From Ontological Traits to Validity Challenges in Social Science: The Cases of Economic Experiments and Research Questionnaires

María Caamaño-Alegre is currently an Associate Professor of Philosophy of Science at the University of Valladolid (Spain). Her main research interests are in general philosophy of science, methodology of science, philosophy of language and epistemology. She has focused on topics at the junction of these fields, like incommensurability, experimental validity, evaluation of theories, and inter-theoretical relations.

José Caamaño-Alegre is Associate Professor of Political Economy and Public Finance at the University of Santiago de Compostela (Spain). His research has mainly focused on public budgeting and financial management, and he is the author of an extensive monograph and numerous articles in the field. Other subjects he is interested in are the economics of the criminal justice system, and the methodological and philosophical aspects of economics and public management.

Postal Address for correspondence with authors:

María Caamaño-Alegre

Departamento de Filosofía

Facultad de Filosofía y Letras

Universidad de Valladolid

Plaza del Campus s/n,

47011 Valladolid, Spain

Abstract

This article examines how problems of validity in empirical social research differ from those in natural science. Specifically, we focus on how some ontological peculiarities of the object of study in social science bear on validity requirements. We consider these issues in experimental validity as well as in test validity because, while both fields hold large intellectual traditions, research tests or questionnaires are less closely connected to natural science methodology than experiments.

Keywords: experimental validity, test validity, economic experiments, survey research, social domain

1. Introduction

One way to address the issue of the divergence between natural and social sciences is to examine how problems of validity in empirical social research differ from those in natural science. For the sake of simplicity, we are going to put aside some undoubtedly significant aspects, like the role of interests behind social research, and focus instead on how ontological differences between the object of study in natural science and that relevant to social science bear on validity requirements. To the best of our knowledge,

this kind of approach has not as yet been broadly and systematically developed. Our analysis shows that some deep ontological peculiarities of the social domain underlie many everyday obstacles and confounds that social scientists deal with. Tracing the ontological roots of validity challenges does not prejudge the ease or difficulty of the problems posed at the methodological level. There may be ready solutions to any of them, and method refinements will probably enable the future solution to some currently unsolved issues. Overall, however, the problems pinpointed here make it hard to envisage a social science providing the same validity warrant in empirical research than that guaranteed in most natural sciences.

In emphasizing the contrast between social and natural sciences, we shall often take physics as a typical example of the latter. We are aware, however, that other natural sciences may share, to a lesser or greater extent, some of the ontological traits and validity challenges presented here as specific to social sciences. The economic experiments and research questionnaires that we have chosen as illustrative cases also contribute to highlight the above-mentioned contrast. Economic experimentation tends to focus on explaining individual human action and, at this level, subjects' intentionality and mental states are an unavoidably issue. Questionnaire research, moreover, ultimately amounts to an attempt at getting direct access to such mental states. In addition, our choice of cases covers a wide disciplinary scope in the search for validity, from examples in experimental economics to instances of questionnaire research in psychology and its related sciences. By considering issues in *experimental validity* as well as in *test validity*, we capture a wide range of problems related to empirical social research, for, in comparison to experiments, research tests or questionnaires are less closely connected to

natural science methodology. Although both experimental and test-based social research hold large intellectual traditions in dealing with validity problems, our analysis reveals an interesting similarity between the problems found in the two fields, a similarity that steams from their confrontation with the same ontological peculiarities of the social domain.

Our paper is structured in three main sections. Section 2 provides a rough overview of the notion and kinds of validity, as well as of the main threats to validity, in empirical social research. In section 3 we analyze the main sources of difficulty arising from the special nature of the social domain, taking into account actual practices from experimental economics and questionnaire research. Section 4 concludes the paper by showing the more general philosophical significance of the above issues for the current normative framework shaping social science.

2. The Standard Approach to Validity in Empirical Social Research

Traditional philosophical approaches to validity were especially concerned with both theory validation or theory testing and the attribution of logical rationality to science (Messick 1989). The two classical accounts of validity in Philosophy of Science — putting aside the more recent contributions by philosophers of experiment like Hacking (1983), Franklin (2005), Galison (1997), Steinle (1997) or Mayo (1996)— respectively revolve around the notions of verifiability or confirmability and falsifiability. Here we pay attention to the enlarged view on validation coming from the social sciences. Within this field, where the discussion embraces more than just those research components

devised for the purpose of testing a theory, validity and reliability are characterized both as logically independent notions and as commonly associated properties of measurements and procedures. Reliability concerns the extent to which an experiment, test, or any measuring procedure yields consistent results internally, temporarily and across observers (Pelham and Blanton 2003, 70-77, Carmines and Zeller 1979, 11-13), while the validity concerns the degree of success in attaining the purported outcome (that is, in determining the variable under study). The common association between reliability and validity is due to the fact that the first is usually required in order to establish the validity of the procedure or just to guarantee its useful applicability.

The notion of validity was originally developed from two different traditions in social science, namely, experimental and test research (Table 1). In the 1950s, the basic distinction corresponding to the first tradition was that between internal and external validity (Campbell 1957). As for the second tradition, the main kinds of validity were criterion, content, and construct validity (Cronbach and Meehl 1955). In an attempt to cope with different methodological challenges, test community gradually embraced an enlarged and unitary concept of validity (Angoff 1988, 25; Sireci 2009), one based on a comprehensive notion of construct validity which comprises all sorts of empirical support for test interpretation and use (Messick 1989). Contrary to this, Shadish, Cook, and Campbell (2002) kept the primary association of validity with the truth of knowledge claims, and integrated construct, external and internal validity, along with statistical conclusion validity, in other unitary framework applicable to either kind of social empirical research involving causal inference.

Table 1 about here

Like in Barron et al. (2008), we adopt the typology by Shadish, Cook, and Campbell (2002) as an umbrella under which we can integrate those test validity notions (for instance, construct validity) relevant for purposes of social scientific research. In the case of two variables, *statistical conclusion validity* implies the appropriate use of statistics to infer whether the presumed independent and dependent variables are correlated[1], whereas *internal validity* refers to whether the covariation between such variables results from a causal relationship. Noise and confounds are the main general threats to both kinds of validity, whose achievement is difficult to obtain even in experimental settings where variables can be partially isolated to test their impact on a single dependent variable. *Construct validity* concerns the extent to which a particular empirical indicator (or a set of indicators) represents a given theoretical concept, that is, the extent to which independent and dependent operative variables truly represent the abstract, hypothetical variables of interest to the scientist (Pelham and Blanton 2003, 66; Shadish, Cook, and Campbell 2002, 65). Construct underrepresentation (empirical indicators leave out something that should be included) and construct-irrelevant variance (they include something that should be left out) constitute the two main general kinds of threats faced by construct validity. Finally, *external validity* refers to the appropriateness of generalizations from results obtained in a research setting to phenomena out of such

[1] Although this raises several issues, the emphasis has traditionally been made on achieving enough *statistical power* to avoid Type-II errors or undetected existing correlations, and enough *statistical significance* to avoid Type-I errors or apparent correlations that do not exist (García-Pérez 2012).

setting. It thus concerns the extent to which a set of research findings provide an accurate description of what typically happens in the real world (Pelham and Blanton 2003, 64). Some of the typical factors jeopardizing this kind of validity are the contrived nature of the testing settings, and selection biases.

3. Validity Threats Specific to Social Science

A main assumption in the present discussion on validity concerns the special character of the social domain in contrast to that of the natural domain. This, of course, presupposes a distinction between both domains, although not necessarily one that should be understood as a dichotomy. More precisely, we are not committing to the anti-naturalistic view that both domains constitute two radically different, disconnected ontological realms. From our perspective, the distinction makes sense in so far as what is called "the social domain", even if in continuity with what is called "the natural domain",[2] exhibits certain peculiar properties not present in the latter —although here is not the place to elaborate on this topic, it may be worth noting that Davidson's (1974/1989, 229-239) anomalous monism and Kim's (1988, 310-312) criterial naturalism are some of the conciliatory approaches combining a moderate naturalism with a dualism between factual or natural properties, on the one hand, and normative or intentional supervenient properties, on the other. Three characteristic features of the (human) social domain should be emphasized: 1) representational and symbolic capacities typical of human agents (including self-consciousness or self-representation), 2) intentional states and behavior partially

[2] For a recent defense of the "social sciences as life sciences" see Duprè (2016).

determined by values and desires, 3) cultural and conventionally mediated forms of social interaction. We are aware that different kinds of living organisms, from ants to chimpanzees, may exhibit some or all of the above features to some extent, yet, as long as human creatures show those features to a much greater degree, it seems appropriate to deal with the human domain separately.

In what follows, we focus on these ontological differences as a source of validity challenges in empirical social research. Contrary to planets, electrons, magnetic fields or wave-lengths, human beings have representations of both their surroundings and themselves, and hence the recognition of observers or experimenters studying them may affect their conduct. In addition, the intentional character of human actions requires social scientists to identify the subjects' motives, which in turn consist of reasons and intentions —being the latter dependent on both representations and values or desires (Davidson 1963/1989). Reasons add further complexity to the picture, since they entail an epistemic attitude towards representations, i.e., the attitude that a certain representation successfully fits reality. Now, beliefs are commonly considered holistic dispositions to think in certain ways, as something being or not being the case. As famously emphasized by W. V. O. Quine, they constitute an inferential network in which no belief is independently formed, but rather every belief is radically dependent on other beliefs, inherited knowledge, previous experiences, etc. Thus, different subjects should be presupposed different beliefs, which entails, not only a lack of homogeneity among people with regard to beliefs, but also a clear source of uncertainty or instability in the social phenomena. However partial or ambiguous, linguistic communication provides (together with observed behavior) one main access to other people's representations and intentional

states. Furthermore, shared representations and high order beliefs[3] make it possible to establish conventions ruling social behavior. Conventions are only found in the social human domain, where conscious, intentional organisms, with high level representational capacities, make decisions about how to symbolically organize their behavior. Regularities in social behavior stemming from conventions may be as diverse and open to change as the reasons, intentions and cultural idiosyncrasies behind their constitution. As already noted by Weber, these ontological differences originate a methodological one, since only in social science causal explanation will require taking behavior out of the domain of the psychological and putting it into the domain of the culturally determined forms of responding to the world (Weber 1949, 66-76, Huff 1982b, 208-209).

In the rest of this section, we delve into these issues by examining some general threats to validity in empirical social research and by connecting each of them to the respective ontological trait.

3.1. The problem of awareness (on the side of the subject under study)

In an enlightening paper from 1968, J. A. Wiggins carefully analyses the different and frequently ignored sources of extraneous variables occurring during an experiment. Without reproducing his detailed taxonomy here, we are going to draw attention to those kinds of confounds that he points out as resulting from the subjects' awareness of different aspects concerning the experimenter, the manipulations and measurements of an experiment. Also in questionnaire research, the respondents' awareness of several

[3] That is, beliefs about others' beliefs, and about others' beliefs about others' beliefs, and so on.

interviewer and instrument related aspects may put validity into question. We will first consider the confounds generated by the subjects' awareness of the experimenter or interviewer, secondly we will pay attention to the variation in involvement resulting from awareness of the research procedures, and, thirdly, we will discuss the variation in beliefs and behavior as a result of recognizing those procedures and their artificiality. All these different, although related sources of confounds obviously endanger the internal validity of causal inferences made from results obtained through those procedures, for the presumed independent variable may not be the only or main causal factor operating in the research setting. Furthermore, even if the confounds are carefully identified and studied, their presence would still restrict the exportability of the results, i.e., the external validity of the research procedure.

a) Experimenter/interviewer effects

With regard to the experimenter, it is possible that the subjects' recognition of the former's conceptualizations (methodological, theoretical), characteristics (sex, race) and behavior (verbalizations, gestures) affects their behavior.[4] In this case, the experimenter would have an unintentional influence on the subjects' attitude and conduct, with a possible experimenter-modelling effect tending to yield, in the subjects, a behavior consistent with the experimenter's expectation or hypothesis (Wiggins 1968, 399-400).

[4] There are other aspects of the experimenter's behavior that may also affect the subjects' response once the latter have become aware of them. Among these aspects, Wiggins (1968, 400) mentions maturation (experimenter's fatigue or boredom due to increasing familiarity with the experiment), and change in the degree of confidence in the hypothesis during the experiment as the experimenter analyses data between the beginning and the end of the data collection.

This threatens both internal and construct validity, since the confound consisting in the subjects' recognition of the experimenter's expectations would result in misleading, inadequate evidence in favor of the hypothesis. In economics, experimenter-effects have been a serious concern; for instance, experimental auctions in certain developing countries have shown that the limited experience of participants makes it more probable that experimenters influence the bidding process (Morawetz, De Groote, and Kimenju 2011, 264-265). Another example can be found in a field experiment comparing charity lotteries to voluntary contributions as to their degree of effectiveness in raising money. The experiment revealed that one-standard deviation increase in the physical attractiveness of the female solicitor had an impact similar to that of the lottery incentive, which constitutes a clear experimenter-effect (Landry et al. 2006).

There is an obvious parallel between these experimenter-effects and the wide range of interviewer-effects well documented in questionnaire research. Such effects may differ between respondent categories and be influenced by highly complex and variable factors; take, for instance, the ethnicity-of-interviewer effects, shaped by what Van Bochove et al. (2015) refers to as "the multifaceted nature and context-dependency of ethnic identifications". It must also be highlighted that interviewer-effects may influence, not only the means obtained for particular items in a questionnaire, but also the covariances between items, and, thus, they may bias the factor loadings and correlations estimated in order to validate the construct (Beullens and Loosveldt 2014, 2016).

b) Lacking or low involvement

The subjects' awareness of the experimental manipulations may increase the risk of the subjects' low involvement with the independent variables due to their lack of

attention, lack of motivation, or to their awareness of other variables. In experimental economics, the use of financial incentives is conventionally required to deal with this problem, although some methodological studies in the field suggest that such incentives may matter more in some areas than in others. Moreover, Read (2005) argues that the three factors through which these incentives have their effect ("cognitive exertion", "motivational focus" and "emotional triggers") can often be activated without monetary incentives, and incentives are not even guaranteed to activate them. Smith, on the other hand, raises the questions of whether the usual degree of subjects' involvement, i.e., the involvement they exhibit when playing games of redistribution of the experimenter's money, would be the same if they brought or otherwise provided their own money (2010, 13). In a similar vein, Rosenboim and Shavit (2012) argue that providing the rewards at the time of the experiment may lead subjects to view the money as if it were not their own. To change such a view and contribute to a more natural environment, a "prepaid mechanism" is devised.

In questionnaire research, a similar risk is created by the respondents' lacking or low involvement, which may undermine both response rates and response quality, respectively deriving in non-response errors and measurement errors. According to Groves and Singer (2004, 36-38), refusal bias tends to arise when the topic of the survey crucially determines the addressee's decision to participate or not, especially in cases where the topic is closely allied with the mission of the survey's sponsor. Krosnick and Presser (2010) emphasize the wide range of intangible motives the respondents may have

to do a good job in answering a research questionnaire, from desires for self-expression, to intellectual challenge, self-understanding, altruism, or emotional catharsis.[5]

c) Beliefs about research devices and perception of their artificiality

The subjects' beliefs regarding laboratory procedures like the experimenter's attempts to deceive, or about the role of an experimental subject, or even about the experimental hypotheses, may significantly influence the subjects' behavior in unexpected ways. In particular, the subjects' recognition of the artificiality of laboratory experiments may lead to several possible confounds related to their reaction to different factors such as: participating in an experiment (feeling of importance, Hawthorne effect), the unfamiliar experimental task, perceiving the experimental consequences as unreal, or perceiving the ecological arrangements and social arrangements as unreal (Wiggins 1968, 412-414). Again, these problems seriously affect the overall validity of experiments, for unnoticed confounds of this kind not only undermine the internal validity of an experiment, but also produce misleading evidence that decreases both construct and external validity. Analogously, in survey research, the respondents' beliefs regarding what is behind the items' features and the instrument as a whole may call into question the validity of the responses obtained. The same applies to how respondents react to the

[5] They also distinguish three levels of involvement in answering questions: optimizing, weak satisficing, and strong satisficing. *Optimizing* means thoroughly and unbiasedly performing the cognitive tasks involved in answering questions. *Weak satisficing* is to offer the first answer that seems acceptable after incompletely and/or biasedly performing such cognitive task. *Strong satisficing* consist of selecting an answer without performing such tasks, and without reference to any respondent-internal cues relevant to the question, either looking for a cue in the question wording or making an arbitrary choice.

very fact of being surveyed, as well as to the lack of familiarity with the task of answering certain questions.

An interesting example of how economic experimenters confront the risk of unexpected effects due to the subjects' beliefs regarding laboratory procedures can be found in their proscription of deception. Indeed, this has been settled on a public good argument, to avoid the distortions induced by the experimental subjects' fear to be deceived.[6] With regard to the behavioral impact of the subjects' beliefs concerning experimental hypotheses, Levitt and List (2009, 15) point out that if one were interested, for example, in exploring to what extent race or gender influence the prices that buyers pay for used cars, seeking informed consent could directly interfere with the ability to conduct the research. As to the subjects' recognition of the artificiality, Bardsley (2005) criticizes, among others, Alm et al.'s (1992) experiment on tax evasion, which remains silent on the fact that people might recognize a civic or legal duty to pay taxes whilst not recognizing a duty to be honest to experimenters in labs, or indeed vice versa. Similar concerns are expressed by Hogarth (2005, 258-259) on market experiments, as it is not clear how well these "replica" models match the characteristics of "real-world" markets on all relevant dimensions —market experiments typically involving a limited number of

[6] Since doing without deception may be costly, McKenzie and Wixted (2001, 424) suggest a different way to avoid the misinterpretation of participants' skepticism induced by his/her deception fear as evidence of non-normative behavior. Their proposal consist in deriving and testing normative models that do not assume full belief in key task parameters. This elegant approach, however, is not found completely convincing by Hertwig and Ortmann (2001, 438), who criticize that it introduces a free parameter into the models —increasing the danger of data-fitting— and assume that it will often not be applicable because of an insufficient, case-sensitive understanding of the distrust effects.

participants as opposed to real markets, which can involve millions. The threats to external validity arisen from lab artificiality have led to a great upsurge in economic field experiments. But, even in field experiments, generalizability across similar field settings is an issue (Camerer 2015) and arriving at artifacts remains possible, either because of submitting the controlled factors to "unnatural" rules of variation or due to the introduction of unnoticed uncontrolled factors that lead to spurious results (Boumans 2016, 143). In fact, the analogy between economic field experiments and clinical trials is only partial in that the former are not framed in any ex-ante knowledge equivalent to that resulting from the preclinical, I and II phases of a therapeutic trial and, therefore, they may fail to provide univocal explanations and clear recommendations (Favereau 2016).

In questionnaire research, it is well known that variations in instrument's features like, for instance, question wording, response options and question order may produce systematic changes in answers.[7] From a certain point of view, these *response effects* suppose an answer variance not attributable to any corresponding variance in the underlying trait being measured and, therefore, they may jeopardize validity. As Groves and Singer (2004, 40-43) explain, these phenomena have been interpreted as instances of "constraint" of the responses by formal features of the questionnaire, and, from the perspective of cognitive psychology, they have been understood as reflecting "implicit instructions" to include or exclude certain phenomena or to interpret the concerned categories and assess their properties in a certain way. Although a panoply of pretesting methods is available to deal with the measurement errors linked to response effects,

[7] This problem has also been extensively discussed under the label of "framing effects". For a typology of valence framing effects, see Levin *et al.* (1998).

pretesting is “itself a mix of science and craft” (Krosnick and Presser 2010, 295) and response effects remain to be a serious concern, as shown in Greenhill et al.’s (2014) and Yount et al.’s (2011) recent examples.

Ultimately, the main problem resulting from the subjects’ awareness of measurement procedures is *reactive measurement*, which has to do with the fact that the process of measurement may itself change its target. Although this problem also affects physical experiments, like for example those in quantum physics, the width and degree of measurement sensitivity on the side of the physical target is not comparable to that of the social target. As explained later, the wider scope of uncertainty in social experiments is closely connected to the representational and cultural features characteristic of social agents. Granting that observation, the questionnaire and the interview are the three most commonly used measuring instruments in social science, Wiggins (1968, 418) draws attention to the possible uncontrolled influence that artificiality of measurements may have on the subject’s behavior, therefore risking the external validity of the experimental findings. The resort to premanipulation or pretest measurements of dependent variables (subject’s motivations, beliefs, real-world behavior) does not solve the problem, for such measurements may also affect the subject’s performance on postmanipulation measurement (maybe providing some unintended hints about the purpose of the experiment). In the case of questionnaire and interview methods, the reactive measurement may also be prompted by the interaction between two variables (subject variation and instrument variation) centered on the same individual.

Last but not least, awareness of theoretical constructs also hampers the possibility of determining some target variables that are assumed to fall under some of those very

constructs. We are referring to what Ian Hacking (1995/2011) calls the 'looping effects of human kinds', which we think can be straightforwardly applied to theoretical constructs classifying individuals in social science. Although Hacking illustrates the looping effects mainly with medical, psychiatric or sociological constructs like child abuse, teenage pregnancy and multiple personality, they could also be illustrated with constructs from other fields, like poor/rich population, marginal citizens, unemployed people, immigrant, etc. The crux of the matter is that social constructs, as opposed to natural constructs, change the target variable. Unlike humans, mass is indifferent to our theoretical constructs. A given mass is not going to lose or gain atoms depending on our construct regarding its number of atoms, but human beings may change their behavior depending on whether they are aware that they fall under the construct 'poor marginal immigrant'. In fact, the looping effect consists in changing so that our features agree with those characteristic stablished by construct. This problem clearly affects statistical and internal validity, since the target domain scientists intend to study is being modified in an unintended and hardly traceable way. Interestingly, an illusion of construct validity could emerge from this situation, as looping effects reinforce the fit between the construct and the target domain.

To sum up, the subjects' awareness of different aspects concerning the experimenter or interviewer, the constructs, the manipulations and measurements poses serious threats to internal validity in the form of different sorts of noises and confounds. Awareness of the artificiality of the research setting, on the other hand, dramatically threats external validity. Construct validity, in turn, is affected by misleading evidence

stemming from the subjects' recognition of different aspects concerning the research procedure and/or the theoretical construct.

3.2. The problem of identifying motives as causes of behavior

Explicative hypotheses about human behavior postulate motives as causes of such behavior. As pointed out earlier, this entails postulating intentions and reasons to act in a certain way. Ultimately, postulates may refer to the agents' representations, values, inferential dispositions and intentions. On the one hand, however, none of the latter is an observable kind of thing, nor something amenable to (strict) measurement; and, on the other hand, none has a simple univocal origin, nor a simple, univocal observable manifestation in behavior.[8] In contrast to what is more common in the natural domain, where similar effects usually have similar causes, in the social domain, the same behavior may be caused by very different motives, and the same motive, in turn, may cause highly heterogeneous behavior given the intertwined and culture-dependent character of social behavior. For this reason, the same pattern of behavior —let us say, buying a car— may arise from different motives —trying to impress other people, replacing an old, ill-functioning car, helping a relative with a car business, etc. And, conversely, different patterns of behavior may stem from the same motive, as, for instance, when different people take different actions in order to become rich —some may play lottery, others

[8] 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).

may speculate with stocks, or rather they may save as much as possible. In addition to this, the same motives may be determined by very different combinations of representations and desires. Thus, because of the unreliability and ambiguity affecting behavior, linguistic access to the subjects' intentional states provides one of the main means to gather information about those states. Consequently, questionnaires and interviews take on special significance as tools for motives discrimination.

Ultimately, some striking peculiarities of the social domain boil down to the mind-dependent nature of some key elements from such domain. The idea that social objects themselves (not just their representations) are mind-dependent has been raised by Uskali Mäki, who argues that their very existence requires minds. As he puts it:

> "While it may be plausible to claim that galaxies and quarks exist mind-independently, this does not seem a good idea in the case of, say, people's preferences and expectations or a society's institutions and organisations" (Mäki 2008, 336).

In a similar vein, Hacking stresses the value-ladenness, not just of social science, but of its very object of study. According to him, human kinds like "chronic unemployed", "juvenile delinquents" or "homeless" are themselves laden with moral values, i.e., with a mind-dependent feature (Hacking 1995/2011, 34-35). Moral values, however, are not the sort of thing that can be easily isolated and measured. In fact, Hacking's main concern is related to how the peculiar, complex dynamics of the social realm results from the individuals' awareness of moral values socially ascribed to human kinds.

3.2.1. *Lack of relevant structural homogeneity between individuals with respect to their psychological properties*. Let us consider in more detail this first key aspect involved in the inscrutability of motives as causes of behavior —an inscrutability aggravated by the above-mentioned spurious inter-individual divergence and convergence in the behavioral manifestation of motives. The lack of structural uniformity between individuals' internal properties has been emphasized by Suppes (1982, 247):

> "The great success of physics and chemistry has depended upon the structural identity of substances (…). It is a plausible thesis that we do not have in the case of persons or even other animals anything like such uniformity of structure; rather, one person's internal structure at a given moment is in no interesting way isomorphic with the internal structure of another person. By 'interesting way' I mean of course in terms of psychological properties and not gross physical properties. If the situation is as hopeless as I am inclined to think it is, this means that the methodology of the social sciences and the development of causal theories must take quite a different direction than that which has been so successful in the physical sciences".

The lack of isomorphism between individuals' internal structures poses serious problems for the empirical determination of independent variables, thus affecting both statistical and internal validity, not less than construct validity. As opposed to physical independent variables like gravitation, that would entail some empirically determinable isomorphism between mechanical systems with respect to properties like distance and mass, psychological independent variables, like the desire to increase one's profit, do not

have any specific structural configuration determinable by available empirical methods. As a consequence, psychological independent variables in social science are not only unobservable, but also unmeasurable, since strict measurement presupposes the homogeneity of the measured attribute from one instance to another. In so far as the empirical indicators for postulating psychological properties come mainly from phenomenological factors, the empirical basis for the postulation of causes remains highly qualitative and dependent on rather ambiguous behavioral indicators.

The above inter-individual mismatch between motives and behavior directly affects both internal and construct validity. High behavioral correlations, like, for instance, that between summer season and spending more money in purchases, do not allow a causal inference between both things, since different motives, like enjoying shopping as a holiday hobby or taking advantage of sales, may be independently playing a causal role and only contingently connected to the summer season. This situation poses some problems for construct validity as well, since it may be difficult to cover all the potentially relevant kinds of empirical indicators for a construct (like "shopping as a hobby"), and to gather convergent evidence accordingly. The search for discriminant evidence is also hampered due to the difficulty in attributing an empirical feature exclusively to one construct —for instance, the empirical indicator of saving money may be connected to both the desire to become rich for one's own benefit and the desire to remain austere for the benefit of the descendants.

Empirical indicators for psychological properties, therefore, are usually not grounded on sound nomological networks where inherited, well-supported empirical laws provide the basis for connecting the empirical properties of the postulated independent

variable with certain empirical effects ascribed to it.[9] Natural scientists assume that, in order to empirically test a correlation between two variables, a clear empirical determination of both variables should be available beforehand. Gravitation, as a cause, can thus be tested by checking whether a planet's trajectory is determined by the respective masses and relative distances between it and some other celestial bodies. In this case, the empirical test will consist in showing that the kinematic effects (i.e., the dependent variable) assigned to the gravitational force (i.e., to the independent variable) are certainly correlated to other empirical indicators (i.e., mass, distance) also assigned to such force. Within social science, however, no such prior empirical determination of psychological independent variables seems attainable; instead, the empirical support for those independent variables seems to come only from their expected correlation with the full range of corresponding dependent variables. This empirical gap affecting independent variables highly increases the empirical under-determination of theoretical constructs, making it difficult at the same time to empirically determine correlations between specific, clear-cut psychological attributes and behavioral patterns. Psychological attributes may accommodate the scope of phenomena that led to its postulation in the first place, but they may not find sound support beyond that scope. Even in barely predictive sciences like geology, independent support is required and must be obtained for a theory to be acceptable. Despite the fact that it accommodated a whole range of heterogeneous phenomena, it took half a century for the theory of continental

[9] Nomological networks play a pervading role in natural science, both in the form of tree-like networks of specializations rooted in the same basic core, and in the form of inter-theoretical links between a theory and independent, well-established previous theories providing the empirical concepts and laws to determine the empirical basis of the former.

drift to be accepted, an acceptance that only occurred after some unexpected oceanographic evidence supporting the theory was gathered.

In conclusion, the correlations between internal structure and external effects, which are the building blocks of natural science, have no analogue in social science. In contrast to the hierarchical nomological network characteristic of natural science (built from more directly measurable attributes to less directly measureable ones, as well as from more empirical to less empirical laws), nomological networks in social science typically lack a clear empirical hierarchy. This leaves too much room for construct irrelevant variance, since the same dependent variables may be easily ascribable to different, highly conjectural theoretical constructs. The above remarks 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, explicative correlations —i.e., those ultimately concerning intentional dispositions and psychological attributes— cannot be as firmly empirically grounded as they are in natural science. It must be noted that the above issues affecting construct validity directly bear on statistical conclusion and internal validity. In particular, correlations involving psychological variables will be neither strictly testable —on the basis of the structural features of the attributes—, nor unequivocally interpretable as a causal relation —given that, without an empirically accessible internal structure, psychological attributes pointed as causes cannot be empirically isolated from others.

In experimental economics, the empirical under-determination 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 complemented and/or combined with nonchoice-data provided by psychological studies (Schotter 2008, Ross 2010). In order for economists to satisfy their current tendency to dig further into social motivations, there is no other way than to examine psychological constructs such as justice, trust, reason, desire, belief, among others (Tyler and Amodio 2015, Dietrich and List 2012). Carpenter, Connolly and Myers (2008), for instance, used survey measures of altruism to test the construct validity of their experimental protocol for a representative dictator experiment, where each participant had to choose a charity and then divide a given amount of money between the charity and herself. The fact that there was a positive and statistically significant association between the altruism factor scores from the survey on the one hand, and the amounts of money given in the dictator game on the other, was taken as evidence supporting the construct validity of the experimental protocol. In checking auxiliary assumptions in economic experiments, psychological constructs prove also of the highest relevance. Some of the latter, like beliefs, expectations or moods, should be acknowledged as essential to examine, for instance, central auxiliary assumptions concerning the *environment*, particularly those about the agents' characteristics and their preferences. The same goes for precepts like that of *dominance*, whose verification as well would require 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.[10]

[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 (nonsatiation, saliency, dominance, privacy).

In questionnaire research, the empirical under-determination of psychological constructs and lack of sound nomological networks seriously challenge test validity. Denny Borsboom et al. (2009, 135-170) have 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. 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 are significantly correlated in humans (Borsboom et al. 2009, 156-157). The rare existence of tight nomological networks in the social sciences, especially in those areas more prone to use questionnaire research, has imposed a shift to a "weak form" of construct validity. As emphasized by Kane (2006b: 442), nomological networks were initially envisaged by construct validity theorists like Cronbach and Meehl (1955) as formal theories of the kind exemplified by Newton's laws in physics. However, given that such theories are generally nonexistent in disciplines like psychology, the initial requirement was relaxed so as to demand only open-ended collections of relationships for each construct. To use Cronbach's (1989) words, this entails a shift from a "strong form" to a "weak form" of construct validity. As later claimed by Borsboom et al. (2009), the weak form of construct validity made it very difficult to evaluate the construct's fit to the network, for the collections of relationships for each construct could be both vast and ill-defined.

3.2.2. *Actions holistic dependence on the individuals' complete past (or the non-applicability of the Markovian condition).* The potential relevance of the individuals' complete past highly increases the difficulty in identifying motives as causes of behavior.

In searching for the causes of a given action, the complexity of independent variables connected to the individual's past, as well as to her retrieval of such past by memory, may become unmanageable. Suppes (1982, 246-248) draws our attention to this point as he states the following:

> "(…) in considering the current phenomenological properties of a physical object (…). It is not necessary to know anything about the history of the object from a theoretical standpoint if we know the current microstructure. (…) This radical truncation of the past is one of the most essential general concepts in the physical science. (…) What may be characteristic of the social sciences and what makes them scientifically very difficult is that we shall not be able to move from chains of infinite order in terms of phenomenological variables to well-defined enlarged Markov processes with underlying theoretical states that render knowledge of the past otiose".

As stressed in the passage, the problem related to the predictive relevance of an individual's complete past is partly connected with the one discussed earlier, namely, the lack of an empirically determinable internal structure. Without a recognizable psychological structure, one that can be systematically associated with certain behavioral patterns, social scientists are just left with an indefinitely open number of phenomenological variables corresponding to an indefinitely open number of causally relevant factors from an individual's past. The main reason why this is so has to do, not so much with scientific limitations, but with the extremely dynamical nature of intentional creatures like human beings, a feature that turns out highly potentiated by what Rosenberg (2009, 62-64) call the "reflexive" nature of psychological processes, i.e.,

the fact that any new experience or information gathered by an individual may significantly change the psychological properties of such individual. This very feature makes it impossible to postulate an internal structure satisfying a Markovian property, i.e., one such that a full knowledge of the past would not give rise to any change in the prediction of future behavior. A person's context of decision regarding a particular issue includes all past experiences, thoughts and pieces of information that are relevant with respect to that issue. All the complexity of human memory, together with its interaction with feelings, beliefs, desires, etc., is transferred into the problem of understanding individuals' actions. For example, one's present decision to donate money to a non-profit organization may be influenced by a wide variety of experiences, pieces of information and thoughts —some long kept guilt feelings, increasing social pressure, a recent documentary on developing countries, and the possibility of easily donate through internet. Tomorrow's decision to stop making donations may come from some rumors that non-profit organizations are unreliable and from a sudden resolution to save money in order to go on a trip.

In experimental economics, the Nobel prized Vernon L. Smith (2010, 3) acknowledges that "human motivation may be so inextricably bound up with circumstances that embody previous experience that it is not even meaningful to separate them." Moreover, he asserts that the powerful *context effects* found in economic experiments depend on the subjects' previous experience, as the latter determines how subjects react to any experimental circumstances other than explicit payoff. If people's decisions are often as sensitive to the specific context as to variation in the structure of the game, it is because each context leads them to search, into their past experiences, the

potentially relevant knowledge for the task they face. When experimental subjects with a given cultural experience are confronted with an unfamiliar situation, their autobiographical knowledge is filtered by contextual circumstances in the search for relevant information concerning the decision at hand (Smith 2010, 11).

By appealing to the subjects' testimony, tools such as questionnaires and interviews provide a way to avoid a main threat to the internal validity of the experiment, such as the ambiguity of experimentally observed behavior. However, linguistic intervention and testimony may be also highly misleading. Experimental attempts to convey controlled pieces of information may face two main obstacles: the incompleteness and the unintended character of the information conveyed to 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 interpretative discrepancies between the experimenter and the subject. In short, it is highly difficult to control the communication process so as to guarantee that the information provided by the experimenter constitutes the cause of the subject's predicted behavior.

C*ontext effects*, like, for instance, 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. 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.[11] In any case, what matters here is that context effects are caused by the unobserved interaction between the subjects' stored beliefs and triggers generated by the survey instrument (Morgan and Poppe 2015). Along the lines of Smith's statement on experimental subjects, it could be said that each survey respondent's autobiographical knowledge is filtered by contextual circumstances for relevance to the decision on responding or not as well as to the choice among response options. The former decision results from the interplay between personal attributes developed over the subject's complete past, social environment, survey features, and interviewer's introductory expressions (Groves, Singer, and Corning 2000, Groves and Singer 2004). The response choice depends on the sample of the 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). In so far as survey responses would thus not result from preexistent answers or true attitudes, the divide between genuine opinions and survey artifacts becomes quite blurred.

3.3. Instability of social settings: highly variable, holistic and recursive nature of social interaction

[11] A more flexible notion includes both semantic and serial order effects, the latter resulting from the location of a question in a sequence of items —for instance, the fatigue effects of the later items. In this vein, Billiet, Waterplas and Loosveldt (1992, 131) define context effects as "response effects coming from one or more preceding questions (and answers) or from response scales belonging to previous questions."

In the previous section we have emphasized the fact that individuals' motives are extremely sensitive to the individuals' complete history. Now, even acknowledging the uniqueness of each individual's past, it seems important to explore the social context for possible sources of shared influences. The problems of holism and reflexivity, however, reappear also at this level, since events are not only interconnected in each individual's past, but also in the social setting shared by different individuals. As pointed out by A. Rosenberg (2009, 62-66), the reflexive (or recursive) character of social processes constitutes a main factor accounting for such holistic feature. Reflexive processes entail that one individual's or population's activity provides the basis for another individual's or population's activity. Any form of social organization or exchange, for example, may prompt many different kinds of responses (rejection, competition, collaboration, etc.), which may in turn cause new ones, and so on. In arguing that the predictive limitations of social science are due to the reflexive character of social processes, Rosenberg (2009, 64) states the following:

> "Most human interactions are strategic and so provide occasions for the selection of new strategies that upset temporary equilibria early and often. Since genetically encoded design solutions vary much less and vary much more slowly than neurologically encoded design solutions, an evolutionary equilibrium model will be closely approximated among biological lineages for appreciable periods, certainly for long enough periods that biologists will be able to state the model and even to estimate the degree of fit to the world".

Adopting a biological perspective —on which we will not comment here—, Rosenberg claims that some of social science's major predictive limitations arise, not from chaotic or exogenous variables —as it happens in natural science—, but from endogenous ones. Social settings can evolve in indefinitely many ways, not only according to different interests, but also to different conventional arrangements. The conventional aspect most clearly reveals how endogenous variables may cause the variation of social settings. The very possibility of establishing conventions ultimately implies the possibility of creating a social setting at will, just from the very resources provided by the social domain —namely, mutual recognition of shared interest, awareness of such mutual recognition, and intention to socially formalize an agreement favorable to those interests.

The above mentioned features of social interaction make it very difficult to support generalizations about behavioral patterns. Sandra Mitchell (2009, 141-142), who also exploits the parallelism between biology and social science, draws attention to the fact that, as opposed to *ceteris paribus* laws in physics, the ones in social science are open-ended. This means that we can never fully specify the interfering factors limiting the scope of our generalizations concerning social phenomena. As she points out, even if particular causal explanations of particular phenomena may be successfully applicable, our attempts at providing general explanations for general phenomena are systematically limited by the complex and dynamical nature of the social domain. As a consequence, even basic correlations between targeted psychological attributes and behavior will have a very narrow scope and quite ambiguous evidence, thus posing some serious problems for both the statistical conclusion and internal validity of the correlations. If basic

correlations turn out highly problematic, any attempt at devising predictively valuable explanations seems even less promising. Finally, because of the highly intertwined and dynamical nature of social interaction, laboratory experiments performed by social scientist always face a high risk of artificiality, which inevitably leads to a low external validity.

Taking a step further, Leonardo Ivarola (2017) contends that both the open-ended nature of socioeconomic processes and the inapplicability of the *ceteris paribus* model of scientific law in social science call into question "the logic of stable causal factors", still underlying recent accounts like Cartwright's discussion on capacities and nomological machines. Whereas, in a mechanism, the activities that mediate causal relationships are supposed to be stable, in the social realm these activities are people's actions not necessarily stable. Indeed, these actions may be unstable due to the joint action of the interpretations that agents make of signals of the world (newspaper information, financial rumors…) and contextual or structural conditions (institutions, culture, environment, moral principles…). There would not be "capacities" to isolate that stably contribute to the production of a result, because "causal contributions would have no predetermined capacity, but a set of *potential* capacities, which crucially depend on the multiple activities agents are able to perform." (Ivarola 2017, 217) Also the possibility of "nomological machines", protected from any perturbing factors, would be hampered in the social realm by the presence of open systems exposed to disrupting and unexpected exogenous factors, as well as to "endogenous" problems too. Therefore, Ivarola (2017, 218) concludes that the basic structure of socioeconomic processes is better captured by the notion of possibility-tree or open ended result than by that of nomological machine.

It could be argued that many physical processes, like the motions of billiards on an oval table or atmospheric changes, are also highly chaotic —which is to say, their dynamics are sensitive to arbitrarily small differences in initial conditions. The difference between these processes and their counterparts in the socioeconomic realm, however, lies not only in the degree of complexity and instability of the exogenous factors (initial conditions), but also in the degree of complexity and instability of the endogenous factors, like the dynamical recursive conditions mentioned by Rosenberg. In contrast to what happens in the above-mentioned examples of chaotic physical systems, where we can specify the kinds of factors and relations between them that prove most relevant in originating certain variation, the sources of uncertainty related to socioeconomic systems are themselves uncertain to a very high degree, being the recursive nature of social phenomena clearly behind this higher order uncertainty.

In sum, the open-ended nature of social phenomena ultimately leads to a recurrent problem of external validity of both models and experiments, since both of them tend to drastically restrict and (sometimes arbitrarily) fix the range of variables acknowledged as causally relevant. Of course, this is not to say that stable causal contributions cannot provide the basis for explaining social phenomena, but rather that they may be both extremely hard to identify and plurally connected to a single variable. The task of identifying causal contributions, and, consequently, the search for internal validity, poses several difficulties; some related to the possibility that stable causal contributions are not exercised (due to lack of some mediating activities), the other concerning the possibility that, even if exercised, stable causal contributions are not manifested at the level of events (due to interfering factors).

Economic experiments are problematic for the very reason that they do not represent in an appropriate way the variable and intertwined nature of economic behavior. To this regard, Hogarth (2005, 259) criticizes the basic logic of classic, factorial experimental designs, which revolves around the orthogonal variation of variables, even though outside the laboratory variables are not orthogonal and all other variables are not constant. The additive separability of the causal factors would be implicitly assumed even in field experiments. Yet, as Guala (2005, 132-134; 2017, 113) notes, randomization can lead to confounding results when there is a non-additive interaction between uncontrolled factors and the main treatment variable. Another problem, which lies behind Bardsley's (2005) examples of artificiality from experimental economics, is the impossibility to implement in the lab, without deception, and unchanged, those mediating conventions shaping social interaction. Following Greenwood (1982), he emphasizes that the nature of phenomena examined in economic experimentation is "relational" and depends on a "construction of social reality", because it is jointly determined by the existence of certain relationships between people and on people's perceptions that the corresponding relational criteria are fulfilled. For instance, a person only becomes a manager of a firm after she has been appointed to it; consequently, her actions will only count as managerial actions after she has been informed of that appointment. Bardsley concludes that, in cases like this one, a laboratory experiment cannot implement certain sorts of actions without deception, since subjects are aware of the activity established by the experimenter (Bardsley 2005, 241). Similarly, given the context specific nature of human learning, subjects' learning from the field should not be expected to prove readily transferable to their lab behavior if the

experimental environment deprives them of their customary contextual cues (Kagel 2015).

In questionnaire research, Groves and Singer (2004, 34) notice that cross-variations in certain social environmental attributes may affect accessibility to respondents, giving rise to non-response error and differences in the respondent pool's composition across surveys with disparate contact rates. According to them, variations in a set of social environmental influences, and even in specific social roles played by respondents, seem to impinge on the heuristics that guide decisions on responding or not to a survey —resulting, again, in non-response errors. They acknowledge that applying a standardized measurement for a whole population or for its representative sample, when subsets of the population have completely different conceptual frameworks regarding the issue under study, is a methodological procedure subject to "untenable positivistic assumptions". As they argue, the complicated interactions of human cognition, social norms, and social stratification lies at the bottom of many of the causes of survey error (Groves and Singer 2004, 56). On the other hand, it is not always possible to build a question context closely resembling the real context to which inference will be made, for, among other things, in many cases there is no single real-world analog (Krosnick and Presser 2010, 294). Even when the latter is available, the impossibility to implement, in a standard questionnaire-research design, those social processes and mediating conventions shaping social interaction may also originate artifacts or "pseudo-opinions" instead of relevant answers, i.e., answers with significant implications for the real-world stakes and behaviors of the target population (Malone, Dooley and Bradbury 2010).

The problem of the external validity of economic experiments and questionnaire results has been tackled by applying different strategies, some of which favoring triangulation and the search for robustness. As any experimental set-up and social-research instrument have their errors and biases, robustness can be achieved through empirical triangulation if similar results are obtained under change in the experimental or questionnaire- design (Ivarola 2017, 221). This strategy applied to seek external validity can also be useful to assure internal validity, since, the more a result appears under different designs, the more implausible the result is found to be an artifact of the research design (Boumans 2016, 144). But empirical triangulation can contribute to construct validity too, given that it constitutes a source of convergent and discriminant evidence for theoretical constructs. Nevertheless, to the extent that the chances of applying such empirical strategy are limited, theoretical approaches may be a necessary complement. To this respect, it is important to stress the relevance of the distinction between neglibility or heuristic assumptions and domain or substantive assumptions,[12] as Musgrave (1981) has noted in criticizing Friedman's approach. On the one hand, contrary to what happened in natural sciences, where neglibility assumptions are often enough to isolate the causal factor, in social sciences it is usually necessary to rely on substantive assumptions concerning some relevant features of the domain. This makes it essential, therefore, to check the empirical soundness of the latter in social research. Unlike what occurs with neglibility assumptions, if the substantive assumptions presupposed in the experimental or questionnaire- design are unrealistic, the range of applicability of

[12] The same distinction is often established also in terms of "Galilean" versus "non-Galilean" assumptions.

theoretical constructs remains undetermined, being thus devoid of empirical significance. In order to avoid this problem, and increase both external and construct validity, triangulation proves as a particularly interesting procedure.

4. Concluding Remarks

Along the previous section, we provided an extensive overview of the validity challenges arisen from some primary ontological traits of the social domain. The global picture that emerges from our journey through these validity issues specific to social research seems like a constant, and perhaps never ending, fight between social scientists' invention and sophistication and a social reality that is reluctant to the intended kind of "scientific" reduction. Piecemeal approaches, ceteris paribus clauses and experimental devices, as well as questionnaire design, pretesting and validation appear to contribute to the progress of knowledge and enable us to disregard certain formulations as naïvetés. However, the success of such a natural science-like approach to social research is limited by, among others, two fundamental reasons.

First, the intertwined and recursive nature of the social phenomena is also often present in the ontological traits that pose validity challenges to social science research. For instance, experimenter/interviewer effects obviously arise from subjects' awareness, but they may also be conditioned by subjects' previous experience (biographical holism) or by social-dependent identifications (social holism). Similarly, the problem of lacking or low involvement from subjects concerns awareness, but it may also result from a failure in implementing, either in an experimental setting or in a questionnaire-research

design, those cultural and conventionally mediated forms of social interaction and recursivity relevant to reveal the subjects' true behaviors or attitudes. Second, social scientists' responses to a validity challenge often tend to exacerbate or give rise to another validity challenge. This is the case, for instance, of the proscription of deception in experimental economics, which avoids possible distortions emanating from deception fears at the expense of exacerbating those linked to artificiality. The same occurs when questionnaire researchers, for instance, counterbalance choice order to handle response order effects, at the cost of creating a problem of response variance due to systematic measurement error (Krosnick and Presser 2010, 281).

Understanding social phenomena from a natural science-like approach tends to require continuous unpacking of black-boxed dialogues among research instrument features, surrounding circumstances and subjects' experiential backgrounds. Given the joint effect of subject awareness, motive inscrutability, and variability and holism of social settings, empirical evidence in social sciences often remains ambiguous, leaving much room for interpretation. It is therefore understandable that large bodies of social science evidence frequently coexist with enduring controversies on substantive issues. Certainly, philosophy of social science has revolved around the axis between naturalism and interpretation (Rosenberg 2015). But what follows from our inquiry is that a naturalistic approach to the social domain inevitably derives into a hefty dollop of interpretation. Evidences from a piecemeal, naturalistic social science are perhaps fated to become nothing but building blocks for relatively holistic and truly enlightening explanations or understandings.

From the above a somehow skeptical view seems to arise regarding the possibility that the validity challenges posed by social research, and examined in the previous sections, can be solved within the current normative, natural science-like framework shaping science. Current limits of our knowledge are admitted by some of the leading social scientists mentioned in previous sections. We are aware that setting limits to the future is a risky endeavor but, to the extent that such challenges remain partly insurmountable or intractable by more sophisticated tools or method refinements, two philosophical hints emerge. The first is distinguishing cases in which the above problems develop into pseudoscientific practices (hidden violation of scientific norms) from those where the result is just an un-scientific inquiry (neither violation of scientific norms nor satisfactory fulfilment of scientific requirements yet). The second consists in relaxing or redefining scientific norms (at least) for social research. Obviously, a proper and in-depth treatment of these issues would require a fine grained analysis of the different developments within social science, an analysis that would prove sensitive enough to both the ontological and methodological peculiarities of fields as diverse as econometrics, experimental economics, psychometrics, social psychology, social anthropology or socio-biology, to mention just a few examples. As such a fascinating task requires extensive discussion, we will have to leave it for future papers.

References

Akerlof, A. G. and Yellen, J. L. (1990) “The fair wage-effort hypothesis and unemployment”, *Quarterly Journal of Economics*, 105: 255–283.

Alm, J., G. H. McClelland, and W. D. Schulze. 1992. "Why do people pay taxes?" *Journal of Public Economics* 48 (1): 21-38.

APA (American Psychological Association, Committee on Test Standards). 1952. "Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal." *American Psychologist* 7: 461-465.

Angoff, W. H. 1988. "Validity: An Evolving Concept." In *Test validity*, eds. Howard Wainer and Henry I. Braun, 19-32. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bardsley, N. 2005. "Experimental Economics and the Artificiality of Alteration." *Journal of Economic Methodology* 12 (2): 239-251.

Barron, K. E., A. R. Brown, T. E. Egan, C. R. Gesualdi, and K. A. Marchuck. 2008. "Validity". In *21st Century Psychology: A Reference Handbook*, eds. S. F. Davis and W. Buskist, 55-64. Thousand Oaks, CA: Sage Publications.

Beullens, K., and G. Loosveldt. 2014. "Interviewer effects on latent constructs in survey research." *Journal of Survey Statistics and Methodology* 2 (4): 433-458.

Beullens, K., and G. Loosveldt. 2016. "Interviewer effects in the European Social Survey." *Survey Research Methods* 10 (2): 103-118.

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.

Binham, W. V. 1937. *Aptitudes and aptitude testing*. New York: Harper.

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.

Boumans, M. 2016. "Methodological ignorance: A comment on field experiments and methodological intolerance." *Journal of Economic Methodology* 23 (2): 139-146.

Camerer, C. F. 2015. "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." In *Handbook of Experimental Economic Methodology*, edited by G. R. Fréchette and A. Schotter, 249-295, Oxford: Oxford University Press.

Campbell, D. T. 1957. "Factors relevant to the validity of experiments in social settings." *Psychological Bulletin* 54: 297-312.

Campbell, D. T. 1986. "Relabeling internal and external validity for applied social scientists." In *Advances in quasi-experimental design and analysis*, edited by W. M. K. Trochim, 67-77. San Francisco: Jossey-Bass.

Carmines, E. G., and R. A. Zeller. 1979. *Reliability and Validity Assessment*. London: Sage Publications.

Carpenter, J., C. Connolly, and C. K. Myers. 2008. "Altruistic behavior in a representative dictator experiment." *Experimental Economics* 11 (3): 282-298.

Cordeiro-dos-Santos, A. C. 2006. *The Social Epistemology of Experimental Economics*. Thesis to obtain the degree of Doctor from the Erasmus University Rotterdam.

Cronbach, L. J. 1989. "Construct validation after thirty years." In *Intelligence: Measurement, theory, and public policy*, edited by R. E. Linn, 147-171. Urbana, IL: University of Illinois Press.

Cronbach, L. J., and P. E. Meehl. 1955. “Construct validity in psychological tests.” *Psychological Bulletin* 52: 281-302.

Davidson, D. 1963/1989. Actions, Reasons, and Causes. In *Essays on Actions and Events*, 3-19. Oxford: Clarendon Press.

Davidson, D. 1974/1989. Psychology as Philosophy. In *Essays on Actions and Events*, 229-239. Oxford: Clarendon Press.

Dietrich, F., and C. List. 2012. “Mentalism Versus Behaviorism in Economics: A Philosophy-of-Science Perspective.” *Munich Personal RePEc Archive*, URI:http://mpra.ub.uni-muenchen.de/id/eprint/43231

Duprè, J. 2016. “Social Science: City Center or Leafy Suburb”. *Philosophy of the Social Sciences* 46 (6): 548-564.

Favereau, J. 2016. “On the analogy between field experiments in economics and clinical trials in medicine.” *Journal of Economic Methodology* 23 (2): 203-222.

Franklin, A. 2005. *No Easy Answers: Science and the Pursuit of Knowledge*. Pittsburgh: University of Pittsburgh Press.

Galison, P. 1997. *Image and Logic: a Material Culture of Microphysics*. Chicago: University of Chicago Press.

García-Pérez, M. A. 2012. “Statistical conclusion validity: some common threats and simple remedies.” *Frontiers in Psychology* 3: 325.

Greenhill, M., Z. Leviston, R. Leonard, and I. Walker. 2014. “Assessing climate change beliefs: Response effects of question wording and response alternatives.” *Public Understanding of Science* 23 (8): 947-965.

Greenwood, J. D. 1982. “On the relation between laboratory experiments and social behavior: causal explanation and generalisation.” *Journal of the Theory of Social Behavior* 12: 225-249.

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. 2005. *The methodology of experimental economics*. New York, NY: Cambridge University Press.

Guala, F. 2017. “Experimental methodology on the move.” *Journal of Economic Methodology* 24 (1): 108-114.

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

Hacking, I. 1995/2011 “The looping effects of human kinds”. In *The philosophy of social science reader*, edited by D. Steel & F. Guala, 24-38, London: Routledge.

Hertwig, R., and A. Ortmann. 2001. “Experimental practices in economics: A methodological challenge for psychologists?” *Behavioral and Brain Sciences* 24: 383-451.

Hogarth, R. M. 2005. “The challenge of representative design in psychology and economics.” *Journal of Economic Methodology* 12 (2): 253-263.

Huff, T. E. 1982a. “On the Methodology of the Social Sciences: A Review Essay-Part II.” *Philosophy of the Social Sciences* 12 (1): 81-94.

Huff, T. E. 1982b. “On the Methodology of the Social Sciences: A Review Essay-Part III.” *Philosophy of the Social Sciences* 12 (2): 205-219.

Ivarola, L. 2017. “Socioeconomic processes as open-ended results. Beyond invariance knowledge for interventionist purposes”. *Theoria. An International Journal for Theory, History and Foundations of Science* 32 (2): 211-229.

Kagel, J. H. 2015. “Laboratory Experiments: The Lab in Relationship to Field Experiments, Field Data, and Economic Theory.” In *Handbook of Experimental Economic Methodology*, edited by G. R. Fréchette and A. Schotter, 339-359. Oxford: Oxford University Press.

Kane, M. 2006b. “In praise of pluralism. A comment on Borsboom.” *Psychometrika* 71 (3): 441-445.

Kelley, T. L. 1927. *Interpretation of educational measurement*. Yonkers-on-Hudson, NY: World Book Co.

Kim, J. 1988. “What is ‘Naturalized Epistemology’?” In *Epistemology: An Anthology*, edited by E. Sosa and J. Kim, 301-313. Malden, MA: Blackwell Publishing.

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.

Landry, C. E., A. Lange, J. A. List, M. K. Price, and N. G. Rupp. 2006. "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment." *Quarterly Journal of Economics* 121: 747-782.

Levin, I. P., Schneider, S. L., and Gaeth, G. J. 1998. "All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects." *Organizational Behavior and Human Decision Process* 76(2): 149-188.

Levitt, S. D., and J. A. List. 2009. "Field experiments in economics: The past, the present, and the future." *European Economic Review* 53: 1-18.

Mäki, U. 2008. "Scientific Realism and Ontology". In *The New Palgrave Dictionary of Economics*, 2nd ed., vol. 7, edited by S. N. Durlauf and L. E. Blume, 334-341. Basingstoke: Palgrave Macmillan.

Malone, E. L., J. J. Dooley, and J. A. Bradbury. 2010. "Moving from misinformation derived from public attitude surveys on carbon dioxide capture and storage towards realistic stakeholder involvement." *International Journal of Greenhouse Gas Control* 4 (1): 419-425.

Mayo, D. G. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: The University of Chicago Press.

McKenzie, C. R. M., and J. T. Wixted. 2001. "Participant skepticism: If you can't beat it, model it." *Behavioral and Brain Sciences* 24 (3): 424-425.

Messick., S. 1989. "Validity." In *Educational Measurement*, third edition, edited by R. L. Linn, 13-103. New York: American Council on Education, Macmillan Publishing Company.

Mitchell, S. 2009. “Complexity and Explanation in the Social Sciences.” In *Philosophy of the social sciences: philosophical theory and scientific practice*, edited by C. Mantzavinos, 130-145. Cambridge, New York: Cambridge University Press.

Morawetz, U. B., H. de Groote, and S. C. Kimenju. 2011. “Improving the Use of Experimental Auctions in Africa: Theory and Evidence” *Journal of Agricultural and Resource Economics* 36 (2): 263-279.

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.

Musgrave, A. 1981. “Unreal Assumptions in Economic Theory: The F-Twist Untwisted.” *Kyklos* 34: 377-387.

Pelham B. W., and H. Blanton. 2003. *Conducting Research in Psychology. Measuring the Weight of Smoke*. Belmont, Wadsworth: Thomson Learning.

Read, D. 2005. “Monetary incentives, what are they good for?.” *Journal of Economic Methodology* 12 (2): 265-276.

Rosenberg, A. 2015. *Philosophy of Social Science*, 5th edition. Boulder, CO: Westview Press.

Rosenberg, A. 2009. “If Economics is a Science, What Kind of a Science Is It?.” In *The Oxford Handbook of Philosophy of Economics*, edited by H. Kincaid and D. Ross, 55-67. Oxford: Oxford University Press.

Rosenboim, M., and T. Shavit. 2012. "Whose money is it anyway? Using prepaid incentives in experimental economics to create a natural environment." *Experimental Economics* 15 (1): 145-157.

Ross, D. 2010. "Why economic modelers can't exclude psychological processing variables." *Journal of Economic Methodology*, 17 (1): 87-92.

Rulon, P. J. 1946. "On the validity of educational tests." *Harvard Educational Review* 16: 290-296.

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

Schotter, A. 2008. "What's So Informative about Choice?" In *The foundations of positive and normative economics: a handbook*, edited by A. Caplin and A. Schotter, 70-94. Oxford: Oxford University Press.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sireci, S. G. 2009. "Packing and Unpacking Sources of Validity Evidence: History Repeats Itself Again." In *The Concept of Validity: Revisions, New Directions, and Applications*, ed. Robert W. Lissitz, 19-37. Charlotte, NC: Information Age Publishing.

Smith, V. L. 1982. "Microeconomic Systems as an Experimental Science." *American Economic Review* 72 (5): 923-955.

Smith, V. L. 2010. "Theory and experiment: What are the questions?" *Journal of Economic Behavior & Organization* 73: 3-15.

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.

Søberg, M. 2005. “The Duhem-Quine thesis and experimental economics: A reinterpretation.” *Journal of Economic Methodology* 12 (4): 581-597.

Steinle, F. 1997. “Entering New Fields: Exploratory Uses of Experimentation.” *Philosophy of Science*, 64 (Supplement): s65-s74.

Suppes, P. 1962. “Models of Data.” In *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, edited by E. Nagel, P. Suppes, and A. Tarski, 252-261. Stanford: Stanford University Press.

Suppes, P. 1982. “Problems of Causal Analysis in the Social Sciences.” *Epistemologia. Rivista Italiana di Filosofia della Scienza*. V: 239-250.

Suppes, P. 1985. “Explaining the unpredictable.” *Erkenntnis*. 22: 187-195.

Suppes, P. 2008. “A Revised Agenda for Philosophy of Mind (and Brain).” In *Lauener Library of Analytical Philosophy*, 19-51. Frankfurt: Ontos Verlag.

Tagiew, R. and Ignatov, D. I. 2014. “Reciprocity in Gift-Exchange-Games.” *Cornell University Library*, arXiv: 1402.5593v1 [cs.AI]. Accessed July 15, 2017. https://arxiv.org/abs/1402.5593v1.

Thurstone, L. L. 1932. *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.

Tyler, T. R. and Amodio, D. 2015. "Psychology and Economics: Areas of Convergence and Difference." In *Handbook of Experimental Economic Methodology*, edited by G. R. Fréchette and A. Schotter, 181-196, Oxford: Oxford University Press.

Van Bochove, M., J. Burgers, A. Geurts, W. de Koster, and J. van der Waal. 2015. "Questioning Ethnic Identity: Interviewer Effects in Research About Immigrants' Self-Definition and Feelings of Belonging." *Journal of Cross-Cultural Psychology* 46 (5): 652-666.

Weber, M. 1949 *The Methodology of the Social Sciences*, trans. E. A. Shils and H. A. Finch. Glencoe, IL: The Free Press.

Wiggins, J. A. 1968. "Hypothesis Validity and Experimental Laboratory Methods." In *Methodology in Social Research*, edited by H. M. Blalock, Jr. and A. B. Blalock, 390-427. New York: McGraw-Hill.

Yount, K. M., N. Halim, M. Hynes, and E. R. Hillman. 2011. "Response effects to attitudinal questions about domestic violence against women: A comparative perspective." *Social Science Research* 40 (3): 873-884.

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.