Empirical underdetermination: A bigger problem for the social sciences?

María Caamaño[1]

(forthcoming in the *Proceedings of the 17th International Congress of Logic, Methodology and Philosophy of Science and Technology*, College Publications).

Proyecto PID2021-126416NB-I00 financiado por MICIU/AEI /10.13039/501100011033 y por FEDER, UE

Abstract

The now familiar idea that the detection of an empirical phenomenon is inferred from a complex collection of data (Bogen & Woodward 1988, Woodward 1989, 2000, 2010, McAllister 1997, 2011, Glymour 2000, Harris 2003, Massimi 2007, Leonelli 2015, 2019, Bokulich 2020) entails the recognition that not only theories, but also the description of empirical phenomena is underdetermined by evidence. Empirical underdetermination, understood as the underdetermination of empirical phenomena by data, emerges as a major challenge still to be fully acknowledged and carefully approached in the philosophy of science.

To face this challenge, it is essential to be able to identify the multilevel theoretical assumptions underlying the production of data models and thus the inference to empirical phenomena. Despite the many difficulties, this kind of analysis has already been attempted with some success in the case of the natural sciences (Kaiser 1991, Leonelli 2009, Karaca 2018, Bokulich & Parker 2021, Antoniou 2021), where background knowledge about instruments and empirical procedures is often explicitly available. However, the situation seems quite different in the case of the social sciences, where the opacity of instruments (Borsboom *et al.* 2009) and the

[1] Department of Philosophy, University of Valladolid (Spain).

highly conjectural nature of background assumptions, renders the challenge of empirical underdetermination more dramatic.

Keywords: underdetermination, evidence, background assumptions, multiple realizability, instrumental opacity.

1. Introduction

Theoretical underdetermination is a central issue in the philosophy of science, whose discussion has developed from the early 20th century. The so called "Duhem-Quine problem" has been used as an umbrella term to refer to several problematic features emerging from the lack of a biunivocal correspondence between theory and evidence. In what follows I am going to argue for two main claims:

- underdetermination reaches beyond the relation between theories and evidence (it also affects the relation between evidence/established empirical phenomena and data).

- this problem is bigger for social sciences due to the special difficulties to accumulate background knowledge.

The paper is structured in two main sections. In section 2, I recall the main features and kinds of theoretical underdetermination and show how empirical underdetermination parallels the theoretical one. After this (section 3), I explain why this issue is a bigger problem for the social sciences. In particular, I emphasize the difficulties in establishing background knowledge for empirical research in the social sciences. Different sources of these difficulties are acknowledged (identified) as especially relevant, i.e. the ambiguity of behavioral observations [exacerbated by the special complexity of the social background, A role for psychology], and the instrumental opacity in experimentation and survey research in the social sciences.

1. From theoretical to empirical underdetermination

In the philosophy of science, debates about underdetermination have predominantly revolved around theoretical underdetermination, which is usually understood as the possible or actual coexistence of alternative theoretical explanations given the available evidence. The underlying intuition behind the acknowledgment of this problem is that theoretical explanations of what we observe necessarily go beyond what we observe, otherwise the observed would be transparent, self-explanatory as to its causes and nature. But this is not the case, so we need to develop conjectures or theories that go beyond the observed. Now, we can do that in different ways**,** because the available set of observations (the *explanandum*), is always compatible with more than one theoretical explanation (the *explanans*). In addition, the same theoretical explanation may or may not be in agreement with the same set of observations, depending on the background assumptions made.

The discussion of this issue has a long history that I am not going to present here.[2] I am only going to focus on some central features attributed to theoretical underdetermination, to show that very similar features can be recognized in what I am going to call empirical underdetermination. My claim is that, since the empirical domain is complex and multilayered, it involves also a relation between an *explanandum* in the form of data, and an *explanans*, in the form of assumed empirical phenomena. Not only theories, but also assumed empirical phenomena (or models of data) are vulnerable to underdetermination. I am going to argue that this problem is particularly challenging in the social sciences, where the assumptions behind observations are often implicit and escape scrutiny.

1.1. <u>The standard approach to underdetermination</u>

[2] For systematic presentations of this issue see Bonk 2008, Biddle 2013, Turnbull 2017, Stanford, 2021.

The problem of underdetermination is usually referred to as "the Duhem-Quine problem", which is an ambiguous label for two different sorts of underdetermination, namely, the holistic and contrastive ones, corresponding, respectively, to the problems of confirmation holism and empirical equivalence (or confirmatory equality) of alternative theories (Stanford, 2021). In its holistic version, a theoretical hypothesis is always underdetermined by evidence in the sense that it cannot be empirically tested in isolation by a given set of observations (Duhem 1906/1991, Quine 1951, Okasha 2002, Dietrich & Honenberger 2020). A hypothesis always needs to be conjoined with background assumptions for it to be testable. But different sets of background assumptions about the world, the functioning of the instruments, etc. are in principle compatible with the same hypothesis and initial set of observations. Consequently, depending on the choice of background assumptions, the same set of observations may or may not support the hypothesis.

In its contrastive version, available evidence is never enough to determine the truth of a certain theoretical hypothesis versus another, provided that both are empirically adequate (Quine 1975, van Fraassen 1980, 1983, Sklar 1981, Bonk 2008, Godfrey-Smith 2008, Lyre 2011, 2018, Acuña & Dieks 2014, Stanford 2021). Considered together, both versions of underdetermination entail a lack of biunivocal correspondence between a given hypothetical assumption and a single set of empirical indicators providing the empirical evidence for it. For the sake of clarity, it is useful to further distinguish, from the beginning, between the traditional sense of contrastive underdetermination, which essentially involves empirical equivalence, and what has been called "transient" underdetermination" (Sklar 1975, 1981) or "practical underdetermination" (Biddle 2013, Turnbull 2017), which refers to alternative theories that–at least for a certain period of time–are equally well confirmed, but not empirically equivalent. However, since in principle any form of underdetermination can be transient in a literal sense, in *lieu* of Sklar's terminology, it seems preferable to call the second sort of contrastive underdetermination 'confirmatory equality', in contrast to 'empirical equivalence' underdetermination.

Pierre Duhem (1906/1991) convincingly argued that it is impossible to test a hypothesis in isolation, since, in order to derive empirical consequences from a hypothesis, the latter needs to be conjoined with many other

assumptions and hypotheses about the world, the functioning of measuring instruments, the environmental conditions, etc. These holistic features of confirmation lead to the acknowledgment of holistic underdetermination, since there are in principle multiple possible choices of auxiliary assumptions to be conjoined with a hypothesis. A classic example of holistic underdetermination was provided by Duhem in his 1906 work. As he points out, in testing hypotheses from thermodynamics, we need to be able to empirically determine changes in temperature by correlating the latter with changes in some other quantity. If we use a mercury thermometer to this end, we need to assume that changes in the length of the strand of mercury is what is relevant to be able to establish changes in temperature and to endorse numerous assumptions about how mercury expands or contracts as the temperature rises or falls. According to Duhem, this type of measurement depends on the assumption of certain laws of nature, like linear expansion, according to which, change in length is directly proportional to the change in temperature. Also, there are assumptions on the conditions under which a temperature reading, as given by a mercury thermometer, should be disregarded, for example, if the mercury thermometer is placed in a strong magnetic field. As it is well known, Duhem emphasizes, as an important implication of his view, that confirmation holism precludes the possibility of performing crucial experiments, thus denying that there had been a crucial experiment leading to the rejection of the particle theory of light in favor of the wave theory of light.

Willard Van Orman Quine did not only acknowledge confirmation holism (1951), but took Duhem's argument a step further and asserted that a theory can always avoid refutation by changing the auxiliary assumptions conjoined with it (Quine & Ullian 1970).[3] While accepting the very fact of confirmation holism, Popper rejected the implications drawn from the so called Duhem-Quine thesis, in particular, the idea that, when a false prediction is derived from a hypothesis conjoined with auxiliary assumptions, it is not possible to identify where the mistake lies (Popper 1963, 322-25). Against this "holistic dogma", he claimed that it is always possible to pinpoint the logical connections between hypotheses or assumptions and refuted predictions. The way to do that would be similar

[3] Two enlightening analyses of the problem of the scope of confirmation holism can be found in Ariew 1984 and Moulines 1986.

to the one applied to prove the independence of axioms in formal systems, which would involve finding out a model that satisfies all axioms but the independent one. When some refuting evidence is gathered, such evidence may provide a model that satisfies several assumptions while not the main hypothesis that happens to be conjoined with them. If so, even in non-axiomatized systems, we could identify the source of error by conjoining a different hypothesis to the same assumptions and check whether the previously refuting evidence is now a model of the new system sharing the same auxiliary assumptions with the old system. In that case, if a positive result is obtained, we have good grounds to infer that the assumptions were not the source of error in the first place, when they were conjoined with the old hypothesis. Hence, as it is argued here, the more analyzed a theoretical system is, the better for methodological purposes.[4] Interestingly, the Quinean holistic claim that a theory can always be immunized against contrary evidence connects both forms of underdetermination, for, once a theory is "immunized" in a holistic fashion –i.e. by changing some auxiliary assumptions associated to it– there is no conclusive way to discard it in favor of some rival theory, hence favoring contrastive underdetermination.

On the other hand, a famous example of contrastive underdetermination was provided by Bas van Fraassen (1980), who described a case in which we have two alternative Newtonian cosmologies with the same predictive capacity: one of them would include the Newton's laws of motion, the law of universal gravitation and the assumption that the universe, as a whole, is stationary; the second cosmology would only be different in that the assumption added to the laws is the opposite.[5] Contrastive underdetermination occurs when rival theories are all empirically adequate (Quine 1975). It happens when alternative theories are empirically highly confirmed, whether or not they are empirically equivalent (Sklar 1981). It has to do with the (transient, recurrent or

[4] Deborah Mayo (1996, 2010, 2018) has famously developed a severe testing approach to theory evaluation on the basis of Popper's ideas.

[5] A peculiar as well as controversial feature of van Fraassen's example is that empirical equivalence occurs as a result of an epistemic limitation, and, what is more, one that seems impossible to overcome. In principle, we cannot detect constant absolute motion of the universe as a whole from our position in it. We cannot even envisage how the potential empirical evidences supporting the corresponding hypothesis would look like. In this case, empirical equivalence would come hand in hand with the inclusion of empirically vacuous assumptions, which, for that reason could be regarded as superfluous theoretical components. The same problem arises in connection with van Fraassen's (1983) and Kukla's (1996, 145) use of theory-producing algorithms to generate an empirically equivalent rival theory for every theory in science.

permanent) incapability to discard a theory in favor of another on the basis of evidence. Darren Belousek (2005, 670) applies the same argument as van Fraassen's in arguing for the observational indistinguishability between rival theories in quantum mechanics, emphasizing that "every experimental test that (dis)confirms 'orthodox quantum mechanics' (dis)confirms Bohmian mechanics, and vice-versa". Jeremy Butterfield (2012) illustrates the same situation in the case of cosmological models that share identical empirical consequences.

On the other hand, as several authors have noted (Sklar 1981, Kitcher 1992, Laudan & Leplin 1991),[6] if we assume that it is always possible to gather new evidences at some point in the future, we could then consider all cases of contrastive underdetermination as transient. Yet, by the same token, this sort of underdetermination could be recurrent (Godfrey-Smith 2008, Acuña & Dieks 2014). Hence, underdetermination can always strike again, even in the form of potential unconceived alternatives −that may or may not emerge in the future (Stanford 2021).

1.2. Empirical underdetermination

The problem of empirical underdetermination has rarely been explicitly formulated, although it is implicitly acknowledged in different discussions, like those around Duhem's holistic thesis, the theory-ladenness of observation, models of data, or the incommensurability of experimental practices. The very notion of empirical underdetermination has not been fully articulated as characterizing a separate form of underdetermination. However, a remarkable step towards this articulation is Thomas Bonnin's (2021) account of "pervasive underdetermination", that he characterizes as follows:

> "I showed that the phagotrophic and syntrophic explanations of the origin of eukaryotes display entrenched disagreements about (a) the constitution and interpretation of the evidence, (b) the scaffolds

[6] Larry Laudan and Jarrett Leplin (1991, 451–455, 461–465) provide a well-known critique of arguments for empirical equivalence underdetermination. They point out that the accumulation of scientific knowledge over time results in some theories often gaining evidential support over others. Further, they argue that empirical consequences are not the only source of evidential support for theories.

> upon which these explanations are built and (c) the interpretation of background theories in the constitution and evaluation of hypotheses. In the case discussed, it is not that the choice between hypotheses is underdetermined by the available data, but rather that (a) the constitution and interpretation of the evidence is underdetermined; (b) the choice of theories used as scaffolds is underdetermined and (c) the correct interpretation of background theories is underdetermined. Underdetermination reaches deeper" (*ibid*., 146).

Talking about empirical underdetermination only makes sense if the multilevel nature of the empirical domain is acknowledged. Just as theoretical underdetermination presupposes the distinction between the empirical and the theoretical level, empirical underdetermination presupposes a distinction, within the empirical level, between data and empirical phenomena. It is important to distinguish one form of underdetermination from the other, since they occur at different levels and have different epistemological and methodological implications. We could certainly conflate both issues and just talk about 'theoretical underdetermination' to refer, in general, to cases where the same evidence (whether data or empirical phenomena) can be explained on the basis of alternative theories (whether low-level or high-level theories). An all-encompassing/embracing notion of underdetermination (one including both the holistic and the contrastive versions) can be characterized by a non-biunivocal relation between observation and theory, such that: the same set of observations can be explained by alternative theories, and depending on the choice of background assumptions, the same theory may or may not be confirmed/refuted by the same set of observations.

However, this way of speaking could lead to misidentifying cases of empirical underdetermination, in which the same data can alternatively be understood as being originated by alternative empirical phenomena, as cases of theoretical underdetermination, where the same empirical phenomena can be interpreted as caused by alternative theoretically postulated phenomena.

To use a classical example: we can distinguish between, on the one hand, the transient theoretical underdetermination affecting the Ptolemaic

geocentric model and the Copernican heliocentric model, given the shared observations at the time where both were rival theoretical models, and, on the other hand, the (transient) empirical underdetermination affecting the determination of parallax, since the lack of observed parallax at the time was considered a refutation of the Copernican model by its opponents, but not so by its advocates. In this case, the same empirical phenomenon, i.e. the lack of parallax, was considered to be the case or not to be the case on the basis of the same data about the positions of the fixed stars relative to the observer. Because of the uncertainty about how distance would affect parallax, the empirical phenomenon of parallax was underdetermined with respect to the available observations. In other words, the acceptance of the empirical phenomenon of parallax was underdetermined with respect to the available observations. This example is a case of holistic empirical underdetermination. But empirical underdetermination, like theoretical underdetermination, may also occur in a contrastive form: a) different bodies of data support different inferences to different empirical phenomena, b) the same body of data supports different inferences to different empirical phenomena. In both cases, different epistemological, interpretative strategies are being followed, although, in the first case, the differences also affect the data being produced.

1.2.1. Clarifications on the empirical/theoretical distinction and on background assumptions

Before digging further into the problem of empirical underdetermination, let us make an important to clarification about the empirical/theoretical distinction. In a minimal characterization of the distinction, two main features have special relevance:

1) functional role relative to a theory;

2) epistemological status relative to a theory.

1) refers to the role played by one side/component/dimension of inquiry with respect to the other, a role which is determined by the relation between the explanans and the explanandum. This is the view advocated

by Horwich and Ben-Menahem, who support the idea that the theory/observation distinction is relative (Okasha 2002, 316).

2) concerns the stronger or weaker conjectural character depending on the role played, a character that depends on the higher or lower conjectural nature of one component of inquiry in comparison to the other.

We usually expect both features to be aligned, i.e. that the explanandum is less conjectural than the explanans. It is important to note that the data/(empirical) phenomenon distinction shares the features 1), 2). In other words, data are the observations (evidence) supporting beliefs about empirical phenomena (Suppes 1969), Bogen & Woodward 1988, Woodward 1989, 2000, 2010).

It is also important to introduce a clarification regarding two sorts of background assumptions, depending on their relation with the primary theory. Some background assumptions are embedded as part of the primary theory, while some others are not. In the first case, background assumptions are not a variable component with respect to the primary theory, whereas, in the second case, they are so.[7] This distinction, therefore, is important in analyzing holistic underdetermination, for only in the second case the background assumptions involved in testing the primary theory could be replaced by others leading to different testing outcomes for the same primary theory. By contrast, if background assumptions changed in the first case, this would entail a change in the empirical meaning of the primary theory, as those assumptions are embedded as part of the empirical meaning of the theory. For instance, if we grant that background assumptions about the conditions for the observability of parallax are (or not) embedded in the Copernican System, then the lack of observed parallax at the time could be accommodated or not in the primary theory depending on the choice of background assumptions regarding observability conditions for parallax. In the Phlogiston Theory, the laws of classical mechanics where embedded in the theory and assumed as a background theory in measuring changes of weights of different portions of substances before and after combustion. By contrast, background assumptions on initial conditions, such as those

[7] A similar distinction is drawn in Balashov's (1994) "string" model of scientific tests.

involved in estimating the effect of heat in the apparatus, where not embedded in the theory, and depending on the choice of such assumptions, anomalous results could or could not be reconcile with the primary theory.

1.2.2. The parallelism between theoretical and empirical underdetermination

Theoretical and empirical underdetermination alike may come in a local or global manner. In the first case, the same evidence constitutes a shared *explanandum,* for which different (high-level/low-level) explanations are provided. In the second case, different bodies of evidence constitute non-shared *explananda,* for which different (high-level/low-level) explanations are developed.[8]

Both theoretical and empirical underdetermination are originated by the same basic problems emerging at different levels. These problems have to do with the ampliative nature of explanatory inference and the different choices of background assumptions involved in such inference. On the theoretical level, the inference from observations to theoretical (postulated/conjectured) phenomena explaining the observed is not univocal. On the empirical level, the inference from the data to empirical phenomena explaining such data is not univocal neither. Different empirical phenomena can explain the same data (contrastive underdetermination) and assumptions about the occurrence of certain empirical phenomenon may or may not be supported by a given set/collection of data depending on the choice of background assumptions playing a role in the explanation of the data on the basis of the assumed

[8] The second case has been usually discussed in connection to the issue of methodological incommensurability. Both Kuhn (1962/1970) and Feyerabend (1970) draw attention to extreme cases of incommensurability where alternative (rival) theories are supported on the basis of alternative (rival) evidences. have been exemplified : metallizing principle, mental illness
Extreme cases of incommensurability have also been acknowledged from other philosophical standpoints, quite far from the Kuhnian tradition. Hacking's literal incommensurability (1983, 1992), as well as Pickering's *machinic* incommensurability (1984, 1995), claim that, in laboratory sciences (like particle physics), alternative theories are truth of different kinds of phenomena, which are "created" in the laboratory with different kinds of instruments. In their view, rival theories are applied and justified through disjoint sets of measurement procedures. In addition, the creation of the instrumentation would be very often subordinated to theoretical interests.

empirical phenomenon (holistic underdetermination). To put it shortly, the main problem arises from the fact that both inferences are ampliative, i.e. both empirical and theoretical phenomena are postulated to some degree, both are explanatory (although at different levels). In addition, there is (holistic) underdetermination whenever there is a need for background assumptions, since choice of assumptions is always multiple. If we acknowledge that background theories are involved in the identification of phenomena from a given set of data, then phenomena are holistically underdetermined by data.[9]

2. A bigger problem for the social sciences?

In contrast to the hierarchical nature of evidential support characteristic of natural science (where evidence is built from more directly measurable attributes to less directly measurable ones), evidential support in social science typically lacks a clear empirical hierarchy. This leaves too much room, not only for alternative highly conjectural theoretical constructs, but also for alternative lower level assumptions about empirical phenomena.

These considerations do not imply that empirical correlations, like the ones pointing to behavioral patterns, cannot be successfully established in social science; they mean only that the most interesting empirical phenomena—i.e. those ultimately concerning intentional dispositions and psychological attributes—cannot be as firmly empirically grounded as they are in natural science.

In experimental economics, the theoretical underdetermination of psychological constructs tends to make the interpretation of results doubtful, and even to hamper the test of auxiliary assumptions. Since behavioral choice data are ambiguous, they need to be supplemented and/or combined with non-choice data provided by psychological studies (Ross 2010; Schotter 2008). In order for economists to satisfy their current inclination to dig further into social motivations, there is no other way but

[9] By the same token, if background theories are involved in the determination of processed data, processed data are underdetermined by less processed data. This would lead us to an even wider notion of underdetermination that, despite its relevance, it is not going to be developed here

to examine psychological constructs such as justice, trust, reason, desire, belief, among others (Dietrich and List 2012; Tyler and Amodio 2015). However, the empirical underdetermination of psychological constructs tends to make the interpretation of results doubtful, and even to hamper the test of background assumptions.

Furthermore, the verification of precepts like that of dominance requires the determination of psychological constructs, in this case, related to the reward structure capable of offsetting the subjective costs or values involved in individual decisions (Søberg 2005, Cordeiro-dos-Santos 2006).[10] In economic experiments, psychological constructs prove also of the highest relevance to examine, for instance, central auxiliary assumptions concerning the environment, particularly those about the agents' characteristics, like beliefs, expectations or moods, and their preferences.

All this points to the problem of background knowledge in the social sciences and its relevance for avoiding empirical underdetermination, both in the observation of behavior and in the devise and use of experimental and survey methods.

2.1. A more serious lack of background knowledge in the social sciences?

Accumulation of background knowledge does not only happen in natural science, but it occurs also in the social sciences (for instance, with regard to biases, learning effects, fatigue effects...). However, such accumulation seems very scattered in comparison to how it looks in natural science (think about core elements of chemistry, optics or electrical engineering). As a result, this accumulation of such scattered knowledge seems insufficient to build a layer of well-established empirical phenomena in the form of background knowledge, which could be the evidential basis for higher level hypotheses. The problem of background knowledge in the

[10] Søberg (2005) and Cordeiro-dos-Santos (2006) independently formulate a set of auxiliary assumptions respectively based on Smith's (1982) distinction between three ingredients of a lab experiment (environment, institution, design) and his distinction between four precepts in economic experimentation (non-satiation, saliency, dominance, privacy).

social sciences is underlying the issue of empirical underdetermination, both in the observation of behavior and in the devise and use of experimental and survey methods.

Many behavioral economists seem to agree with Matthew Rabin's view that 'economists should aspire to make our assumptions about humans as psychologically realistic as possible' (Rabin 2002: 657). Obviously, psychology would be necessary for these purposes. But, although this idea is indeed worth pursuing, it seems unclear that psychology can avoid the very same problems affecting social sciences like economics. As Francesco Guala points out:

> "This programme is based on a scientific gamble: that economic theory (and choice theory in particular) can be substantially improved by studying the psychology of human decision. The gamble will pay off if there are indeed robust and systematic mechanisms that help understand and predict behavior more accurately than the standard theory does" (Guala 2019, 389).

A consequence of the lack of background knowledge in social science is that, even if researchers can still distinguish between the empirical and the theoretical on functional grounds, the difference in function does not seem to go together −or to be aligned− with an epistemological difference in the degree of conjectural character exhibited by one level or the other. A reason for this is what Rosenberg has labeled the "hermeneutical circle" of beliefs, desires, and actions, in which coherence among the three variables is the criterion of explanatory adequacy' (Rosenberg 2012, 50). He notes that, in explaining action, we inevitably face a threesome of belief, desire, and action, where we would need to fix two of them in order to determine the third. According to Rosenberg, folk psychology may be summarized in the following general statement that he labels [L]:

> "If any person, agent, individual, wants some outcome, d, and believes that an action, a, is a mean to attain d under the circumstances, then x does a" (Rosenberg 2012, 39).

He notes, however, that, in order to determine the initial conditions of an action, it is always necessary to take into account more desires and beliefs

than the ones referred in [L].[11] [L] proves less amenable than most physical models to empirical correction and subsequent refinement by way of improving the measurement of the initial conditions. In measuring initial beliefs and desires, social scientists confront 'a regress problem' that is hardly instantiated in natural science: they have to resort to the very model that is being tested, that is, they need to rely on [L] to be able to measure the beliefs and desires involved in the initial conditions. In short, psychological features like beliefs often play an empirical role in the social science research, despite the fact that they are not an observable kind of thing, nor something amenable to (strict) measurement; and possess neither a simple univocal origin, nor a simple, univocal observable manifestation in behavior.

The "hermeneutical circle" problem affects the main sources of evidence in the social sciences, which are the observation of behavior and the data obtained through experimental or survey research. But, beyond and above hermeneutical circle, the problematic nature of evidence in the social sciences is ultimately related to the combination of two hampering factors: too much background and too little background knowledge. The initial conditions of an observed behavior are extremely complex, and therefore, they are difficult to determine. Without knowledge of the initial conditions restricting the interpretation of overt behavior, the latter becomes highly ambiguous. The same lack of knowledge drastically severs the possibility of identifying confounds in experimental and survey contexts, overshadowing the way experimental and survey methods function. The rest of the paper is devoted to explain how empirical underdetermination emerges from the ambiguity of overt behavior and the opacity of experimental and survey methods for empirical research. Although most of the time I will be referring to the "too little background knowledge problem", acknowledgement of the "too much background problem" is presupposed.

2.2. The ambiguity of behavior: the problem of "overlapping multiple realizability"

[11] For a similar idea see Guala 2019, 390.

The non-uniform behavioral exercise of the same mental dispositions had been already noticed by Gilbert Ryle in his path-breaking work, *The Concept of Mind*:

> "Now the higher grade dispositions of people with which this inquiry is largely concerned are, in general, not single-track dispositions, but dispositions the exercises of which are indefinitely-heterogeneous" (Ryle 1949/2009, 32).

Ryle's point is often stated in terms of the multiple realizability of metal properties or states in different courses of action. Being angry or being intelligent can be realized in different behavioral tendencies depending on the particular individual and context. From Ryle's approach it is possible to argue that the ambiguity of overt behavior is due to the phenomenon of "overlapping" multiple realizability. My way of being angry may sometimes be behaviorally indistinguishable from your way of being sad, your preference for historical novels may be behaviorally indistinguishable from my desire to impress my friends by collecting historical novels.

In considering the problem of ambiguity with regard to the observation of behavior, I am going to place the focus on data about choices and data about preferences. Choices are a kind of action or behavior, usually employed to empirically check hypothesis about preferences, which are established as theoretical constructs. In this kind of context, preferences are invoked as theoretical constructs referring to overt behavior, that is, to observable and measurable patterns of choice. But even the more basic description of the choice behavior, presupposes that the options are understood in a certain way by the subject. This, in turn, can only be established by making assumptions about the beliefs of those subjects. For one thing, the description of the choice behavior presupposes a certain conceptualization or description of the options, hence in describing a subjects' choice we assume the conceptualization of the options to be shared by the subject and us. But, of course, the adequacy of this assumption depends on whether it is plausible to assume that we share the relevant beliefs about the options. In choosing between two different brands of oranges, this may not be a big deal, but in choosing between two legal procedures, financial products, educational systems, etc… It is not that clear that we share the relevant beliefs. And, again, to

check those beliefs is not easier than checking the primary theory we want to check by gathering observations about choices.

But, even if we did share the relevant beliefs about options, empirical underdetermination would reemerge at more sophisticated levels of inquiry. Very often, theories do not invoke preferences to explain (patterns of) choices, but they invoke different bias, heuristics, etc. to explain (patterns of) preferences. Here, preferences constitute the empirical domain, for instance, with respect to prospect theory (Guala 2019, 392-4). Now, preferences are even more affected by empirical underdetermination than choices are. Because evidence for preferences inherits the underdetermination of the description of choice behavior by the observed selection of an option, and, in addition, is vulnerable to the variability in the manifestation of preferences due to different beliefs about the implications of choosing one option over the other. Even if I prefer anonymous voting over non-anonymous voting, I may (or not) choose the second in a panel decision, just to show (or not) that I have nothing to hide. In this case, my pattern of choice would be deceiving (ambiguous) with respect to my pattern of preference. So, the pattern of preference would not be a well-established empirical regularity, thus providing a misleading (an indeterminate) *explanandum*. In other words, the empirical basis for theorization would be misconstrued (underdetermined).

Ultimately, social science faces a pervading twofold problem with respect to observation: multiple realizability of the same phenomena (even in the case of what are regarded as basic empirical phenomena), and equal manifestation of different phenomena. The combination of these two sides of the problem result in what we could call the issue of "overlapping multiple realizability". This happens when two different multiple realizable phenomena converge (overlap) in the some of their realizations or occurrences. Let us think of preferences, again, as theoretical constructs. As recently emphasized by Guala, they are multiply realizable:

> "Preferences in the economic theory of choice are dispositions that can be realized in different ways depending on the circumstances of choice and on the characteristics of the decision-maker" (Guala 2019, 384).

As he points out, if we take for granted that consumers do not have perfect information about the objects of choice, there cannot be a one-to-one correspondence between preference and choice. He provides the following example:

> "Consider the following two cases:
>
> (a) Tony prefers the restaurant Pizza Vesuvio to Pizza Bella Napoli. He believes that Pizza Vesuvio is closed tonight. Therefore, he goes to Pizza Bella Napoli
>
> (b) Vincent prefers Pizza Bella Napoli to Pizza Vesuvio. He believes that both are open. Therefore, he goes to Pizza Bella Napoli.
>
> Although their choices are identical, it would be a mistake to conclude that Tony and Vince have the same preferences. The reason is that, in the standard theory of economic choice, behavior is determined both by preferences and by beliefs. As a consequence, the same choice (behavior) may result from different preferences, if beliefs also differ." (Guala 2019, 385-386).

This case is one of theoretical underdetermination if preferences are taken as part of the explanans (as a theoretical construct), but it would be a case of empirical underdetermination if preferences are taken as empirical evidence to be explained on the basis of different theoretical constructs. For instance, preferences for Pizza Vesuvio (established on the basis of patterns of choice) are taking as evidence for the hypothesis that people prefer traditional restaurants over more modern ones.

Guala explains this twofold role of preferences as follows:

> "We can thus distinguish between two explanatory tasks: at one level, economists are interested in providing explanations of behavior; at another level, many (but not all) economists believe that improving such explanations requires that we are also able to explain preferences. Another way to put it is to say that preferences are part of the explanans (what does the explaining) at the former level, but constitute the explanandum (what is to be explained) at the latter" (Guala 2019, 391).

Guala's argument amounts to distinguishing two levels of explanation and two corresponding levels on the empirical side. For instance, while choice

theory explain behavior by invoking preferences, behavioral economics explain the shape of certain preferences in the models of behavioral choice theory by invoking psychological mechanisms. The two levels are shown in table 1:

Explanans	preferences (choice theory)	biases, heuristics… (prospect theory)
Explanandum	choices (observation of behavior)	preferences (in risky choice experiments)

Table 1: two levels of *explanantia* and *explananda*

In order for economics to provide explanations of preferences, it needs to embrace psychological explanations, since preferences themselves are important psychological traits targeted by psychological theories.[12] An example of a theory where preferences constitute the *explanandum* and, hence, play an empirical role is prospect theory. According to this theory, the causal basis for preferences exhibited in risky-choice decisions are biases like loss aversion and reference dependence in the evaluation of outcomes (Guala 2019, 393).[13]

If we acknowledge the obvious fact that preferences are belief-dependent dispositions, then this inevitably leads us to the problem that beliefs are unobservable states included in the initial conditions (Guala 2019, 390). The inclusion of internal, unobservable and highly conjectural states in the initial conditions is a common fact in empirical "observations" in the social sciences. Rosenberg's hermeneutical circle emerges as a recurrent issue leading to an ambiguous, underdetermined empirical basis for research in this field.

[12] "Like most economists, I will take psychology to be engaged in the explanation of human preferences, rather than in the explanation of behavior by means of preferences (which is the specific task of choice theory). Moreover, I will argue that no psychological explanation of preferences is likely to cover all the domains in which choice theory is applied" (Guala 2019, 388).

[13] "Prospect Theory uses psychological information to attain a more accurate representation of preferences, whereas traditional choice theory tends to sacrifice realism in favour of normative considerations (its agents are rational decision-makers, first and foremost).What is the causal basis of preferences according to Prospect Theory? Terms like 'loss aversion' and 'reference dependence' refer to systematic biases in the evaluation of outcomes" (Guala 2019, 393).

2.3. Instrumental opacity in experimentation and survey research.

Empirical underdetermination in the social sciences is also connected to the problem of instrumental opacity, which challenges both experimental and test validity. A strong voice on this issue is Borsboom *et al*. (2009, 135–170), who has shown how, as long as the measurement 'instruments' in social science remain opaque, the same empirical data can be accommodated into many alternative interpretative frameworks, both theoretical and empirical. A measurement instrument for an attribute must have the property of being sensitive to differences in the attribute; that is, differences in the attribute must be the main cause of the differences in the outcomes generated by the measurement procedure. As emphasized by Borsboom *et al*., this has a straightforward methodological implication: scientists must have enough knowledge of the causal chain or network involved in the working of the measurement procedure (Borsboom et al. 2009, 148).

To use their own example, if the functioning of mechanical weight scales were unknown, results obtained from the scales could be easily interpreted as height measurements instead of weight measurements, given that both properties, weight and height, are significantly correlated in humans (Borsboom *et al*. 2009, 156–157). By virtue of this correlation, measurement with a measure stick would validate the interpretation of measurements with scales in terms of height. Ignorance about how the measurement instrument works leads to pervasive, empirical underdetermination. Conversely, knowledge of how the instruments work enable us to gather meaningful (non-ambiguous) empirical information, which can be useful as evidence.

In general philosophy of science, the hierarchy of models approach advocated by Patrick Suppes (1962) and later refined by Deborah Mayo (1996), constitutes an explicit defense of the need to identify the theories presupposed in experimental practice. Different theories would be assumed at different levels of inquiry, and each of them would determine models at the corresponding levels (from top to bottom): theoretical, models of experiments, models of data, experimental design and *ceteris paribus* conditions. The emphasis on the presupposed theories is not meant to imply that there is any ultimate, unifying empirical basis for science, since both Suppes and Mayo are well aware of the dynamical

nature of the empirical basis and the fragmentary, disjointed character of many areas of natural science. What is important to stress here is, rather, that the hierarchical structure and support of empirical phenomena in natural science has not analogue in some areas of social science, where hierarchies of models underlying (and supporting) empirical or evidential claims can seldom be identified as present to the same degree. If concepts like those of ether and phlogiston have been dismissed is because certain empirical properties (interference with the speed of light and weight loss in combustion, respectively) were unequivocally attributed to them and, in addition to this, the experimental determination of these properties was unequivocal too, partly by virtue of the well-established character of (optical, mechanical) laws presupposed in the use of the corresponding experimental instruments (interferometer, balance).[14]

Let us consider, por example, the case of opacity in public goods experiments. In these experiments, each subject chooses how much to contribute to a common pool which returns benefits to all participants equally. The typical results show initial high levels of contribution that decrease over time. In order to explain this typical tendency, several alternative hypotheses have been developed (Andreoni 1988, 1989, 1990, Levine 1998, Santos 2009), giving rise to a clear case of theoretical underdetermination. The same results can be explained by appealing to competing free riding hypothesis (one assuming strategic play, the other assuming a learning effect), although impure altruism, rule of thumb, and regret effect could also be invoked to explain the same results. Making significant progress in testing these different hypotheses would require the empirical determination of some of the conditions that are assumed to hold in the public good experiments. For example, it would be very relevant to be able to establish how the subjects understand the different parts of the game (for instance, the final round). The lack of background

[14] There is certainly a dispute on whether theoretical knowledge of how the instruments work is indeed needed or rather some low level practical knowledge about such working is enough to obtain meaningful empirical information. Ian Hacking (1983), Allan Franklin (1986), and Jim Woodward (1989, 2000) are some of the authors who have questioned the need for theoretical knowledge of the instruments, vindicating instead the role of practical knowledge. Hacking's well-known example of the light microscope is intended to show how, by manipulating the specimen in known ways and observing corresponding changes in the image, we can establish that the instrument suitable to determine empirical features of the target phenomenon. Elucidating these issues goes beyond scope of this paper, but what matters for present purposes is that some knowledge about the instruments, whether theoretical or practical, is needed to avoid empirical underdetermination and hence obtain meaningful empirical knowledge.

knowledge on this matter is difficult to avoid. Just asking the subjects is always an option, but introspection may not help them become aware of how they interpret a highly artificial environment such as the experimental situation. Again, subjects' beliefs remain both empirically relevant and underdetermined. This does not preclude the increase of background knowledge on other important aspects of public-good experiments. For example, it has been successfully established that there is a group-size effect in public goods experiments, but the opposite to the one expected, i.e., the bigger the group, the higher the percentage of subjects contributing (Lipford 1995).

Opacity also affects survey research in social sciences. By appealing to the testimony of subjects, tools such as questionnaires and interviews provide a way to face or counterweight the ambiguity of experimentally observed behavior. However, linguistic intervention and testimony may be also highly misleading in so far as knowledge about the instruments (i.e. the workings of the surveys) is missing, or false beliefs about the functioning of the instrument are commonly adopted. Social science constructs are very often validated on the basis of survey generated data, whose opacity in terms of validity has been noted by several authors (Boulier & Goldfarb, 1998, 5–6, Meyer, Mok, and Sullivan 2015, Caamaño & Caamaño 2019).

Attempts to convey controlled pieces of information through the use of surveys may face two main obstacles:

1. the incompleteness of the information conveyed to/from the subject.
2. the unintended character of the information conveyed to/from the subject.

These obstacles may be due to the simultaneous manipulation of several independent variables, to different contextual factors involved in communication, or even to discrepancies between the experimenter and the subject regarding how to understand the survey questions. In short, it is highly difficult to control the communication process so as to guarantee that the information which the experimenter intends to convey constitutes the cause of the subject's response, and the other way around. The problem is not only that there is "too much background", nor even "too much possible background", like in cases where the set of causally relevant variables potentially affecting the target phenomenon is not only

numerous and heterogenous but highly variable in what it includes. All this happens also in natural science, in areas like meteorology or even in the study of ordinary mechanical phenomena like the trajectory of the golf ball after the stroke. The ultimate difference in the case of social science is related to the unknown or highly conjectural nature of those background factors, which is what leads to a severe form of empirical underdetermination.

Once again, progress has been made in the recognition and knowledge of contextual factors affecting the use of surveys. But the background knowledge so obtained is too limited and scattered. For instance, context effects, like those found in attitude measurement, have been widely studied in questionnaire research. A large body of evidence shows that many survey respondents strongly react to different contextual features, i.e. to features other than the questions' real core or essence, like surrounding circumstances, question order, and contextual wording, as well as the order of the response options and response scales belonging to previous question (Billiet, Waterplas, and Loosveldt 1992, 131, Smyth, Dillman, Christian 2009). A broad notion of context effects embraces all influences on question answers that are due to information passed on to the respondent from the survey environment (Smyth, Dillman, and Christian 2009). A more restricted notion specifies a subset of response effects that Krosnick and Presser (2010) label as semantic order effects. These effects result from the location of a question in a sequence of meanings. What matters here is that context effects are caused by the unobserved interaction between subjects' stored beliefs and triggers generated by the survey instrument (Morgan and Poppe 2015). The choice among response options results from the interplay between personal attributes developed over the subject's complete past, social environment, survey features, and interviewer's introductory expressions (Groves and Singer 2004; Groves, Singer, and Corning 2000). The response choice depends on the particular mix of ideas accessible by memory, some of which are made salient by the questionnaire itself and the recent events experienced by the subject (Zaller and Feldman 1992).

Hence, despite the progress made in background knowledge related to the use of surveys −like in the case of framing effects (Gamliel & Kreiner 2013, 2019, Caamaño 2021) −there is still a large gap in background knowledge concerning contextual features.

All in all, in so far as knowledge of instruments and contextual factors is not good enough to enable a univocal interpretation of survey answers, their value to provide empirical evidence of the genuine opinions of respondents is into question.

3. Conclusions

It is important to acknowledge empirical underdetermination as an issue separate from theoretical underdetermination, occurring at a different level of inquiry. Although underdetermination at both levels has some features in common in terms of the reasons why it occurs −ampliative reasoning, choice of background assumptions −and the forms in which it occurs−holistic, contrastive−, the conflation of empirical and theoretical underdetermination could make us oblivious to the problematic, conjectural nature of evidence. The inference from data to empirical phenomena - which is always a prerequisite for the inference from empirical phenomena to theory - is worthy of separate consideration.

The relative, functional nature of the theoretical/empirical distinction must be recognized and distinguished from the epistemological dimension of the distinction in order to characterize both types of underdetermination at each level. It is assumed that both the functional and epistemological sides of the distinction are aligned. That is, the explanandum (a collection of data or an empirical phenomenon) is assumed to be less conjectural than the explanans (the description of an empirical phenomenon or a theory). While this is usually the case in natural science, the situation is quite different for social science, which is more heavily affected by empirical underdetermination.

It is not just the fact that there is more background, but also more conjectural knowledge about the background, that makes empirical underdetermination a bigger problem for the social sciences. For example, psychological properties such as beliefs or memories not only add complexity to the background of social phenomena, but also lead to more partial and conjectural background knowledge insofar as it depends on knowing psychological properties. There are two main problems underlying empirical underdetermination in the social sciences: the ambiguity of behavior, whose observation is affected by the overlapping

multiple realizability of social phenomena, and the opacity of methods for empirical research in the social sciences.

Despite these difficulties, by focusing the search for background knowledge on conditions that are particularly relevant to social research, the challenge of empirical underdetermination can be met. The challenge of empirical underdetermination can also be met by a focus on the search for background knowledge which is particularly relevant to social research. Especially in the social sciences, evaluating theories before evaluating evidence would be like putting the cart before the horse.

Acknowledgements

I am thankful to the participants in the 17th edition of the International Congress on Logic, Methodology and Philosophy of Science and Technology (Buenos Aires, August 2023), where I had the chance to get a valuable feedback on different aspects of this work. Many thanks are due to Agnes Bolinska for very illuminating insights on a previous version of this work. This research was financially supported by the research projects The Philosophy of Statistical Modeling Science, (PID2021-126416NB-I00, Spanish Ministry of Science and Innovation), and Laws and Models in Physical, Chemical, Biological, and Social Sciences (PICT– 2018–03454, ANPCyT, Argentina).

References

Andreoni, J. 1988. “Why Free Ride? Strategies and Learning in Public Goods Experiments”, *Journal of Public Economics* 37: 291-304.

Andreoni, James (1989). "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence". *Journal of Political Economy*. **97** (6): 1447–1458. doi:10.1086/261662.

Andreoni, J. 1990. “Impure Altruism and Donations to Public Goods: A Theory of Warm-glow Giving.” *Economic Journal* 100: 464–477.

Balashov, Y. (1994). “Duhem, Quine, and the multiplicity of scientific tests”. *Philosophy of Science* 61 (4):608-628.

Biddle, J. (2013), “State of the field: Transient underdetermination and values in science”, *Studies in History and Philosophy of Science*, 44: 124–133.

Billiet, J. B., L. Waterplas, and G. Loosveldt. 1992. “Context Effects as Substantive Data in Social Surveys.” In *Context Effects in Social and Psychological Research*, edited by N. Schwarz and S. Sudman, 131-147. Berlin: Springer.

Bogen, J.,and J.Woodward.1988.“Saving the Phenomena.” *Philosophical Review* 97:303– 52.

Bokulich, Alisa (2020). “Towards a Taxonomy of the Model-Ladenness of Data”. *Philosophy of Science, 87*(5), 793-806. doi:10.1086/710516 .

Bokulich A, Parker W. (2021) Data models, representation and adequacy-for-purpose. Eur J Philos Sci. 2021;11(1):31. doi: 10.1007/s13194-020-00345-2. Epub 2021 Jan 29. PMID: 33584875; PMCID: PMC7846529.

Bonk, Thomas, (2008) *Underdetermination. An Essay on Evidence and the Limits of Natural Knowledge*, Dordrecht: Springer.

Bonnin, T. (2021) “Monist and Pluralist Approaches on Underdetermination: A Case Study in Evolutionary Microbiology”, *Journal for General Philosophy of Science*, 52:135-155.

Borsboom, D., A. O. J. Cramer, R. A. Kievit, A. Z. Scholten, and S. Franic. (2009). “The End of Construct Validity.” In *The Concept of Validity: Revisions, New Directions, and Applications*, edited by R. W. Lissitz, 135–170. Charlotte, NC: Information Age Publishing.

Boulier, B. L. & *Goldfarb, R. S.* (1998). “On the use and nonuse of surveys in economics”, *Journal of Economic Methodology*, 5 (1): 1-21.

Caamaño-Alegre, María & Caamaño-Alegre, José (2019) “From Ontological Traits to Validity Challenges in Social Science: The Cases of Economic Experiments and Research Questionnaires”, *International Studies in the Philosophy of Science*, 32:2, 101
127, DOI: 10.1080/02698595.2019.1682773

Caamaño-Alegre, María (2021). “On glasses half full or half empty: understanding framing effects in terms of default implicatures”. *Synthese* 199 (3-4): 11133-11159. DOI: 10.1007/s11229-021-03282-6

Dietrich, M., Honenberger, P. (2020) “Duhem's Problem Revisited: Logical versus Epistemic Formulations and Solutions”, *Synthese*, 197: 337–354.

Feyerabend, P. K., (1970). “Against Method: Outline of an Anarchistic Theory of Knowledge”. In M. Radner and S. Winokur (ed.), *Analysis of Theories and Methods of Physics and Psychology*, (Minnesota Studies in the Philosophy of Science, Volume IV), Minneapolis: University of Minneapolis Press, pp. 17–130.

Franklin, Allen (1986), *The Neglect of Experiment*. Cambridge: Cambridge University Press.

Gamliel, E., & Kreiner, H. (2013). Is a picture worth a thousand words? The interaction of visual display and attribute representation in attenuating framing bias. *Judgment and Decision Making, 8*(4), 482–491.

Gamliel, E., & Kreiner, H. (2019). Applying fuzzy-trace theory to attribute-framing bias: Gist and verbatim representations of quantitative information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000741

Godfrey-Smith, P. (2008). “Recurrent transient underdetermination and the glass half full”. *Philosophical Studies* 137 (1):141 - 148.

Groves, R. M., and E. Singer. 2004. “Survey Methodology.” In *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, edited by J. S. House, F. T. Juster, R. L. Kahn, H. Schuman and E. Singer, 21-64. Ann Arbor, MI: University of Michigan Press.

Groves, R. M., E. Singer, and A. Corning. 2000. “Leverage-saliency theory of survey participation: Description and an illustration.” *Public Opinion Quarterly* 64: 299-308.

Guala, F. (2019). “Preferences: Neither behavioral nor mental”. *Economics & Philosophy, 35*(3), 383-401. doi:10.1017/S0266267118000512

Hacking, I. (1983). *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, I. (1992). “The Self-Vindication of the Laboratory Sciences”, *Science as Practice and Culture*, A. Pickering (ed.). Chicago, University of Chicago Press:29–64.

Hansson, Sven Ove and Till Grüne-Yanoff, "Preferences", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/preferences/>.

Hirokazu, T. (2019). “Laboratory Experiments in Game Theory”. In: Kawagoe, T., Takizawa, H. (eds) Diversity of Experimental Methods in Economics. Springer, Singapore. https://doi.org/10.1007/978-981-13-6065-7_2

Jiménez-Buedo, M. (2015). “The Last Dictator Game? Dominance, Reactivity, and the Methodological Artefact in Experimental Economics”, *International Studies in the Philosophy of Science*, 29:3, 295-310, DOI: 10.1080/02698595.2015.1179042

Kaiser, Matthias (1991). “From rocks to graphs — the shaping of phenomena”. *Synthese* 89 (1):111 - 133.

Karaca, K. (2018). “Lessons from the Large Hadron Collider for model-based experimentation: the concept of a model of data acquisition and the scope of the hierarchy of models”. *Synthese, 195*(12), 5431–5452.

Krosnick, J. A., and S. Presser. 2010. “Question and Questionnaire Design.” In *Handbook of Survey Research*, 2nd ed., edited by P. Marsden and J. Wright, 263-313. Bingley, UK: Emerald Group Publishing, Ltd.

Kuhn, T. S., (1962-1970). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Leonelli 2009. “On the Locality of Data and Claims about Phenomena.” *Philosophy of Science* 76 (Proceedings): 737–49.

Leonelli (2015) “What count as scientific data? A relational framework” *Philosophy of Science*, Vol. 82, No. 5 (December 2015), pp. 810-821

Leonelli (2019) “What distinguishes data from models?” *European Journal for Philosophy of Science* (2019) 9 (22). https://doi.org/10.1007/s13194-018-0246-0

Levine, D.K. (1998). “Modeling Altruism and Spitefulness in Experiments”. *Review of Economic Dynamics, 1*: 593-622.

Lipford, J. W. (1995) “Group Size and the Free-Rider Hypothesis: An Examination of New Evidence from Churches.” *Public Choice*, vol. 83, no. ¾: 291–303.

Massimi, M. (2007). “Saving Unobservable Phenomena”. *British Journal for the Philosophy of Science* 58 (2):235-262.

McAllister, J. W. (1997). Phenomena and patterns in data sets. *Erkenntnis, 47*, 217–228.

McAllister, J. W. (2011) “What do patterns in empirical data tell us about the structure of the world?” *Synthese*, 182:73–87

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives*, 29 (4): 199-226**.**

Morgan, S. L., and E. S. T. Poppe. 2015. “A Design and a Model for Investigating the Heterogeneity of Context Effects in Public Opinion Surveys.” *Sociological Methodology* 45 (1): 184-222.

Norton, J. D. (2008). “Must Evidence Underdetermine Theory?” In *The Challenge of the Social and the Pressure of Practice: Science and Values Revisited*, ed. M. Carrier, D. Howard, and J. Kourany, Pittsburgh: University of Pittsburgh Press: 17-44.

Okasha, S. (2002). “Underdetermination, holism and the theory/data distinction”. *The Philosophical Quarterly, 52*, 303–319.

Okasha, S. (2016). On the Interpretation of Decision Theory. *Economics and Philosophy*, 32(3), 409-433. https://doi.org/10.1017/S0266267115000346

Pickering, A. (1984) “Against Putting the Phenomena First: the Discovery of the Weak Neutral Current”, *Studies in History and Philosophy of Science* Part A 15 (2):85 (1984).

Pickering, A. (1995). *The Mangle of Practice*, Chicago: University of Chicago Press.

Rabin M. 2002. “A perspective on psychology and economics”. *European Economic Review* 46, 657–685.

Rosenberg, A. 2012. *Philosophy of Social Science*. 4th ed. Boulder, CO: Westview Press.

Ryle, G. (1949/2009) *The Concept of the Mind*, London: Routledge.

Santos, Ana C. (2009) “Behavioral experiments: how and what can we learn about human behavior”, *Journal of Economic Methodology*, 16:1, 71-88, DOI: 10.1080/13501780802684278

Smyth, J. D., D. A. Dillman, and L. M. Christian. 2009. “Context effects in Internet surveys: New issues and evidence.” In *Oxford Handbook of Internet Psychology*, edited by A. N. Joinson, K. Y. A. McKenna, T. Postmes, and U.-D. Reips, 429-446. New York: Oxford University Press.

Soler, L. (2008) “The Incommensurability of Experimental Practices: An Incommensurability of What? An Incommensurability of a Third Type?”. In Léna Soler, Howard Sankey, Paul Hoyningen-Huene (eds.) *Rethinking Scientific Change and Theory Comparison: Stability, Ruptures, Incommensurabilities?*, Boston Studies in the Philosophy of Science, volume 255, Boston: Springer, 299-339.

Stanford, Kyle, (2021) "Underdetermination of Scientific Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/win2021/entries/scientific-underdetermination/>.

Suppes, P. 1962. “Models of Data.” In *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes, and A. Tarski, 252–61. Stanford, CA: Stanford University Press.

Turnbull M. G. (2017) “Underdetermination in science: What it is and why we should care”, *Philosophy Compass*, 2018;13:e12475.https://doi.org/10.1111/phc3.12475TURNBULL11 of 11

Woodward,J.1989.“Data and Phenomena.” *Synthese* 79:393–472.

Woodward, James (2000). “Data, phenomena, and reliability”. *Philosophy of Science* 67 (3):179.

Woodward, J. (2010) “Phenomena, Signal, and Noise”, *Philosophy of Science* , Vol. 77, No. 5 (December 2010), pp. 792-803.

Zaller, J., and S. Feldman. 1992. “A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences.” *American Journal of Political Science* 36 (3): 579-616.

Author biography

María Caamaño-Alegre is an Associate Professor of Philosophy of Science at the University of Valladolid (Spain). She joined CSLI at Stanford University as a Fulbright post-doctoral fellow before starting her teaching career at the Philosophy Department (UVa). Her research focuses on topics at the junction of general philosophy of science, methodology of science, philosophy of language and epistemology. She has published extensively on incommensurability, experimental validity, evaluation of theories, and inter-theoretical relations. Her current interests include the relation between natural and social sciences and the methodological differences between both fields in the search for empirical support.