-stretched those inferences need to be for our practical purposes).

We want to put here the emphasis on a fact that usually goes unnoticed: in terms of the Campbellian validity typology, speaking *indistinctively* about the validity of an experiment and about the validity of the inferences from an experiment does imply that there are no issues regarding *construct validity*, i.e., "the validity of inferences about the higher order constructs that represent sampling particulars". In other words, to speak indistinctively about experiment or inference presupposes that there are no concerns regarding the match between the experiment's operations and the constructs used to describe or to design those operations. Again, in the jargon of social science research, constructs are not phenomena, but part of the theoretical framework to analyse the social world, and they are used to make choices at the measurement and data collection stage. They are, in this sense, the concrete operationalization of the concepts whose causal properties we are investigating. Typical threats to construct validity include (but are not restricted to) the following (all of which deal with issues that are often most debated or debatable in any particular social scientific experiment)[2]:

- the inadequate explication of constructs;
- experimenter expectancies;
- subjects' reactivity to the experimental situation;
- treatment diffusion, which occurs when the treatment provided to the treatment groups ends up administered to some of the members in the control group.

[2] See appendix for the full list of threats in the typology as per Shadish et al. (2002). It is important to note that the list of threats conceptually differs from the idea of error in statistical modelling, and especially Type 1 and Type 2 errors and that it is not a check list to ensure, in a quasi-automatized way, that validity is met. The lists of threats are meant to guide the researchers through the process of design and evaluation experiments.

All of these threats can invalidate our capacity to make meaningful inferences from a given experiment and they can be rather ubiquitous, but it is often the case that when speaking loosely, we associate instead our concerns about construct to either internal or external validity issues. Take for example the threat of an inadequate explication of constructs: If an effect is found in some experimental setting that suggests that A does cause B, are we sure that it is not because A or B (or both) have been incorrectly operationalized? Would this causal effect generalize to different specifications of A or B?

We argue here that, because of the common tendency to think of the internal/external validity pair as if it exhausted the inferential realm of experiments, we are all too accustomed to deal with the questions regarding constructs as questions that pertain to either internal or external validity, even though this is not so in the Campbellian classification. According to the validity typology, to find out if the concrete *a*s and *b*s in the experiment do or do not stand for the target As and Bs that we are ultimately interested in is, prima facie, a question that has to do with the *validity of our constructs*, not with the "degree to which a causal relationship found in a given study generalizes across [conditions]" (i.e., external validity). Eminent methodologist Blalock (1961/2018) framed the question in terms of building an *operational model* based on a *conceptual model*, the challenge being precisely to find the 'right' variables to measure or to proxy the concepts we are interested in. Similarly, the worries associated with the confounding of an association with causality due to misspecification of either As or Bs is a question of construct validity, and not one "referring to inferences about whether the observed covariation between the presumed treatment and the presumed outcome reflects a causal relation." (i.e., internal validity). As Shadish et al. (2002) explain, construct validity does relate to questions both of confounds and of generalization, but the emphasis here is on the *validity of the constructs*, as opposed to the emphasis given to whether the causal hypothesis involved holds 'inside' or 'outside' the experiment.

To those versed in some of the discussions both in the philosophical and in the experimentalists' circles this might come as a surprise, for often, in both of those realms the internal/external distinction is used as if it jointly exhausted the validity questions that can be posed about a given experiment in relation to the inferences that we can make from it. Though statistical conclusion validity is also often forgotten as part of the validity typology, it is nevertheless picked up in the debates around hypothesis testing in experimentation. In the case of construct validity, however, the questions around it are not so much forgotten as dissolved into the questions of internal and external validity. A consequence of this dissolving away is the slippage into the idea that experiments come with their own inferences under the sleeve, thus oversimplifying the relation between experiments and the inferences that they allow. To ignore construct validity is to ignore that, when we experiment, we do need to make inferences (that in turn, can be valid or invalid) regarding whether the independent and dependent variables in our experiment (say, a and b) correctly represent the higher order constructs that are the target of our research question (A and B). Note, and this is crucial, that these inferences are not (necessarily) causal, and they are, in any event, different from the causal inference that we aim to test (which will relate to the general question of whether As cause Bs).

Once we conflate all these inferences about constructs into the dichotomy between internal and external validity, it becomes easier to assume, 'for convenience', that there is an equivalence between an experiment, understood as a material intervention, and the inferences that it allows. But this kind of thinking, we contend, is actually pernicious to a proper understanding of the relation between experiments and their inferential import. We call this view of experimentation, in which experiments correspond univocally with certain causal inferences, for convenience, the "one experiment, one inference" (1e:1i) assumption, which we examine in detail later in Sect. 3.

Is sum, through the years, the Campbellian validity classification became the standard in the specialized literature on research design in social sciences and gradually permeated the language of what was then a still restricted domain of social scientific experimentation, mostly limited to psychology and some areas in educational research. As experimentation became more popular in neighbouring fields such as epidemiology, part of the typology of validity—the distinction between internal and external validity—eventually came to be widely used in other social sciences, in biomedical domains, policymaking circles, and it has now become part of the standard vocabulary of the sciences and even, as we see next, in the corresponding philosophical debates. While the Campbellian project always considered the four types of validity, it is not an exaggeration to say that, whereas part of their validity project became universalized (the internal and external validity dyad), another part (statistical-conclusion and construct validity) got lost to many practitioners and experimental commentators, and also, to philosophers.

2.3 The philosophical debate around the internal–external validity dyad

The terms internal and external validity got gradually passed on to Philosophy of Science at the turn of the century, most notably popularized by Cartwright (1999, 2007), especially regarding her ideas on the type of evidence needed to inform policy decisions, by Guala (2003, 2005), who has discussed extensively the concept of external validity in reference to the practices of experimental economists and the problem of "artificiality" in the lab, and by Steel (2008), who has popularised 'external validity' in terms of extrapolation in a variety of research settings. Perhaps because both sets of problems (evidence for interventions, and artificiality of experimental settings) were connected to the broader issue of extrapolation, it is around this time that the notion of external validity becomes associated with extrapolation and often used interchangeably in some of the philosophical circles interested in social scientific practices. The "problem of external validity" thus became part of the philosophical lingo of those interested in extrapolation and in the policy implications of causal evidence (see for instance the debate between Guala and Steel on extrapolation and the role of analogical reasoning (Guala, 2010; Steel, 2010).

At present, and when used in a philosophical context, the terms external and internal validity tend to be used without reference to the broader typology in which Campbell and his collaborators inserted them. Often, the philosophical use

is not limited to the original methodological context but generalized to the relation between model and target, case study to universe of interest, or animal models to humans. In the philosophical discussion, then, internal and external validity often acquire broader, less specific meanings than those intended by their originator: internal validity is often invoked to convey the idea of reliability of inferences about genuine causal relations *within a study*, and external validity is associated with the idea of the generalizability of findings, thus *outside a study*.

The internal–external validity dyad was adopted rather easily by philosophers of science perhaps, in part, because it has been often seen as a clear and simple way to distinguish a realm where scientists enjoy some degree of control (the experiment, the model, the simulation even), versus the domain to which they would like to extrapolate their findings and where scientists have less control (the broader population, the target, the real world). However, this broader sense was not exactly coincidental with the original intent and meaning of Campbell's methodological project. The fact that the two sets of meanings for the terms internal/external validity (the technical, original senses, and the more general denotation) cohabitate is itself a source of problems and confusions, as we show next.

In recent years, some philosophical analyses have begun to look at these problems. However, most of the criticisms or reservations vis à vis Cook and Campbell's validity distinction have been directed to the internal/external validity dyad, understood mostly as separate from the methodological project in which it was inserted. Jiménez-Buedo and Miller (2010) and Jiménez-Buedo (2011), for example, have laid out some of the problems of these notions when used to assess experimental findings. Jiménez-Buedo and Miller have analyzed the paradox that emerges when we examine the idea, present in the literature, that there is a trade-off between internal and external validity and we contrast it with the idea, also prevalent in the literature, that internal validity is instead a precondition of external validity. Jiménez-Buedo (2011) deals with some conceptual inconsistencies found in standard interpretations of the internal/external validity distinction. More recently, there seems to be a growing unease with the formulation of extrapolation in terms of the "problem of external validity". Reiss (2019) for example, has claimed that the current reasoning on external validity puts undue emphasis on knowledge about the inferential source, at the expense of the knowledge of the target, and thus it encourages poor reasoning about evidence. Also, Deaton and Cartwright (2018) have found problems in some current interpretations of external validity in relation to the travelling of evidence from Randomized Controlled Trials, and they have deemed the term ambiguous.

More recently, Nagatsu and Favereau (2020) have also addressed the multiplicity of meanings of external validity. They describe the recent history of field experiments in economics as coming from two distinct intellectual traditions. The first strand in field experimentation can be seen as an extension of laboratory experiments, whereas the second strand instead is instead heir to a long-standing tradition from social sciences, i.e., the Campbellian tradition of field tryouts of ameliorative programs. Nagatsu and Favereau's historical argument points to the fact that the concept of external validity represents two different sets of problems in these two different intellectual strands (i.e., artificiality and generality,

respectively), and that these two issues are sometimes conflated in the literature by virtue of using the same term to refer to both.

Against this background, we propose instead a critical look at the concept of internal–external validity by putting it in relation to the *four-fold validity typology* in which it was meant to be inserted, and by looking at its specific context of application: experimentation in applied social research. We want to emphasize that internal and external validity are part of large methodological project, initiated in the context of applied research settings. Yet, the way philosophy of science has borrowed and discussed these terms (1) simplified the methodological context a great deal, moving from a four-fold treatment of validity to a dyad and (2) in doing so, connected internal–external validity with a supposedly predetermined set of inferences an experiment licences. In the next section, we suggest returning to the methodological project, emphasising the importance of construct validity.

3 Construct validity, objectivity, and the 1e:1i view of experiment

3.1 The structural design dimensions of experimentation and the list of threats to validity

Aware of the difficulties of reducing all inferential problems of a given experiment to the distinction between internal and external validity, Cambpell had moved away early on from the initial dyadic distinction between internal and external validity (1963), into the fourfold typology (1979). Nevertheless, he and his collaborators were cognisant that there was a generalized tendency to go back to thinking in terms of a dyadic distinction between internal and external validity that collapsed the fourfold typology into two. To avoid this collapsing, and partly, to give construct validity the role it needed, Campbell and his collaborators adopted, as an important addition to the Campbellian project, Cronbach's classification of the *structural design dimensions of experiment*, or the elements that can be varied among designs in *Units, Treatments, Observations and Settings (U.T.O.S.)*. This allowed the Campbellian scheme to differentiate between distinct domains relevant to the inferential space of a given experiment. Campbell thus adopted Cronbach's distinction between the different (material and inferential) domains relevant in empirical research. In this way, they started distinguishing between three relevant domains. First, the domain from which the data is collected in a particular study (utos). This is the domain of the concrete **u**nits, **t**reatments, **o**bservations and **s**ettings in a given experimental intervention (or more generally, in an observational design). Second, the domain at which the research question is asked (UTOS). This is the domain of the [higher order] constructs in which we are interested, but that we can only study, or access, through our more limited, empirical implementation. Third, and last, there was the domain of other potential questions of interest. These pertained to questions that our study did not intend to study directly, but where the findings of our study could in principle apply or be considered relevant (*UTOS).

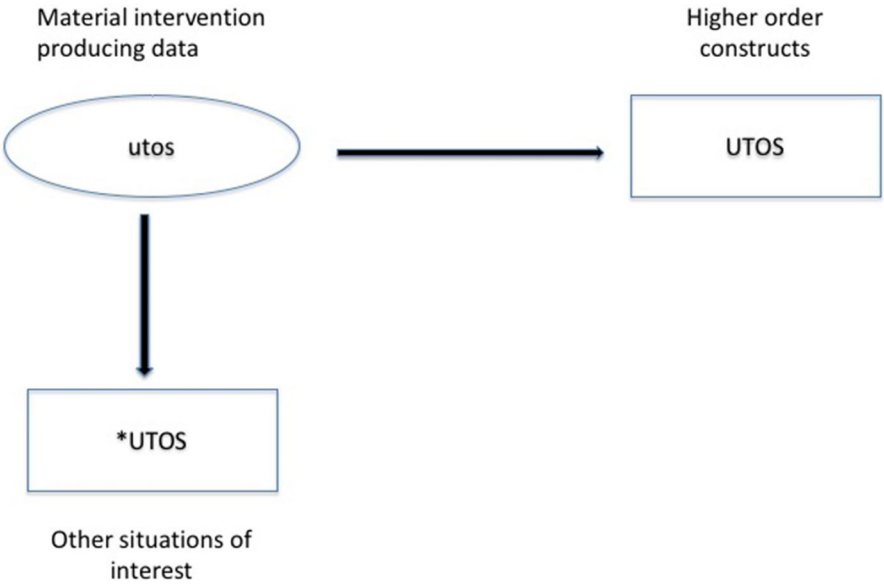

Fig. 1 Experimental inferences and the structural design dimensions of experiment (utos, UTOS, *UTOS)

Under this classification, questions about internal or statistical conclusion validity were restricted to inferences about the concrete experimental data (utos). Questions of construct validity revolved instead around the relation between utos and UTOS. Finally, other questions of interest regarding the possible generalization of findings to other contexts (i.e., external validity) would have to do with the relation between utos and *UTOS. So, for example, if we wanted to find out a whether extra-curricular tutorial programs run by voluntary parents can help the performance of students from disadvantaged social backgrounds (a question pertaining to the domain of higher order constructs, UTOS), we would collect the relevant data in a particular intervention study (utos), and could, perhaps, additionally, hypothesise about whether a related type of tutoring program could be also relevant in alternative scenarios, with different units, treatments, observations or settings [so, for example, we can wonder if a tutorial program run by voluntary parents can also have a positive effect on children's performances in more well-off communities (*UTOS)].

This addition to the Campbellian methodological project illustrates well how, in practice, the project did not actually presuppose a straightforward dichotomous distinction between the "inside" and the "outside" of a given experiment: both inferences about constructs (from utos to UTOS, as in construct validity) and inferences about other relevant units, treatments, etc. (from utos to *UTOS, as in external validity) are ampliative, extrapolative, inferences that allow to go from one domain to the other (see Fig. 1). In this way, and while construct validity inferences are ampliative, they are neither clearly *internal nor external* in the dichotomous sense in which often researchers (and philosophers) assume a neat distinction between the inside and the outside of an experiment.

It is worth insisting on the fact that Campbell was also preoccupied with the issue of a dichotomous interpretation of internal and external validity and the surrounding misunderstandings regarding these two labels (internal and external validity). This is reflected, for example, in Campbell's own proposal to rename internal validity as "local molar validity", to reflect a meaning close to its original intent: internal validity meant that we could infer that a *local*, concrete, "fat-handed" intervention (thus the *molar* bit), was in fact responsible for any kind of change in our output variable (Campbell, 1986). Internal validity, therefore, was not associated to a neat, well-delineated, causal hypothesis, but instead was associated to something more basic or brute. Internal validity had thus to do with the "on site", empirical detection of "some kind" of effect, i.e., any sign of causal efficacy (as seen in changes in the output variable).

As we can see, this type of conceptual understanding is rather different from the one that is often assumed (both by practitioners and, perhaps more upsettingly, by philosophers), in which internal validity came to be associated to the idea of a neatly distilled cause identified *within* an experiments, while external validity was associated to the problem of extrapolating this supposedly neatly defined causal claim to the messier, real *outside* world. But the Campbellian project had not been meant to provide a clean conceptual slate to differentiate between these alleged two opposed realms (inside/outside). It was meant, instead, as a methodological guide for social scientists relating causal identification questions to optimal research designs and it did so by studying systematically the possible confounders that actually mine the social scientific field, and that therefore threaten the capacity to correctly isolate the effect of interventions on outcomes of interest.

To be sure, this methodological project did contain a four-fold validity typology, but it is our contention that the classificatory aspect of the project was perhaps not as important as the fact that the typology provided researchers with a list of threats, understood as a practical tool that could remind researchers of possible confounds that they needed to take into account both in designing their research and later, in going from their data to their conclusions.

The practical orientation of the Campbellian project is best exemplified in the fact that through the years, new threats to validity were added to the different versions of the list that ensued. In turn, the fact that the classificatory aspect of the validity typology was secondary to the goal of listing known confounds is illustrated by the fact that some of the threats listed in the typology were reclassified among different validity types throughout the years. For example, "mono-operation bias" (defined as the bias that occurs when a given operationalization underrepresents a construct and measures irrelevant constructs) was initially considered a threat to external validity and later reclassified as a threat to construct validity. The same happened with the threat known as "compensatory rivalry" (which occurs when subjects selected to the control leg of a study are motivated to show that they can perform as well as those selected in the treatment group), which was initially considered a threat to external validity but ended up classified as a threat to construct validity.

In short, the Campbellian project was grounded in the practice of researchers, and its main aim (and main achievement) was to provide a practical guide of common confounders faced by social science practitioners, rather than a neat, analytical, and

definitive distinction between the types of inferences that we can draw from a given study (be it experimental or observational).

3.2 The internal/external distinction and the 1e:1i view of experimentation

Despite Campbell's best efforts to clarify some of the confusions regarding the validity typology, a dyadic interpretation of the internal/external validity has come to dominate much of the contemporary debate about experiments, including the philosophical debate. It is our contention that this interpretation of the internal/external validity dyad has helped encroach a distorted view of experimentation, one that overstates their capacity to adjudicate objectively among debated causal conclusions. But how does this interpretation of the internal/external validity distinction contribute to an inflated view of the "objectivity" of experiments?

In a nutshell, we argue that: (1) the dyadic understanding of the internal/external validity ignores the issue of construct validity. In turn, (2) ignoring construct validity opens the door for an oversimplified depiction of the relation between a given experiment (understood as a material intervention) and the inferences we can make from it (the 1e:1i view). We contend that once one slips into the 1e:1i view, (3) it is easy to overestimate the extent to which experiments can objectively adjudicate between contentious causal claims.

We have already seen in detail how interpreting the distinction between internal and external validity as a dyad describing exhaustively the inferential domain of an experiment implies collapsing away or ignoring the issue of construct validity ((1) above): To recall, and in the language of the *structural design dimensions of experiment:* speaking solely about the internal and external validity of an experiment means ignoring that inferences from a given concrete intervention (utos) are ampliative not only when they pertain to other related domains of potential interest (*UTOS), but also, when they refer to the constructs that a given experiment is aimed at representing (UTOS).

What about the 1e:1i view of experimentation, and its connection to construct validity? Certainly, no philosopher of science (and for that matter, no social scientist if he or she were to properly examine the idea) would really want to defend, as generally applicable, the notion that there is a univocal correspondence between an experiment and its (sole possible) related inference. Yet, our contention is that it is easy to inadvertently slip into this view once we ignore or assume away the relevance of construct validity. This is perhaps best seen in the fact that many philosophers and social scientists speak interchangeably about the internal/external validity of experiments and of the inferences from those experiments (recall, that Shadish, Cook and Campbell's even licence away that way of speaking "for convenience").

We argue that the experimental situations in which we can speak indistinctively about an experiment and the inferences that stem from them (the 1e:1i view) are one limiting case: one that only takes place in some (mostly applied) settings in which there are no fundamental questions that can raise regarding construct validity. Section 4 below presents one such case study, in the context of an ameliorative social

program, which is representative of the contexts in which the Campbellian project came about.

In Sect. 4 below we show by means of a case study how in some applied settings, where construct validity is not a matter of debate, it can sometimes be ok to assume that for a given experiment there is only one causal inference that is naturally connected to it). However, this same 1e: 1i view can lead to serious confusion and distortion when transported into other experimental settings in which construct validity is one of the very issues at stake in presenting one's experimental results. In Sect. 5 below, we illustrate this by presenting a second case study based on a lab experiment in behavioural economics.

The contrast between these two case studies is intended to show that, while slippage into the 1e:1i view can sometimes be less problematic, in cases in which construct validity is not at stake, it is however always deleterious to a proper philosophical elucidation of the inferential import of experiments. Next, we show how, in particular, it contributes to the misunderstanding (and overstating) the sense in which experiments can *objectively* elucidate or adjudicate among alternative causal claims (3). We argue that this elucidation rests on shared background assumptions, without which experiments can lead to different conclusions, just as any other (non-observational) research method or design.

3.3 The objectivity of experimental results and the neglected role of background assumptions

As we have seen, many of the current uses of the internal/external validity distinction assume that it can refer, indistinctively, both to a given experiment and to the inferences that we can make from it. But by embracing the idea that we can talk, interchangeably, about experiments (understood as concrete material interventions) and the inferences we make from them (in short, the 1e:1i view) we are flattening the complex relationship between the experiment and its inferential import. This false equivalence between experiments and the inferences that we can make based on them has another unwelcome consequence: it leads us to misrepresent by overstating it, the (limited) sense in which experiments can objectively adjudicate among contentious causal claims, by underplaying the role of background knowledge or assumptions. Background assumptions are a crucial element in the mediation between an experiment and the causal conclusions we can draw from it, so taking their role into account is also crucial in order to acknowledge that different observers of the same experiment can actually make different inferences from the same results. This, as we see next, is often ignored in the discussions of experimental results that use the internal/external validity distinction in its dyadic form.

Consider the famous World Bank Intensive Nutrition Programme in Tamil Nadu, TINP (as per Cartwright's 2012 account), where *mothers* were provided food support and education in order to improve the nutritional outcomes of children. The material intervention in India's Tamil Nadu was successful in the trial's more immediate objectives, as children's health improved due to the intervention but, what should have been inferred from this experiment? One could infer that providing food

support to mothers improves the nutrition of their children, and this is what was inferred by some officials and social scientists, who then went on to unsuccessfully replicate the results in Bangladesh. But perhaps an observer well-versed in Bangladeshi family structure could have inferred, alternatively, that the Tamil Nadu trial showed that providing support to "the person in charge of food supply" in a disadvantaged household does help to improve nourishment levels of the children: that was another of the possible inferences that could have been made from this experiment. In this case, the construct "person in charge of food supply" would have been involved in describing the inferences to be made from the experiment, an inference that could perhaps been made by the person with the right background assumptions (in this case, an observer who knew that in Bangladeshi households it is often the father or the mother in law who are in charge of managing the household's food supply).

The case is well-known in philosophical circles because this is not what was inferred from TINP by the Bank officials who went on to reproduce it in the Bangladeshi trial (BINP), and which consequently failed to improve children's outcomes by providing food support and education to their mothers. The researchers and contractors in charge of implementing the BINP must have made a description of the same intervention that is different to the one we now think is relevant, and their different background assumptions led them to make an inference that we now, in hindsight, known to be wrong. The example illustrates how interpreting the same experiment as involving the construct "mothers", rather than a more general "people in charge of the household food decisions", and having different background assumptions, can lead to drawing different conclusions or inferences from the same trial.

This example illustrates well the fact that different actors can make different inferences out of the same experiment, depending on the constructs used, which in turn, depend on their different background assumptions. Yet, this intervention is often invoked in the literature as an example of the external validity problems of RCTs (see e.g. Clarke et al., 2013; Howick et al., 2013; Marchionni & Reijula, 2019). In our view this way to look at the issue impoverishes our thinking about the relationship between RCTs and the inferences that we can derive from them, by ignoring the issue of construct validity, and indirectly resting on the idea that from a given trial a unique (causal) inference follows (1e:1i).

We also contend that this kind of reading of the evidential import of RCT, supposedly lacking in external validity, has indirectly contributed to a neglect of the crucial role of background assumptions in determining the inferential import of experiments, and in so doing, it contributes to a misunderstanding the role of RCTs in providing *objective* causal knowledge.

Cartwright, who popularized this example, is an important exception in this regard, since she has advocated the importance of gathering additional evidence about the array of "supporting factors", or background conditions, that can help the success of an intervention, and in so doing she too has emphasized the importance of background knowledge in securing the relevant claims from an experiment. However, some of her criticisms to RCTs, and in particular some readings of her claim whereby RCTs are good at telling us what "worked there" yet not good at telling us whether the same intervention would "work here" (for us), may have contributed to

encroach an overly simplistic picture of the relation between RCT results and the inferences we make from them. Her example of the Tamil Nadu and Bangladeshi programs is often invoked to show the "external validity problems" faced by RCTs and it does in fact show that it is often difficult to know whether a given program will "work here" (for us). What we want to stress is that this example, though this is not what is normally used for, *also shows* that it is often not obvious to agree on the description of how or whether it "works there" (its original implementation), and that different researchers, depending on their different background assumptions may come to interpret differently the constructs involved in a given intervention. Though only recently Cartwright has noted that "external validity" is a somewhat confusing term, she has used it within a dyadic internal/external interpretation, and helped popularized it among philosophers. She has mainly used it to describe the limitations of RCTs for extrapolative purposes, as part of an important and necessary effort to appease some of the overly enthusiastic versions of EBP. But this dyadic use of the internal/external validity distinction has too contributed to the neglect of construct validity, and to an overly simplistic view of the relationship between an experiment and its inferential import, by helping to obscure the fact that the same experimental intervention can be conceptualized as representing different constructs, and that often background assumptions is crucial in determining the inferences that we make from one and the same experiment.

The popular view that RCTs are "good in internal validity but bad in external validity" is a good example of the problems that a poor conceptual basis can contribute to, by disregarding that the inferences that we make from experiments are not predetermined and that they depend a great deal on our background assumptions and on our inductive caution or lack thereof. To avoid this kind of inoperable cliché we suggest avoiding an overly simplified view of how experiments relate to potential inferences. One such view needs to underline that extrapolation is actually a difficult business across the board (whether the evidence is experimental or observational) and, especially, that the same experimental intervention can be conceptualized as representing different constructs, and that background assumptions can be crucial in determining what the appropriate constructs are. Because background assumptions can determine the inferences that we make from an intervention, experiments can only contribute to solve causal controversies if the agents involved already share a common set of background assumptions, i.e., if there is a pre-existing consensus about the validity of the constructs involved in a given experiment. So, in this sense, experiments can contribute to the objective adjudication of causal controversies, only by piggybacking on a great deal of pre-existing shared or consensual background assumptions.

The next two sections of the paper illustrate the limits of common (dyadic) uses of the internal and external validity distinction. We contend that the notions of internal and external validity as commonly understood may be of some heuristic use in some very applied settings (where experiments do not rely heavily on tentative or contentious background theories). We also show how some uses of the internal/external validity dyad can nevertheless be a source of confusion when applied to experimental practices in which construct validity is a matter of contention, or open for discussion. Section 4 provides an example of field experimentation in the

context of ameliorative programs, the context where the Campbellian project came about, and where debates regarding construct validity would typically be about specific empirical issues. Section 5 instead deals with a case of behavioural economics, where both the operationalizations and the constructs generally under some degree of theoretical dispute. Our goal is to show that in the latter case, the labels of internal/external validity are of very limited use and even, sometimes, counterproductive to the understanding of the workings of an experiment's inferential import.

4 Undisputed construct validity in an ameliorative programme

To recall, the Campbellian program is an ambitious methodological project that is pragmatically oriented, born in the context of educational research, and then developed and refined on the basis of the practice and experience of those directly involved in the implementation and assessment of a wide array of ameliorative programs. In this section, we show an example that illustrates the kind of pragmatic concern and the intellectual milieu in which Campbell and his collaborators came up with their notions of internal and external validity and we argue that, under some conditions, the use of these terms may be justified pragmatically. We contend, however, that because this only shows a limiting case (one in which there is an easy or straightforward correspondence between an experiment and the inferences we can make from it), the terms do not have a relevant role to play in a broader epistemological project regarding the inferential import of experiments generally.

To illustrate this, let us consider an example of an applied intervention, a case study in which a trial is designed to delay the age of initiation of first sexual intercourse in a population of urban junior high school students (Aarons et al., 2000). Through this intervention, six Washington D.C., junior high schools were randomly assigned to either the intervention or the control condition for an educational program, where three health professionals (one per intervention school) implemented a programme consisting of reproductive health classes and some educational activities during two school years. The outcomes of 'initiation into sexual activities', 'attitudes toward delayed sex and childbearing', and 'sexual knowledge and behavior' were assessed at four time points to determine whether the intervention had an impact in the outcomes mentioned.

This particular intervention provides us with some particular operationalization (utos), that we can think of, as representing the higher constructs relating causally the educational activities developed and the output variables related to delaying of sexual activity (UTOS). The causal inferences drawn from the analysis of the intervention can be then, perhaps, generalizable to other units, treatments, outcomes and settings (*UTOS). As exemplified in this case, the Campbellian scheme of validity is all about how to establish inferences from the experimental data (utos) to the domain of the question of interest (UTOS), or to even other potentially relevant domains (*UTOS), but the relationship between the very local (utos) and the research question (UTOS) is, so to say, unproblematic *by construction*: the Campbellian project is one of *design analysis*, and its aim is, precisely, to systematize the way in which we convert a research question (e.g., 'Are educational programs

encouraging delayed sexual activity efficacious?') into a particular, concrete, design that we can test on the ground—e.g., 'Does a particular program with a given curriculum, and implemented by program about contraceptives provided by health professionals delay the initiation toward sexual activity in 7th and 8th graders in a set of schools of Washington DC?'.

To be sure, there are many things that can go wrong when we go from the domain of questions to the domain of particular operationalizations (and thus, all the array of threats to internal, statistical conclusion, or to construct validity), but, in the type of field trial of ameliorative programs that this first wave of experimenters were implementing, these concerns were *pragmatic problems* for which intersubjective agreement could in principle be reached. The idea is that researchers (or anyone, for that matter) would normally not expect that we can find out, via a single trial, whether*, in general*, information about sexual health delays the age of initiation to sex in adolescent students. Also, and by the same token, most people would expect that by setting up a concrete intervention that (concretely) incarnates that putative general causal relation we are contributing at least *partially to finding out more about the ways* in which A affects B, or in this case, the ways in which sexual health information might or might not contribute to delay the age of initiation to sexual activities in adolescent students.

These are precisely the questions to be addressed in assessing construct validity, as discussed earlier in Sect. 3. In other words, in the Campbellian setting of applied empiricism and research design, there is not necessarily a univocal relation between the concrete operationalization utos and the higher construct UTOS, but there tends to be a sense in which the concrete operationalization of the trial can nevertheless be considered, rather consensually, as partially contributing to knowledge about some relatively well-understood higher construct (in this case, some general notion of "teenage pregnancy" and of "information campaign"). There is therefore a sense in which a given concrete trial can be thought of as empirically contributing to the higher order construct, and a common shared sense in which the trial is classified as belonging to a particular phenomenon. As the TINP and BINP cases illustrate (also concerning ameliorative programs), this is not to say that there cannot be substantive differences in the interpretation of the relevant constructs involved when dealing this type of trial. But, in contrast with our next example, there is at least conceivable scenarios in which we can think of straight forward agreement about the constructs involved.

Regarding the external validity (in this setting, the capacity to extrapolate the findings of a given trial to other, different situations, of interest), there seems to be little to say systematically: researchers and users must use results (no matter how carefully crafted the design) with care, for there is no guarantee, as Cartwright would put it, that what worked "somewhere" will also work "here". Campbell explained throughout the years that external validity could be understood as a principle of proximal similarity: a lot of how we reason about extrapolation does seem to rely on analogical reasoning, a point that was also emphasised in the aforementioned Guala-Steel debate.

The context of ameliorative programs, and the kind of inferential and practical questions they pose, is the milieu in which Campbell and collaborators came up with

their impressive methodological guide. It is also in this context that Campbell and collaborators thought that one may sometimes, "for convenience", speak indistinctively about the validity of either *inferences or designs.* As we have tried to show, this context lends itself to providing examples in which there is a rather straightforward correspondence between a given experimental design and the inferences that we can derive from it. In other words, in the case of these pragmatic interventions, assuming a correspondence between a given operationalization and the causal inferences that stem from it can be seen as uncostly because, in this settings, there will often be enough agreement regarding the background assumptions that goes in deriving an inference from a given trial: the validity of the constructs that are represented by concrete operationalizations is generally straightforward. In our case study, we might generally agree with the idea that the programmes implemented do represent at least a partial aspect of the teenage pregnancy and the prevention campaigns that are our ultimate interest.

The question remains as to whether in these type of applied context one can use meaningfully the internal/external validity distinction as something other than as a list of threats, but rather, as a way to divide the types of inferences that can be made from an experiment as either internal or external, in the common (yet not strictly Campbellian) sense in which internal validity as supposedly inferences about what goes on inside the experiment, and external validity refers to the generalization of the causal relationship identified in the experiment as extending to other, outside situations of interest.

The exercise of classifying the type of inferences in this way seems to us, in any event, of little use; what is important, as we said, is not so much to know whether a specific inference can be internal or external, but to know whether it is valid or invalid. At least, in cases in which construct validity is settled, or unproblematic, it seems that we can identify some sense in which to use this distinction is intelligible: if construct validity in the experiment is unproblematic, internal validity could refer to the causal relation (expressed in the language of the higher order constructs) in the experimental sample. External validity can become this more diffuse notion linked to extrapolation, also expressed, again, in the language of the higher order constructs. We nevertheless want to stress that though this interpretation of the terms can be, in some contexts, used loosely but intelligibly in this way, this is not the original senses in which they were intended, and after all requires an equivalence between an experiment and the inferences we derive from it: the experiment/inference would be internally valid if A caused B in the experimental sample, and externally valid if A caused B outside of the sample, where in our case, this would mean that "information campaigns" cause "a reduction in teenage pregnancies" in either inside the experiment or outside of it.

In the next section we show that, outside of the Campbellian project, the lack of an obvious correspondence between a given intervention and the causal inferences it can licence is far from uncommon and that, under these circumstances (in which construct validity is not a fixed or settled matter), speaking of the types of inferences from an experiment as either internal or external ends up being a source of confusion, rather than clarification.

5 Open ended constructs in behavioural economics

In this section, we pose the question of whether the internal/external validity distinction can relevantly describe social scientific experiments generally, or whether, instead, the dyadic use of the terms may muddle important aspects of current experimental practices that should instead be addressed by discussing constructs and construct validity. We focus on an example from the growing field of experimental behavioral economics, and in particular we relate our example to one of the most commonly used experimental games, the Dictator Game (DG).

5.1 The dictator game and "the power of asking"

In the DG, the experimenter allocates some fixed quantity of money to player 1, the Dictator, who then must decide how much, if any, he or she wants to share with player 2, the Respondent. The game ends there and Dictator and Recipient collect their share of the initially endowed sum according to the Dictator's offer. The results of the standard DG show that roughly half of the Dictators depart from the earnings maximizing strategy and choose to give some money, the mean allocation being 20% of the initial endowment. Moreover, a consistent minority of dictators choose to split the sum in two similar sizes (Camerer, 2003). One of the common uses of the standard DG is to employ it as a baseline treatment against a modified DG, where the latter embodies the intervention of interest: often the modification of the DG consists in adding to a standard DG an extra normative cue (e.g., the modification of the identity of the Recipient- being an anonymous player versus being an identified person in need). In such modified DGs, researchers compare the results of the baseline and the treatment of interest to see if there are relevant differences in the behaviour of subjects. The relevant differences are thus attributed to the normative cue introduced in the modified DG.

Fielding and Knowles (2015) is an example of this kind of intervention. The authors motivate their experiment as one that tries to contribute to knowledge about "the power of asking". More precisely, the experiment wants to test whether the level of generosity is higher when a visual cue is augmented by a face-to face verbal invitation to consider a charitable donation. Subjects in the treatment and baseline conditions earn some payment by completing a task (filling out a survey) in the experimental lab. Upon finishing their task and receiving payment, subjects in the *treatment group* are told by the experiment's administrator that they can, if they so wish, donate some of their money to a well-known NGO before leaving the premises. On their way out, in an adjacent room, subjects are faced with a box where they should deposit their completed surveys. Next to the box, there is also a transparent urn with the NGO's name and logo, and participants can donate some money anonymously. Subjects in the baseline group, in contrast, are not verbally asked if they wish to make any donation but they too enter the adjacent room to deposit their task upon leaving, where they also find the small urn identified by the NGO's name and logo. When compared across treatments, the average donation level is higher

in the condition in which subjects are asked verbally to donate.[3] Also, importantly, the level of donations drops to close to zero in the baseline condition (to recall, the condition in which subjects merely see, but are not asked to donate to, the NGO's donation urn).

5.2 DG, 1e:1i, and construct validity

Fielding and Knowles' intervention has a clear experimental hypothesis (p.723): "When a visual cue is augmented by a face to face verbal invitation to consider a charitable donation, the level of generosity is higher", yet the question regarding the inferential import of the experiment's ensuing result remains open. In the lingo of the Campbellian-Cronbach synthesis, the experimental hypothesis would correspond to the realm of the concrete implementation of the intervention, as embodied in the results produced (represented as an ensemble of data about the particular units (persons), treatment implementation, observation, and setting (i.e., utos). Yet, the domain to which this experimental hypothesis can apply, however, is much less clear: What is exactly the domain about a question regarding "higher order constructs" (UTOS) is asked?, and also: can we differentiate this domain (UTOS) from other units, treatments, variables and settings not directly observed to which we would like to generalize the findings (*UTOS)? It is as if, in this type of experiment, the experimentalists make no commitments with regard to the construct validity of their trials. The relation between the operationalizations in the experiment and their correspondence to higher order constructs remains open: do the "visual cues" and the "generosity" referred to in the experimental hypothesis apply to other games played in the lab? Or does this concrete setting aim at representing behaviour outside the lab, too, as it might be relevant to the behaviour of people when they donate to NGOs? The proponents of the experiment do not commit to any particular set of constructs, nor to any particular domain of application.

This is attested, also, by the fact that the authors motivate their question as potentially relevant to a number of both applied domains and theoretical debates. First, Fielding and Knowles argue that their results are relevant in applied settings like those faced by charities: is it effective to verbally ask for donations, or is it counterproductive? Second, the authors also motivate the relevance of their experiment against a miscellaneous theoretical background: on the one hand recent research in evolutionary biology suggest the "the potential importance of the distinction between visual and verbal cues" in the coordination of mutually beneficial actions On the other hand, the authors conjecture that it is possible that visual cues be easier to ignore than verbal cues when people wish to preserve an altruistic self-image). Third, according to the authors, the experiment also can help explain the standard DG results, or the puzzling contrast between the high level of donations observed in the DGs enacted in the laboratory and the, on average, overall low societal levels of

[3] The article also includes an additional intervention that tests the effects of having more or less "lose change" over the level of donations, but because it is orthogonal to the intervention of interest here we will ignore it, in the interest of brevity of exposition.

charity donation. The intervention is therefore also an attempt to show under what conditions the DG can be a good predictor of donations to charities: if donations are high in the DG it might be, after all, because the DG setting "asks" participants for their money in a particularly persuasive way, even if it only does so implicitly.

Two remarks are in order. First, the experiment and its results licence inferences regarding other Dictator Games. It can be inferred, for example, that as soon as the DG is "naturalized" into a charity game, the level of donations falls dramatically, *unless* subjects are verbally asked to donate. The experiment could also be interpreted as providing additional insights about the standard DG results: high levels of donations seem to depend, in a standard DG, on the fact that the game, by its mere structure, *implicitly* demands these donations of Dictators. In a more naturalized setting like Fielding and Knowles', this demand needs to be verbal to obtain similar results. Thus, under this first interpretation, the domain of higher constructs (UTOS) of which the specific testing is a sample would here be constituted by a population of other Dictator Games.

Second, and alternatively, the experiment can be seen as a charity game that is played in a quasi-natural setting; its results would speak directly to the levels of donation in real charities, showing that adding verbal cues to visual one increases the level of donations in this setting. The domain of higher constructs (UTOS) to which the experimental samples belong however prove difficult to describe: this domain would be constituted by charity donations in which there is both a verbal and a visual request, but this cannot exhaust its description, for there are additional crucial elements in the intervention and they define, too, the complex experimental setting (think of these other crucial elements: the existence of a room in which participants find, prominently displayed, a donation urn with the logo of an NGO (1); a preceding task for which participants receive payment (2), etc.). We can thus not limit our UTOS description to a meagre "charity donations in which there is both a verbal and a visual request", yet, and at the same time, we cannot define this domain of higher constructs so exhaustively that we end up re-describing, again, the experimental intervention. This domain of higher constructs seems, in Fielding and Knowles' experiment, to resist clear boundaries or, put in other words, Fielding and Knowles intervention (and their corresponding description of the experiment) does not seem to be meant to be confined to any particular domain.

Our contention is that, in cases in which construct validity is far from a straightforward matter, using the distinction between internal and external validity of an experiment to describe the inferential import of an experiment leads to confusions and ambiguities. In fact, a common interpretation in the literature suggests that the standard DG has problems of external validity (Bardsley, 2008; List, 2007). We know, however, that this kind of use of the term is problematic, since the proper use of internal or external validity refers to inferences from the experiment, and, cannot refer to the experiments themselves, for we cannot assume that any given experiment can only be used for a predetermined set of inferences. In the case of Fielding and Knowles' intervention, this becomes particularly apparent: the experimental question is clear (do verbal cues—in addition to visual ones—increase the level of generosity of donations?), but there is no clear domain about which this question is asked.

On the one hand, it is clear that the regularity found in the experiment cannot be conceived as holding universally, even not understood as a tendency law held by a ceteris paribus clause, for the authors themselves acknowledge than in some contexts, verbally asking for donations can provoke rejection in potential donors. On the other hand, the phenomenon identified in the experiment is thought to illuminate an aspect relevant to our interpretation of other DG design, and these, in turn, are thought of as providing relevant insights about charity donations. But, if we do not think that there is a univocal correspondence between a given intervention and the inferences we can make from it, then speaking of the internal or external validity of those inferences also becomes of little use: for each of those inferences we will have to argue about whether they are correct or incorrect, but it will be of little additional use to say whether they are also internal or external.

Contrary to what is often assumed, the Fielding and Knowles example shows how in many instances there is not a clear sense in which we can classify the inferences to a given experiment as belonging to either the "inside" or the "outside" of the experimental realm. We argue that this is particularly obvious in cases in which the higher order constructs that are supposed to be represented by particular operationalizations remain an open question. In these cases, the (material, experimental) intervention lends itself to be described in many ways, and under each of these descriptions, and depending on everything else we know, perhaps from previous experiments, which we might consider to be "related" to this particular intervention (i.e., depending on our background assumptions), we might feel licenced to interpret the results in different ways. In other words, the inferences licenced by the experiment will depend on our descriptions of the experiment, and by the same token, the same experiment might licence different inferences depending on how we describe it.

There is not, therefore, a univocal correspondence between the experiment and the higher order constructs that it represents, or the inferences it licences: there is not a "one experiment, one inference" relation that can justify that we can talk indistinctively about the validity of an experiment and the validity of the inferences it licences. We submit that this is not unique to DG experiments or to this particular intervention, but is instead quite widespread across experimental practices and even field experiments in the social sciences. In some settings, mainly, in more applied settings, we can often assume that there can be some sort of default agreement on the type of inference that a given experiment allows, perhaps (or at least) on pragmatic grounds (though this need not be the case, as we have tried to argue in the TINP/BINP case). We should be aware, however, that this type of agreement over the relation between a given experiment and the causal inferences that it can allow has causes that are external to the experiment itself, and are that are often rooted in our shared background assumptions. In the absence of that type of shared background knowledge or assumptions, we are very likely to find a plurality of views regarding what we are licenced to infer from a given intervention. Using the language of internal and external validity, and buying with it the 1e:1i view of experimentation obscures this fact and presents experiments unduly as loci for objective causal claims.

6 Conclusion

The social sciences have been striving to establish the objectivity of the results of their study ever since they started being codified in clear and recognizable methods. This, to be sure, has happened since social sciences methodologists started developing quantitative methods in the second half of the nineteenth century. In this trend, another important milestone has been the introduction of experimental methods in the social domain.

In this paper, we look at experimental methods used in social research and investigate the idea that a certain view of experimentation came to be seen as the objective way to adjudicate among contentious causal claims. We argued that experimental methods really carried out a revolution in the way causal claims are identified in social research. In the methodological literature, one of the most notable contributions has been that of Cook and Campbell in the late 1960s and in the 1970s and their approach gained much popularity, especially via the distinction between internal and external validity.

We revisited the distinction between internal and external validity to show that it is not some of its uses are ambiguous and can lead us into wrong-headed beliefs about what experiments can or should achieve. In particular, the distinction, if used as a dyad, is often based on the idea that the inferential import of an experiment can be neatly distinguished into two realms corresponding to the inside and the outside of the experiment. This view is, in turn, based on but also reinforces the assumption that every experiment has one set of inferences attached to it (what we called the 1e:1i assumption). We explained why this is not the case in many experimental settings, and we further illustrated our point using two case studies (one about teenage sexual health and another about testing the 'power of asking' using the experimental setting of the Dictator Game). We hope this paper shows that, if we are correct in disentangling the 1e:1i assumption as being at the basis of this use of the internal–external validity distinction, it follows that these concepts are not useful to account for experimental practices.

Our argument, however, is not geared towards dismantling an alleged pillar of social science research (namely the internal–external validity distinction). On the contrary, we contend that the Campbellian typology, understood as a way to systematize common confounds faced by social scientific experimenters, remains useful. We instead advocate for the re-embedding of the internal/external validity dyad in the entire four-fold typology of validity (with particular attention to construct validity), and to reflect on the role of background knowledge or assumptions in experimental practices and the alleged objectivity of experimental methods, of which RCTs are admittedly a paradigmatic approach. To this end, we discussed a stock example from the RCT and policy literature, namely the Tamil Nadu Nutrition Programme and the Bangladesh counterpart, to explain the way in which construct validity is crucial.

We have contrasted and compared our arguments with others offered in the recent debate around the limitations of RCTs, and most notably in the work of Cartwright nad Hardie (2012). In fact, there is a sense in which objectivity is also a

core preoccupation motivating Cartwrights' arguments. She has long expressed critical concerns about using RCTs to 'export' results for policy purposes. Her most famous case study concerned blatantly failed interventions to improve child nutrition in Bangladesh by educating women (BINP), based on very successful interventions doing exactly the same in the region of Tamil Nadu (TINP). This is synthetized in Cartwright's motto that while RCTs are good to show us that a given policy or treatment "worked there (or somewhere)" they are of limited use to tell us whether they will "work here". Or, in other words, results of RCTs are inherently local in character.

While Cartwright and colleagues are certainly aware that even to establish results 'here', 'there', or 'somewhere' takes a lot of work and pain, some of the ways in which these expressions have subsequently been used may have contributed to crystalize the idea that experiments can generally rather straightforwardly establish that a given treatment has worked somewhere', (*within* the original research (or experimental) setting, and are thus, in the usual terminology, *internally valid)*, yet that their problem is though that we cannot be sure about whether they will work for us, "here", given that they offer knowledge that is local (and thus, not, in the usual terminology, externally valid). Our diagnosis is that we should also be more aware of what exactly we infer from the success stories, and we should address the fact that those inferences about what works "somewhere" are also influenced by our own incomplete background knowledge but in our view, the contraposition of the notions of internal and external validity as used normally in the philosophical discussion often obscure this fact.

In a sense, the argument we propose invites to further substantiate the claim that even *within* a study the causal inferences can be far from straightforward. But to do that, we need to abandon certain readings of the internal–external validity distinction, and to resume to understand them within the whole four-fold typology of validity. The Campbellian typology contains a list of threats to the various types of validity that has proven to be a great practical tool helping social scientists to assess in a continuous way the whole design and implementation research process. A distorted reading of some of some of its elements (internal and external validity) has been interpreted by some as a conceptual elucidation of the inferential role of randomized trials. This distinction has been consequently used to drawing a sharp divide between the inside and the outside of the experiment, but in so doing it presents a view of experiments that neglects crucial aspects in the relation between the material intervention, our background assumptions, and the inferences we ultimately make.

In a sense, this paper is one of many attempts to re-assess how philosophical analyses of social science experimental methods need to make proper room for aspects that have so far received less attention, such as background knowledge or assumptions, and the development of constructs. We hope that in so doing we can also contribute to creating the intellectual space for a deeper reflection on how epistemic and non-epistemic values are bound to influence (social science) research and its objectivity.

Appendix: List of threats to validity

In this Appendix we report the list of threats to statistical conclusions, internal, construct, and external validity, as they are summarized in a series of table in Shadish et al. (2002).

Threats to statistical conclusion validity (Shadish et al., 2002, p. 45)

1. Low Statistical Power: An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.
2. Violate Assumptions of Statistical Tests: Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.
3. Fishing and the Error Rate Problem: Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.
4. Unreliability of Measures: Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.
5. Restriction of Range: Reduced range on a variable usually weakens the relationship between it and another variable.
6. Unreliability of Treatment Implementation: If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.
7. Extraneous Variance in the Experimental Setting: Some features of an experimental setting may inflate error, making detection of an effect more difficult.
8. Heterogeneity of Units: Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.
9. Inaccurate Effect Size Estimation: Some statistics systematically overestimate or underestimate the size of an effect.

Threats to Internal validity (Shadish et al., 2002, p. 55)

1. Ambiguous Temporal Precedence: Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.
2. Selection: Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. History: Events occurring concurrently with treatment could cause the observed effect.
4. Maturation: Naturally occurring changes over time could be confused with a treatment effect.

5. Regression: When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.
6. Attrition: Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.
7. Testing: Exposure to a test can affect scores on subsequent exposures to that an occurrence that can be confused with a treatment effect.
8. Instrumentation: The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.
9. Additive and Interactive Effects of Threats to Internal Validity: The impact of a threat can be added to that of another threat or may depend on the level of another threat.

Threats to construct validity (Shadish et al., 2002, p. 73)

1. Inadequate Explication of Constructs: Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and construct.
2. Construct Confounding: Operations usually involve more than one construct, and failure to all the constructs may result in incomplete construct inferences.
3. Mono-Operation Bias: Any one operationalization of a both underrepresents the construct of interest and measures irrelevant constructs, complicating inference.
4. Mono-Method Bias: When all operationalizations use the same method (e.g., self-report), that method is part of the construct actually studied.
5. Confounding Constructs with Levels of Constructs: Inferences about the constructs that best: represent study operations may fail to describe the limited levels of the construct that were actually studied.
6. Treatment Sensitive Factorial Structure: The structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used.
7. Reactive Self-Report Changes: Self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made.
8. Reactivity to the Experimental Situation: Participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation, and those perceptions are part of the treatment construct actually tested.
9. Experimenter Expectancies: The experimenter can influence participant responses by conveying about desirable responses, and those expectations are part of the treatment construct as actually tested.
10. Novelty and Disruption Effects: Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as part of the treatment construct description.

11. Compensatory Equalization: When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment, and this action must then be included as part of the treatment construct description.
12. Compensatory Rivalry: Participants not receiving treatment may be motivated to show they can do as well as those receiving treatment, and this compensatory rivalry must then be included as part of the treatment construct description.
13. Resentful Demoralization: Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must then be included as part of the treatment construct description.
14. Treatment Diffusion: Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.

Threats to External validity (Shadish et al., 2002, p. 87)

1. Interaction of the Causal Relationship with Units: An effect found with certain kinds of units might not hold if other kinds of units had been studied.
2. Interaction of the Causal Relationship Over Treatment Variations: An effect found with one treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.
3. Interaction of the Causal Relationship with Outcomes: An effect found on one kind of outcome observation may not hold if other outcome observations were used.
4. Interactions of the Causal Relationship with Settings: An effect found in one kind of setting may not hold if other kinds of settings were to be used.
5. Context-Dependent Mediation: An explanatory mediator of a causal relationship in one context may not mediate in another context.

Acknowledgements The authors wish to thank the editors and reviewers for helpful suggestions. We also would like to thank participants and organizers of the "Objectivity in Social Research" workshop, May 2019 (Bergen University), the INEM19 symposium "Extrapolation and external validity" at the University of Helsinki, and the 2018 External Validity SPSP Symposium, and 2019 Workshop Reasoning about Evidence, both at Ghent University.

References

Aarons, S. J., Jenkins, R. R., Raine, T. R., Nabil El-Khorazaty, M., Woodward, K. M., Williams, R. L., Clark, M. C., & Wingrove, B. K. (2000). Postponing sexual intercourse among urban junior high school students—A randomized controlled evaluation. *Journal of Adolescent Health, 27*(4), 236–247. https://doi.org/10.1016/S1054-139X(00)00102-6

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics, 11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Blalock, H. M. (2018). *Causal inferences in nonexperimental research*. Chapel Hill: University of North Carolina Press.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences, 7*(5), 225–231. https://doi.org/10.1016/S1364-6613(03)00094-9

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation*, *31*, 67–77.

Cartwright, N. (1999). *The Dappled world: A study of the boundaries of science*. Cambridge University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine, 57*(6), 745–747. https://doi.org/10.1016/j.ypmed.2012.10.020

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Davies, P., & Boruch, R. (2001). The Campbell collaboration. *BMJ, 323*(7308), 294–295. https://doi.org/10.1136/bmj.323.7308.294

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Fielding, D., & Knowles, S. (2015). Can you spare some change for charity? Experimental evidence on verbal cues and loose change effects in a dictator game. *Experimental Economics, 18*(4), 718–730. https://doi.org/10.1007/s10683-014-9424-x

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70*(5), 1195–1205. https://doi.org/10.1086/377400

Guala, F. (2005). *The methodology of experimetal economics*. Cambridge University Press.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science, 77*(5), 1070–1082. https://doi.org/10.1086/656541

Hammersley, M. (1991). A note on Campbell's distinction between internal and external validity. *Quality and Quantity, 25*(4), 381–387. https://doi.org/10.1007/BF02484586

Hammersley, M. (1993). Abandoning internal and external validity: A response to swanborn. *Quality and Quantity, 27*(2), 217–218. https://doi.org/10.1007/BF01102735

Heukelom, F. (2011). How validity travelled to economic experimenting. *Journal of Economic Methodology, 18*(01), 13–28. https://doi.org/10.1080/1350178X.2011.556435

Howick, J., Glasziou, P., & Aronson, J. K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics, 34*(4), 275–291. https://doi.org/10.1007/s11017-013-9266-0

Jiménez-Buedo, M. (2011). Conceptual tools for assessing experiments: Some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology, 18*(3), 271–282. https://doi.org/10.1080/1350178X.2011.611027

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*(3), 482–493. https://doi.org/10.1086/519249

Marchionni, C., & Reijula, S. (2019). What is mechanistic evidence, and why do we need it for evidence-based policy? *Studies in History and Philosophy of Science Part A, 73*(February), 54–63. https://doi.org/10.1016/j.shpsa.2018.08.003

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. *New Directions for Program Evaluation, 1986*(31), 47–66. https://doi.org/10.1002/ev.1433

Morgan, K. J. (2016). Process tracing and the causal identification revolution. *New Political Economy, 21*(5), 489–492. https://doi.org/10.1080/13563467.2016.1201804

Nagatsu, M., & Favereau, J. (2020). Two Strands of field experiments in economics: A historical-methodological analysis. *Philosophy of the Social Sciences, 50*(1), 45–77. https://doi.org/10.1177/0048393119890393

Parkhurst, J. O., & Abeysinghe, S. (2016). What constitutes "Good" evidence for public health and social policy-making? From hierarchies to appropriateness. *Social Epistemology, 30*(5–6), 665–679. https://doi.org/10.1080/02691728.2016.1172365

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden: Blackwell Pub.

Reiss, J. (2019). Against external validity. *Synthese, 196*(8), 3103–3121. https://doi.org/10.1007/s11229-018-1796-6

Reiss, J., & Sprenger, J. (2017). 'Scientific Objectivity'. In E. N. Zalta (Ed.), Winter 2017. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Steel, D. (2008). *Across the boundaries extrapolation in biology and social science*. Oxford University Press.

Steel, D. (2010). A new approach to argument by analogy: Extrapolation and chain graphs. *Philosophy of Science, 77*(5), 1058–1069. https://doi.org/10.1086/656543

Tellings, A. (2017). Evidence-based practice in the social sciences? A scale of causality, interventions, and possibilities for scientific proof. *Theory and Psychology, 27*(5), 581–599. https://doi.org/10.1177/0959354317726876

Authors and Affiliations

María Jiménez-Buedo[1] · Federica Russo[2]

Federica Russo
F.Russo@uva.nl

[1] Department of Logic, History and Philosophy of Science, UNED (Universidad Nacional de Educación a Distancia), Paseo de senda del rey 7, 28040 Madrid, Spain

[2] Department of Philosophy, University of Amsterdam, Amsterdam, The Netherlands