Enlarged Empirical Economics and the Quest for Validity: Facing the Ontological Intricacies of the Social Domain

Article *in* Journal of Institutional Studies · March 2019

DOI: 10.17835/2076-6297.2019.11.1.020-040

CITATIONS
0

READS
110

2 authors:

José Caamaño-Alegre
University of Santiago de Compostela
47 PUBLICATIONS **144** CITATIONS

SEE PROFILE

Maria Caamano-Alegre
Universidad de Valladolid
20 PUBLICATIONS **22** CITATIONS

SEE PROFILE

ENLARGED EMPIRICAL ECONOMICS AND THE QUEST FOR VALIDITY: FACING THE ONTOLOGICAL INTRICACIES OF THE SOCIAL DOMAIN[1]

JOSE CAAMAÑO-ALEGRE,

Ph.D., Associate Professor,
Department of Applied Economics,
University of Santiago de Compostela (Spain),
e-mail: jose.caamano@usc.es;

MARIA CAAMAÑO-ALEGRE,

Ph.D., Associate Professor,
Department of Philosophy,
University of Valladolid (Spain),
e-mail: mariac@fyl.uva.es

This article focuses on the impact that the recent widening of empirical economics has on the quest for validity in this field. We begin by summarizing the continuous evolution from a primarily deductive economics to a more empirical one, especially emphasizing the broader experimental and survey-based evidence. Although these developments pave the way for an economics with greater empirical support, they also bring into this field the same validity concerns that mainstream economists naively thought to be avoidable (i. e., concerns with the external validity of experiments and with "test validity" issues largely addressed in other social sciences). We show how, ultimately, such developments force economists to confront some serious challenges and limitations in the quest for validity arising from four ontological peculiarities of the social domain: 1) the awareness of the inquiry on the part of the subject being studied; 2) the lack of relevant structural homogeneity between individuals' shared psychological properties; 3) actions holistic dependence on the individuals' complete past; and 4) the variable and holistic nature of cultural, conventionally mediated forms of interaction. We finally argue that openly acknowledging these problems would help economists to tone down their scientificity claims and avoid pseudo-scientific practices like endorsing assumptions refuted by experience.

Keywords: *external validity; experimental economics; test validity; survey research; pseudoscience.*

JEL: *A12, B41, C83, C90*

Introduction

In the last decades, empirical economics has evolved from a limited econometric testing of poorly specified models to an enlarged and more pluralistic enterprise. The spectrum of evidence currently gathered by economists does not only include sounder econometric findings obtained from enriched models grounded on ever larger datasets, but also encompasses experimental and quasi-experimental results, as well as a variety of newly generated survey data. The label "empirical turn/shift in economics" summarizes an increasingly extended view *(Boettke, Leeson* and *Smith, 2008; Einav* and *Levin, 2014; Hausman, 2018)*, supported by bibliometric evidence *(Hamermesh, 2013; Biddle* and *Hamermesh, 2017; Angrist, Azoulay*

[1] We are thankful to Don Ross for extremely helpful comments on an earlier version of this paper and for generously sharing his rich insights into different issues addressed in our work. This research was financially supported by the research projects "Models and Theories in Physical, Biological, and Social Sciences" (PICT-2014-1741, ANPCyT, Argentina), "Political and Economic Consequences of the Descentralization" (CSO2013-47023-C2-2-R, Spanish Ministry of Economy and Competitiveness).

and *Ellison, 2017)*, although the very adequacy of the label has recently been questioned from different stances *(Syll, 2016; Backhouse* and *Cherrier, 2017; Mäki, 2018)*. Rather than accurately characterizing or sizing up this phenomenon, however, our purpose here is to show how the recent enlargement of empirical economics poses a wide range of validity challenges, typical of social sciences, to mainstream economics. This should be emphasized, since mainstream economics tended to be illusorily seen, by economists, as unaffected by such issues.

Specifically, we focus on a set of challenges arisen from the ontological peculiarities of the social domain. Acknowledging that several validity challenges in economic research are ontologically rooted necessarily imply, neither that all of them are insurmountable, nor even that they are difficult to overcome. We are fully aware that some specificities of the social realm are behind many kinds of everyday obstacles and confounds that empirical economists deal with, with a variable degree of success. However, we do not endorse the orthodox discourse that assumes, a priori and uncritically, that the continual and widespread failures in naturalistic[2] mainstream economics are mere "hallmarks of the activity of science rather than a reason to acknowledge that the entire enterprise may fail to be scientific" *(Morgan* and *Patomäki, 2017, p. 1393)*. Contrary to this soothing and self-indulgent view, our analysis casts some doubts on any expectation that economics achieves the same validity warrant in empirical research than that guaranteed in physics or even medical sciences. Rather, it invites economists to tone down their scientificity claims and avoid dogmatism by critically reflecting on the methodological adequacy of empirical economics to deal with the ontological features of the social domain.

In the following sections, we illustrate the aforementioned validity challenges with examples provided by the literature on economic experimentation and survey use in economics. Our especial concern with these two research tools, putting aside the most prominent one (econometrics), is justified on the following grounds: (1) the growing use of experimental evidence and wider variety of survey data in economics; (2) their contrasting character, as survey research is less closely connected to natural science methodology than experiments are; and (3) the fact that, within social sciences, both fields have long pursued intellectual traditions regarding validity.

Ultimately, our approach vindicates a role for philosophy in discussing validity issues and, therefore, to some extent our view departs from the traditional account of such issues within experimental economics, where they have been understood as purely empirical rather than (also) as a philosophical matter *(Heukelom, 2011, p. 20)*. We, however, do not appeal to the "metaphysical arguments about the nature of economic and social reality" that Guala (*2008, p. 34*) deems "of little utility" to this respect. Instead of postulating or refuting the questionable existence of universal laws —or, at least, of *tendency laws*— in economics, we pinpoint some uncontroversial peculiarities of the social domain which prove potentially threatening for validity. This enables us to put the validity problems in a broader context by rediscovering their deep ontological roots and philosophical implications. Among the latter, we highlight those concerning the debate on the scientific status of economics, as well as on the risk of allowing pseudo-scientific practices and the efficient steering of the collective effort within this sphere.

The remaining of this article is structured as follows: section 1 deals with the enlargement of the empirical economics, section 2 is devoted to rethinking validity in this field, and section 3 focuses on the validity problems arising from some ontological peculiarities of the social domain. The general significance of these issues for the scientific status of economics and the

[2] According to Rosenberg (*2015, p. 30–31*), "naturalist" are all those social scientists commited to mehods adapted from the natural sciences and that find no difficulty in reconciling prediction and interpretation. In his own words: "They believe that there is a causal theory of human behavior and that we can uncover models, regularities, and perhaps eventually laws that will enable us to predict human action".

role of philosophy in elucidating them are briefly discussed in section 4. In the last section we make some concluding remarks.

1. From a primarily deductive economics to a more empirical one

From the mid-twentieth century to the mid-eighties mainstream economics revolved around highly abstract and deductive issues, since it was primarily focused on extending the neoclassical paradigm mainly by proving microeconomic and welfare economics theorems, building, at the same time, marginalist models of consumer and producer behavior, as well as rational expectations and general equilibrium models. It must be emphasized that, over that period, the deductive, empirically minimalist approach vigorously endorsed by Mill, Cairnes and Robbins in their days took precedence over the "much wiser and better balanced practice and principles of Smith, Jevons, Marshall and Keynes" (*Hutchison, 1998, p. 44*). The awareness of this fact is certainly behind one of the most quoted statements by Daniel Hausman *(1992, p. 1*): "The method of economics is deductive, and confidence in the implications of economics derives from confidence in its axioms rather than from testing their implications."

Before the end of the eighties economic experiments played a very marginal role, to the point that laboratory environments were regarded by many economists as being outside the domain of economic theories, which would be intended to explain real behavior rather than lab behavior *(Cubitt, 2005, p. 198*). Survey data, on the contrary, were already used for many purposes in economics during the sixties *(Morgan, 1967*), even if survey research remained "officially" marginalized in the discipline on the basis that people sometimes do not tell the whole truth (*McCloskey* and *Ziliak, 2001, p. 164*). Furthermore, declared behavior was also considered as foreign to the domain of economic theories, for actions, as opposed to testimonies, were assumed to be the real focus of study *(Boulier* and *Goldfarb, 1998, p. 17*). This being so, traditional econometric testing on supposedly "hard" data became the standard practice, and even the almost only legitimate way for empirical economics. However, theoretical shortcomings and tractability issues severely restricted econometric analyses, which tended to rely on institutionally, and behaviorally poor, ill specified models and quite limited datasets. This kind of approach thus provided a very partial and insufficiently valid empirical support for economic science, as it proved unable to satisfactorily accommodate localism and control confounds.

The above picture has substantially changed over the last decades. As Ross and Kincaid have pointed out, the expansion in computational capacity, together with the invention of numerous econometric techniques, has resulted in a wide variety of methodological improvements. Some of them concern the modeling of richer causal hypotheses, while others are related to the more robust testing of models by combining wider ranges of tests and handling larger sets of data (*Ross* and *Kincaid, 2009, p. 12*). These technical advances make it possible to accommodate theoretical developments from institutional and behavioral economics into enriched empirical model specifications. Our paper is focused on two other key enlargements of empirical economics: first, the sharp increase in economic experimentation starting from the beginning of the present century; and second, the extended range of survey data used in economic research, gradually departing from the traditional "just the facts" survey data. We thus put aside, not only the obviously dominant contributions from econometrics, but also other research tools like cases studies, increasingly used in explaining, for instance, economic growth —see Rodrick (*2003*) and, for a methodological approach, Perraton (*2011*) and Ruzzene (*2014*), among others.

Regarding economic experimentation, List *(2009, p. 440*) documents the increase in the number of experimental publications along the previous sixty years, in the top 15 economic journals as established by Kalaitzidakis, Mamueas, and Stengos (*2003*). In his graphic display there are two distinguishable turning points. A first and softer acceleration is recorded in

the mid-eighties, but the real uprush corresponds to the eight first years of the XXI century, with the number of articles being multiplied by four. Furthermore, the variety of topics covered by economic experiments quickly rises along the same period, and institutions become ever more engaged in supporting and financing field experiments. Circumscribing the analysis to all applied micro articles published in the top 5 economics journals in 1951–55, 1974–75 and 2007–8, Biddle and Hamermesh *(2017)* corroborates the expansion of the "experimentalist paradigm" in the first decade of the current century. Bibliometric evidence gathered by Hamermesh *(2013, p. 168)* points to a similar conclusion. He notices that the proportion of experiments in the articles published in the 3 leading economics journals goes from 0.8% in 1983 to 3.7% in 1993 and 8.2% in 2011.

Regarding survey data, Presser *(1984)* and Saris and Gallhofer *(2007, p. 2)* report an increasing percentage of articles with use of survey data in 3 main economic journals from 1949–50 to 1994–95. Nevertheless, their operative definition of survey data includes also statistical data collected for official statistics, which are at least partially based on survey research and on administrative records. More recently, Chetty *(2012)* graphically displays a substantial drop in the percentage of micro-data based articles using pre-existing survey data in 4 leading economic journals, from 1980 to 2010. By "pre-existing" surveys he means micro surveys such as the Current Population Survey (CPS) or the Survey of Income and Program Participation (SIPP) and, consequently, those surveys designed by the own authors of the articles are not included[3]. This relative decline in the use of such kind of survey data could be due to the perception that, in comparison to traditional survey data sources, administrative data offer much larger sample sizes while having far fewer problems with attrition, non-response, and measurement error *(Card, Chetty and Feldstein, 2010)*.

Unfortunately, comprehensive figures on the use of survey data in economic research appear not to be available by now. In a provocative contribution to the *Mises Wire*, Jonathan Newman *(2017)* holds the view that survey data are more popular than ever among economists, and plots the National Longitudinal Survey citations by year since 1968 to proxy the growth in popularity of surveys. A decade before, Porter, Ketels and Delgado *(2007, p. 61)* note that, despite skepticism among some researchers, the use of survey data in economic analysis is becoming increasingly more common. It is undeniable that the list of survey datasets available for social scientists is continuously increasing[4], and the number of them used in the different economic subfields has become overwhelming. Within some of the subfields, researchers report a dramatic growth in their use, as it is the case, for instance, in international political economy (*Jensen, Mukherjee* and *Bernhard*, *2014*). In addition, economists have been gradually enlarging the range of survey data that they use. Moving ahead from the traditional "just the facts" surveys, they began to use surveys also to ask for predictions of future events and for measurements of hard-to-gauge variables *(Boulier* and *Goldfarb, 1998, p. 6*). Moreover, other survey categories which were traditionally rejected by them have also been added to the methodological repertoire —like in the case of questionnaires about mental states, now in use by behavioral economists (*Fudenberg, 2006*).

2. Rethinking validity in empirical economics

Not surprisingly, the enlargement in empirical economics depicted in the previous section has prompted a reconsideration of validity in this field. Therefore, we briefly discuss next how economists and economic methodologists are rethinking validity after the period going from the mid-twentieth century to the mid-eighties, when academic economics was primarily deductive and empirical validity was taken for granted on theoretical grounds. The rather limited validity of former empirical economics is now highlighted and, at the

[3] Moreover, his sample excludes studies whose primary data source is from developing countries.

[4] Gathering data on US government surveys only, Presser and McCullogh *(2011, p. 1023)* find that these surveys "increased from 1984 to 2004 at a rate many times greater than the increase in the US population."

same time, its recent enlargement is taken as an opportunity to seek a more valid empirical basis for economics. Against this background, we finally address the current approach to validity in empirical economics.

2.1. Deductive economics and empirical validity taken for granted on theoretical grounds

For a long time, mainstream economists tended to elude the kind of validity challenges in empirical research largely confronted in other social sciences. The very deductivism endorsed by economists, with its confidence in axioms and its appellations to folk psychology and commonsense evidence (*Rosenberg, 1992*), ended up by neglecting the importance of validity in empirical research. The structural approach to causal inference (*Heckman, 2000, 2005, 2008*), which was dominant for decades, was intended to solve the problem of confounding by applying economic theory to model the causal structure relevant to the question of interest. Within this framework, validity typically relied on a number of assumptions in terms of errors in regression functions, and the bundling of these assumptions made it more difficult to asses their plausibility (*Imbens* and *Wooldridge, 2009, p. 11*). Theoretical devices like those of "representative agent" and "revealed preference" also contributed to "hide" those validity challenges related to the attempt at gaining direct access to psychological variables.

Such devices bespeak that economists were traditionally focused on the abstract sensitivity to changes in incentives shown by the average observed behavior of a given population (*Ross, 2011*); an approach then projected on their empirical research, where the concern with accommodating agent heterogeneity is a quite recent and still ongoing phenomenon. Even in the marginal field of experimental economics, classical lab experiments to test theories were designed to enforce all the theoretical assumptions (*Lucking-Reiley, 1999, p. 1078; apud. Boumans, 2016, p. 143*), and most of these experiments were conducted on very homogeneous samples (usually students from the same background) and with insufficient information on potentially important socioeconomic characteristics (*Siedler and Sonnenberg, 2010, p. 2–3*). Although the issue of "parallelism" between results inside and outside lab was acknowledged among pioneer economic experimenters, their conviction that lab markets are "real" markets led them to establish a favorable presumption that parallelism held, shifting the burden of proof on anybody that put it into question (*Siakantaris, 2000, p. 269*). That is why, to some extent, validity in economic experiments tended to be taken as granted by its advocates and denied by its detractors, depending on their opposed views on the relationship between experimental settings and real markets. Both sides tended to see validity as an all-or-nothing question.

On other hand, the above mentioned renounce to "direct" access to psychological variables in favour of behavioral data, together with the poverty of the econometric models, allowed the prominent use of single item, objective indicators with supposedly no error in measurement to measure constructs in structural models (*Venaik, 2007*). Consequently, economists avoided the "test validity" issues typically addressed in psychometrics and widely faced by social scientists in general. As explained in Backhouse (*2010, p. 167–168*), traditional economic statistics tend either to be viewed as "natural" measurements, like the number of tons of steel produced in a year, or to involve a substantial conceptual input from economists and others, like the gross domestic product. In the first case, validity tends to be considered unproblematic and, in the second, it tends to be assumed on the grounds of a standardized framework backed by official institutions.

2.2. Enlarged empirical economics and the pursue of validity

On an optimistic note, the recent developments set out in section 1 can be viewed as promising advances towards an *empirically based* and *methodologically pluralistic* economics, as the one advocated by Herbert Simon (*1997*). Despite the validity problems

associated with the enlarged empirical economics of our time, the situation seems better than the one during the deductivist period, characterized by its validity elussiveness. By applying experimental methods, economists have improved their ability to elucidate causal relationships, thereby making better progress than when they were just relying on traditional econometric inferences from naturally-ocurring data. Statistical techniques used by them on the latter kind of data can help to determine correlations between economic variables, but often such techniques do not suffice to draw specific causal inferences from the spontaneous variations found in the data, as the ideal conditions typical of laboratory settings are rarely found in the wild unless in so-called "natural experiments" *(Guala, 2008, p. 18)*.

As for survey data, they make it possible for economists to capture the "informed judgments of the actual participants" in the economies, since the stance of decision makers, which ultimately determines economic activity, is reflected by such data *(Porter et al., 2007, p. 61)*. More generally, survey data contribute to reinforce control and validity in economics by providing empirical models with psychological and institutional variables, which are necessary in order to account for the "frames" or "partial representations of their opportunity sets" on which people base their choices *(Ross, 2010, p. 88–89)*. Aditionally, survey data are sometimes used to validate experimental designs or results *(Carpenter, Connolly* and *Myers, 2008; Dohmen* and *Falk, 2011; Bradler 2015)*, the same way that experiments are sometimes used to validate survey instruments *(e.g. Kling, Phaneuf* and *Zhao, 2012, p. 13; Vischer, Dohmen* and *Falk, 2013; Falk, Becker* and *Dohmen, 2016, 2018)*. Mutual validation is openly pursued in some cases, like in Naef's and Schupp's *(2009)* use of the results of a trust experiment to validate their own survey trust scale, and, conversely, their use of the responses to selected items of the same survey to validate their experimental results (provided that a common sample is taken for both their experiment and survey).

Given that econometric analyses, experiments and survey research have each one its own strengths and limitations, taking advantage of their synergies and/or complementarities may very well reveal as a necessary strategy to attain validity in empirical economics. While *evidential variety* in a narrow sense, i.e. use of multiple means to determine a property of interest or confirm specific propositions, may still be a source of credibility for causal inference *(Claveau, 2011)*, evidential validity in a wider sense, i.e. aimed at understanding a complex object of study, has a greater potential. After digging into this issue, Downward and Mearman *(2007)* advocate a *mixed-methods triangulation* in economic research, an approach that favors the combination of different insights into the same phenomenon, thereby reorienting economics towards social science.

2.3. The current approach to validity in empirical economics

As suggested earlier, progress towards methodological pluralism brings into empirical economics the same validity challenges that have been largely confronted in other social sciences, but to which mainstream economists thought to be immune. In this vein, Angrist and Pischke's *(2010)* "credibility revolution" has emphasized sharpness in quasi-experimental design as crucial for valid causal inferences, in contrast to the emphasis on theoretical soundness germane to the structural approach. This is leading to a shift "from models to methods" in the empirical economist's toolkit *(Panhans* and *Singleton, 2017)*, and paving the road for empirical economists to adopt the campbellian notions of validity *(Shadish, Cook* and *Campbell, 2002)* commonly used by other social scientists. Indeed, Campbell's *(1957)* seminal distinction between internal and external validity has already been adopted by economic experimenters and methodologists. Heukelom *(2011, p. 21)* explains this adoption as follows:

"(…) as experimental economics was growing, but its position anything but secured, external validity and later internal validity were introduced to convince a skeptical audience of economists who did not believe in the 'reality' of experiments."

Within the first decade of the 21st century, the rise of experimental economics was causing an increasing concern with external validity, which became more relevant as a greater number of experiments were designed for purposes other than testing theories[5]. Several suggestions have been made to face the external validity challenge in experimental economics, from reproducing certain features of the target system in the experimental system *(Guala, 2003)* or doing stress tests in the lab *(Schram, 2005)*, to randomly sampling target situations *(Hogarth, 2005)* and combining laboratory with field evidence *(Guala, 2005)*. With respect to the latter, it is interesting to note that three in-between types of field experiments have been defined in the continuum from the highest control (achieved in lab experiments) to the highest closeness to reality, namely, *artefactual*, *framed*, and *natural* field experiments *(Harrison* and *List, 2004)*. Given its middle-ground location in the continuum, it is not surprising that field experimental research in economics has recently explored the possible significant connections between laboratory data and naturally-occurring data *(Levitt* and *List, 2009, p. 15)*.

According to Heukelom *(2011)*, these latter developments would have been backed by adopting definitions of internal and external validity stricter than the ones used in experimental psychology —therefore drawing a sharp line between an "inside" world of the experiment and an "outside", real world—, and by understanding the relationship between them as a strict trade-off. Jiménez-Buedo and Miller *(2010)*, however, have questioned this alleged trade-off, and Jiménez-Buedo *(2011)* has argued not only that the internal/external dichotomy entails serious conceptual problems, but also that it does not provide the most useful *practical* guides for scientists like behavioral economics experimenters, who are especially concerned with establishing the *correct* theoretical interpretation of the behavior they observe. She illustrates this point by remarking that, in most ultimatum game experiments, the most pressing questions regarding a given experimental exercise are related, neither to confounds, nor to generalizability. As she remarks, a recurring question revolves around whether:

"(…) the same stylized facts, i.e. robust results in the ultimatum game regarding modal offers and rejection rates, can be interpreted as either coming from an individual preference for altruism (or inequity aversion) or rather, from the desire to adhere to a social norm dictating a particular behavioural pattern conditional on the compliance of other agents." *(Jiménez-Buedo, 2011, p. 279)*

In parallel to this evolution in the case of increasing economic experimentation, the greater use of survey data to capture institutional and behavioral factors also makes economists gradually more concerned with the same validity issues confronted in test validity literature *(Wainer* and *Braun, 1988; Lissitz, 2009; Markus* and *Borsboom, 2013; Price, 2017)*. Economists' new uses of survey data indeed go beyond their traditional resort to single item, objective indicators with supposedly no error in measurement and entail putting at least partially aside their "perennial reluctance to operate with the kinds of highly abstract constructs most common in psychology" *(Ross* and *Kincaid, 2009, p. 26–27)*.

In institutional, political economy, for instance, authors like Alt and Lassen *(2006)* or Aaskoven *(2016)* employ multi-item survey data from budget officers to measure a highly abstract and complex construct such as fiscal transparency. It could certainly be argued that the items' content falls in the category of "just the facts" survey, which only requires recounting some facts that are presumably known by the respondent and, therefore, is acknowledged by economists as providing data with a relatively high "fact status" *(Boulier* and *Goldfarb, 1998, p. 5–6)*. Nevertheless, a couple of considerations must be taken into account here. First, even "just the facts" survey data traditionally used by economists show error in measurement, as Meyer, Mok, and Sullivan *(2015)* show in detail for the emblematic case of household surveys. Second, even if the items of a survey on budget transparency

[5] For instance, purposes such as motivating the development of new theories *(Schram, 2005, p. 234)*.

may reveal relevant dimensions of the construct, some of the items may actually be more indicative of another construct (for instance, financial management sophistication). Fiscal transparency surveys may, therefore, raise the typical issues regarding construct validity largely addressed in test validity literature, *viz.*, construct underrepresentation and construct-irrelevant variance. The most usual approach among those economists that do examine the validity of their fiscal transparency measures does not go beyond loosely linking such measures to any established framework (IMF, OECD…) and estimating their correlations with other transparency measures or outcome indicators. Now, surveys on fiscal transparency are by no means an isolated case, for broad institutional concepts such as "voice and accountability", "government effectiveness", or "rule of law" have often been measured on the basis of their implicit and changing definitions obtained from the available surveys *(Voigt, 2013, p. 22)*, without any serious construct validation.

In behavioral economics, current usage of surveys on mental states requires measuring similarly abstract constructs with both multi-item and remarkably subjective indicators. Even though surveys on subjective well-being, for instance, often consist of a single item on global life satisfaction, higher reliability and validity can be obtained by a multi-item scale *(Powdthavee, 2007, p. 56)*. Current surveys usually include separate items or scales to measure affective and cognitive well-being, which have been revealed as separable and additive constructs *(Powdthavee, 2015, p. 315)*. As already said, the whole range of issues regarding construct validity and test validity in general are involved here. Economists' most common approach to validity assessment has been grounded on the pattern of correlations between declared well-being and other characteristics of individuals, together with its capability to predict future outcomes. Kahneman and Krueger *(2006, p. 9)* have examined the relevant literature and gathered a number such correlates[6]. Krueger and Stone (*2014*) add some more, but also highlight several pending issues of validity and comparability. From a more critical stance, Bond and Lang *(2018, 2019)* object that key results from the happiness literature are reached from parametric approaches that rely on the assumption that all individuals report their happiness in the same way, and this is rejected by both authors in all cases in which they test it. Ultimately, these and other surveys on mental states entail facing, within economics, the same validity challenges as the ones traditionally confronted in neighboring disciplines like marketing and management, where researchers mainly use multi-item, subjective indicators with inherent error in measurement *(Venaik, 2007)*. Consequently, issues like question design, reliability and validation, which have been for long studied in these and other disciplines, should now be incorporated into the economists' academic curricula *(Boulier* and *Goldfarb, 1998, p. 15–16)*.

3. Validity challenges arisen from some traits of the social realm

Researchers working within this enlarged empirical economics are more openly confronted with the perennial, controversial question concerning the limits of the quest for validity in empirical social research and, ultimately, with the scientific status of economics as a social science. Our approach to these issues emphasizes those ubiquitous challenges arising in the new empirical economics due to the ontological peculiarities of the social domain. In particular, we discuss the methodological difficulties caused by the four following factors:

(1) the awareness of the inquiry on the part of the subject being studied *(Wiggins, 1968)*;

(2) the lack of relevant structural homogeneity between individuals with respect to their shared psychological properties *(Suppes, 1982, p. 246–247; Shulman* and *Shapiro, 2009: 125, Borsboom, Cramer* and *Kievit, 2009, p. 157–158)*;

[6] Namely: smiling frequency; smiling with the eyes ("unfakeable smile"); ratings of one's happiness made by friends; frequent verbal expressions of positive emotions; sociability and extraversion; sleep quality; happiness of close relatives; self-reported health; high income, and high income rank in a reference group; active involvement in religion; recent positive changes of circumstances (increased income, marriage).

(3) actions' holistic dependence on the individuals' complete past *(Suppes, 1982, p. 247–248)*; and

(4) the highly variable and holistic nature of cultural and conventionally mediated forms of socioeconomic interaction *(Rosenberg, 2009, p. 62–66)*.

In the three following subsections, we show how these factors have a direct bearing on the quest for validity. To this end, we apply the Shadish, Cook, and Campbell's (*2002*) taxonomy, although subsuming statistical conclusion validity under internal validity[7]. We will, therefore, go beyond the internal/external validity dichotomy to add construct validity as a tertium genus.

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

In experimental economics, the subjects' awareness of different aspects concerning the experimenter, the manipulations and measurements of an experiment may seriously threaten internal validity in the form of different sorts of noises and confounds. Awareness of the artificiality of the experimental setting, on one hand, may dramatically jeopardize external validity. Construct validity, in turn, may be affected by misleading evidence stemming from the subjects' recognition of different aspects concerning the experiment or research procedure. Let us consider, as a first example, the finding by Hoffman, McCabe, and Smith *(1996)* that social distance or isolation of the subjects from the experimenter in dictator games provokes a shift toward lower offers —a case of the so called *experimenter effects*. A second example is related to a problem arising from the risk that subjects' have low involvement with the independent variables due to their lack of attention, lack of motivation, or to their awareness of other variables. The main tool employed by economists in dealing with such risk (i.e., financial incentives) may matter more in some areas than in others, and, putting that aside, it does not even guarantee the achievement of *cognitive exertion*, *motivational focus*, and *emotional triggers* through which it would have the intended effect *(Read 2005)*. As a third example, we can mention the serious difficulties created by the practice of informed consent, as it opens the door to a strategic or distorted response on the side of the experimental subject due to the latter's beliefs about the experimental hypotheses. Levitt and List *(2009, p. 15)* point directly to this problem as they state:

"For example, if one were interested in exploring whether, and to what extent, race or gender influence the prices that buyers pay for used cars, it would seem difficult to measure accurately the degree of discrimination among used car dealers who know that they are taking part in an experiment."

Some thought-provoking examples concerning the problem of how artificiality affects the subjects can be found in Bardsley's *(2005)* insightful approach. He criticizes, for instance, that Alm *et al.*'s *(1992)* experiment on tax evasion is silent on the fact that assuming a real-world civic or legal duty to pay taxes is not the same as recognizing a duty to be honest to experimenters in labs. Similarly, he argues against Ball *et al.*'s *(2001)* claim that social status —defined as "a ranking in a hierarchy that is socially recognized"— was implemented in their experiment, by objecting that there is nothing in the latter that is recognized as a ranking in a hierarchy outside of the experiment[8]. He also rejects Mason *et al.*'s *(1992)* conviction that their findings were about duopolies instead of mere interpersonal strategic choice, arguing that no individual in the experiment would have seen themselves as acting in the capacity of a manager, employee or representative of a firm[9]. In a more recent contribution, Jimenez-

[7] "It is interesting to note —Jiménez-Buedo *(2011, p.. 281)* states— that Shadish et al. have grouped together statistical conclusion and internal validity, opposing them to construct and external validity, suggesting the existence of a broader 'internal' and 'external' validity dichotomy subsuming the dyad under the four kinds validity categorization."

[8] Bardsley vigorously contends that "if any status was imparted, it might have been analogous to that of a 'teacher's pet', with few conceivable real-market consequences."

[9] Where were the required contractual or role obligations to an institution with legal obligations in turn to shareholders? The problem would have remained if the subjects had been businessmen, so it is not one of a subject pool effect.

Buedo and Guala *(2016)* try to disentangle the skein of artificiality by distinguishing three meanings of the term (subjects' awareness of being studied, unfamiliarity with the experimental tasks, and gap between the experimental setting and the target situation), which leads them to conclude that the relationship between artificiality, demand effects, and external validity is far from being univocal.

Survey use in economics is affected by the same kind of problems we have been discussing. *Interviewer effects* and *reactive measurement* may be respectively prompted by the subject's awareness of the interviewer's presence and of the procedure. The characteristics of the interviewer may have an effect on the degree of underreporting in household surveys, leading to item nonresponse, measurement error conditional on responding, or both *(Meyer, Mok* and *Sullivan, 2015, p. 219)*. An interviewer effect may cause a non-sampling bias in the subjective well-being responses if, for instance, people tend to overstate their happiness level in order to make a good impression to the interviewer *(Powdthavee, 2007, p. 57)*. The interviewer effect, together with respondents' beliefs about the questions under inquire, may make them reluctant to admit things like a racial prejudice, an unfulfilled civic duty, or even a lack of an attitude towards whatever they are asked *(Bertrand* and *Mullainathan, 2001, p. 68)*. Respondents' possible unwillingness to report or to do it rightly was soon recognized as a source of error, one related to the risk of them having low involvement, lack of attention or motivation *(Morgan, 1967, p. 248–249)*. In fact, conventional economic reasoning would lead us to assume that a respondent, as a *homo œconomicus*, might refuse to cooperate, or might even give misleading answers if he had no incentive to respond truthfully or thoughtfully *(Blinder, 1991, p. 90)*. These concerns about non-response and measurement errors have thus been largely addressed within certain areas of economics that are especially dependent on survey data. Already by the mid-twentieth century, George Katona, the director of the Survey of Consumer Finance conducted within the Survey Research Center's Economic Behavior Program at the University of Michigan, persuaded the Federal Reserve Board that, to be more successful in extracting financial information from consumers, the survey instrument should have "a few 'can-opener' questions" —by this he meant that the questions were easy, made the respondent feel comfortable, and provided relevant information about future behavior *(Juster, 2004, p. 120)*.

It is interesting to note that, despite having elicited a wider approval from economists, "just the facts" survey data are also vulnerable to some serious problems concerning how the respondents' awareness hinders reliability and validity. To illustrate this, let us consider the Newman's *(2017)* example of self-reported high school grades in the US National Longitudinal Surveys, where there is a binge of high grades that does not correspond to the actual data collected from the official transcripts of many of the respondents. Although this phenomenon might be partly explained by a desire to make a good impression on the interviewer, therefore giving rise to biases in any further statistical analysis —as the errors are not random but correlated with certain personal and peer characteristics— the problem is not only that interviewees may lie or refuse to answer questions, but also that measurements of important phenomena may be extremely sensitive to the way in which questions are asked (*Boulier* and *Goldfarb, 1998)*. For example, the more probing the style of the labor force surveys (*vis-à-vis* the census), the narrower the count of the unemployed. Another case in point is that of the respondents' understating of their actual expenditures on alcoholic beverages in the Survey of Consumer Expenditures[10]. These threats to validity may become more serious as economists increase their use of other kinds of survey data, like those measuring the value of goods and services not traded in the market. A well-known example is that of the contingent valuation surveys, where respondents tend to give strategic responses which are influenced by knowledge on how the survey results could be used in policy decisions.

[10] So-called "bogus pipeline" experiments have certainly shown that subjects who are led to believe that inaccurate answers can be identified (say, by means of a lie detector) are more likely to give answers admitting a socially disapproved behavior such as drinking alcoholic beverages.

3.2. The problem of identifying motives as causes of behavior

Factors (2) and (3) above are both involved in the inscrutability of motives as causes of behavior. The lack of relevant correlations between internal structure and external effects, which are the building blocks of natural science, severely limits the empirical support for theoretical constructs as well as the very possibility of measuring them. This leaves too much room for construct invalidity occurring in the form of construct irrelevant variance, since the same dependent variables may be easily ascribable to different, highly conjectural theoretical constructs. As pointed out by Schotter (*2008*), the use of choice data to make inferences about preferences, costs, beliefs, and other unobservable, psychological variables has been a common practice in empirical economics. As a result, an ambiguous observed economic behavior constitutes the final empirical basis for causal inferences about unobservable variables.

The founding figure of experimental economics, Nobel prized Vernon L. Smith *(2010, p. 3–4)* recognizes that, regardless of whether experimental observations are in accord with the actions we predict, the inference of truth from observation is intrinsically ambiguous. Put in his own words:

"A procedure used to test a theory that leads to correct predictions is not 'right,' nor is one that leads to incorrect predictions, 'wrong.' Rather, we need to ask what the totality of this evidence tells us about the maintained (auxiliary) assumptions of theory and its test framework." *(Smith, 2010, p. 6).*

In game theory experiments, for example, experimenters tended to suppose that subjects would apply backward induction (independently of their reception or not of an instruction in this sense) and would focus on the game structure (independently of the context). Experimental evidence was assessed as confirmatory or disconfirmatory on the assumption that subjects behaved like game theorists. This assumption, however, was proven false and, therefore, the meaning of the evidence with regard to the tested hypotheses became an open question. Going a step further, Smith *(2010, p. 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." He understands the powerful *context effects* found in economic experiments as resulting from the subjects' previous experience, since the latter determines how subjects react to any given experimental context. According to him, the prominent role of autobiographical knowledge explains the fact that people's decisions are often as sensitive to each specific context as they are to variation in the structure of the game. As pointed out by Smith *(2010, p. 11)*:

"Under this hypothesis, each subject's personal autobiographical knowledge is filtered by contextual circumstances for relevance to the decision at hand. This conjunction determines the person's 'set point,' in which personal knowledge also appears to reflect important life style characteristics derived from the human career of sociality."

By appealing to the subjects' testimony, tools such as questionnaires and interviews would provide a means to avoid the ambiguity of experimentally observed behavior, thereby decreasing the threat to experimental validity. However, controlling the communication process so as to fully assure a common understanding of the messages by both sides —i.e., by the experimenter and the subjects under study— is extremely difficult. As an example, Smith *(2010, p. 4)* discusses the use of the word "unfair" by subjects involved in an ultimatum or dictator game. When they describe their experiences in terms of "unfairness", it is not clear whether they mean something about the rules, the outcomes, the outcomes under the rules, or other circumstances surrounding the experiment. The experimenters, by contrast, use "fair" in an outcome, non-procedural sense entrenched in the utilitarian tradition. Given that there is no direct access to what the subjects mean, their reactions to the circumstances of decision constitute the only evidence for the experimenters to interpret their speech[11].

[11] In view of all above, Jones *(2007, p. 168)* complains that ideas about understanding are often implicit and not carefully discussed in the economic literature. His article represents a first attempt at providing a more systematic account of the issue.

The connected problems of identifying motives and determining the meaning of subjects' responses also affect survey use in economics. Graham's *(2008)* emphasis on caution, when it comes to applying the findings obtained from the economics of happiness, is mainly due to the potential biases in survey data and the difficulties surrounding the analysis of this kind of data in the absence of controls for unobservable psychological traits. Subjective survey data may be responsive to the respondents' effort to be consistent with their behavior and past attitude, instead of expressing their actual attitudes. For instance, do respondents from rich backgrounds actually have a greater preference for money or are rather more prone to *report* a preference for money? *(Bertrand* and *Mullainathan, 2001, p. 69–70).* In surveys devised to measure expectations, the risk of supposing that respondents think like economic theorists becomes strikingly clear. As Curtin *(2004, p. 143–145)* warns us, in pursuing an ideal of maximum measurement precision, a shift from verbal scales to numeric probability scales has been promoted, and another one too from the second to the specification of the complete probability distribution for each person. Yet, the numeric response may still be vulnerable to misinterpretation, since two respondents may not only associate different numerical probabilities with a given verbal scale category, but also attribute different meanings to a given numeric probability. Furthermore, after the second kind of shift, an unusually large number of respondents reported a 100 percent probability for a single value —rather than probabilities spread over a range of possible outcomes—, what may be interpreted as an indication that consumers did not understand the survey question rather than as an indication of the absence of uncertainty.

Ultimately, survey use in economics entails dealing with an array of context and framing effects depending on the respondent's previous experience. Reported satisfaction or happiness, for instance, is often strongly affected by earlier questions in a survey. This has been made evident, for example, in a randomized controlled trial ran by Gallup within the Gallup-Healthways Well-being Index poll, in order to test the impact of including or not political questions before asking an evaluative well-being question to the surveyees. When the former questions are included, surveyees disapproving the political direction of their country tend then to report lower life evaluation *(Deaton* and *Stone, 2016).* Thus, global judgments by respondents are very often constructed only once that they are asked and partly as a result of their current mood, memory, and immediate contextual conditions *(Kahneman* and *Krueger, 2006, p. 6).*

3.3. The holistic and historically-sensitive nature of economic behavior

Factors (3) and (4) refer to the holistic and historically-sensitive nature of human agents and social interaction, respectively. These features render knowledge requirements for accurately predicting human behavior hard to be met. Because of the highly intertwined and dynamical nature of social interaction, laboratory experiments performed by social scientists always face a high risk of artificiality, which may seriously jeopardize external validity. In this vein, Loewenstein *(1999, p. F29–F30)* criticized the willingness of many experimental economists to do context-free experiments by removing subjects "from all familiar contextual cues."[12] Although from a different perspective, Hogarth *(2005, p. 259)* also finds economic experiments problematic in that they do not appropriately represent the changing and intertwined nature of economic behavior:

"Economists are adept at handling the requirements of classic, factorial experimental designs. (…) By varying one variable at a time and holding all others constant, one can isolate the effects. However, outside the laboratory all other variables are not constant and

[12] Opposing this view, Jiménez-Buedo and Guala *(2016, p. 12–13)* contend that an abstract or unfamiliar experimental situation may be less prone to send uncontrolled cues that undermine inferential validity, whereas a rich environment may make subjects "react to a variety of stimuli, and the way in which they construe the task may become very difficult to predict." The apparent clash between their approach and Loewenstein's might be taken as a reminder that certain features of an experimental design may reinforce internal validity at the expense of external validity.

variables are not orthogonal. (...) You can design factorial worlds within an experiment but this may not have much to do with what happens outside it."

The impossibility to implement in the lab, without deception and unchanged, those mediating conventions shaping social interaction represents an important side of the above problem, as Bardsley's *(2005)* examples mentioned in subsection 3.1 showed. Following Greenwood *(1982)*, he draws attention to the fact that social phenomena essentially depend on role-based relationships and the corresponding, shared perception that certain relational criteria are satisfied. Outside lab, in the wild, there is no doubt that a manager, for instance, considers herself as manager and is also perceived as such by her employees. Inside the lab, there is no guarantee that whoever is assigned the role of a manager in an experiments is going to be perceived as such. Experimenters, therefore, need to find out how subjects interpret the assigned roles at the lab before being able to correctly understand the experimental results. Ultimately, social positions, their related sets of collective practices, and every components of social reality are *internally related* to other positions, practices and components, which means that they all are ontologically constituted in their mutual interconnection. Experimentally isolating single social constituents does not seem like a feasible option, given the highly holistic and conventional nature of the social domain *(Lawson, 2015a, p. 44; 2015b, p. 318)*.

It has also been noted that some methodological precepts of experimental economics, such as script enactment, repeated measures, performance-based payments, and absence of deception, make subjects conform to the normative expectation that their behavior must be rational and oriented to self-interest *(Vugt, 2001)*. Notwithstanding the numerous "anomalies" found in economic experiments, the impact of the abovementioned precepts on both experimental results and external validity deserves some consideration. Already two decades ago, Siakantaris *(2000, p. 273–274)* talked of *experimental economics trade-off*, in arguing that a precepts-based, narrow definition of the experimental situation tends to severely limit parallelism and, consequently, the generalizability of the findings. Along the same lines, Ana C. Santos *(2009)* alerts that behavioral experiments tend to lose interest as human agency becomes smothered by experimental control; the tighter the control, the higher the probability that experimental results are the outcome of economists' actions.

Survey results may also be affected by the problem of external validity or similar ones. Cases of unreal anomalies, like the ones discussed by Boulier and Goldfarb *(1998, p. 19)* clearly illustrates this point. The use of surveys as a source for identifying potential or alleged anomalies may turn out unsuccessful, as it is possible that some percentage of the seeming anomalies apparently detected through surveys may constitute artefacts of measurement instead of real phenomena. Regarding happiness surveys, Holländer *(2001, p. 243)* admits that the selection of information processed in a retrospective evaluation of one's life for some period is prone to manipulation. Subjective survey data are in general vulnerable to simple manipulations, for example, though ordering and wording of questions or through designing response scales, which can affect how people process and interpret questions *(Bertrand* and *Mullainathan, 2001, p. 67–68)*. For instance, when a certain political question is placed before questions about well-being and other issues, the former has spillover effects, and these *context effects* are problematic in different ways. Such effects exhibit the following features:

- they are large, not easily removed, as they may affect answers to questions asked much later in the questionnaire;
- they differ depending on the difficulty of the questions affected, and also across population groups;
- they may coexist with high test-retest or internal reliability coefficients; and
- they may mislead those interested in monitoring national well-being over time *(Deaton* and *Stone, 2016)*.

The generalizability of survey results is also questioned by Groves and Singer *(2004, p. 24),* who claim that, when subsets of the population are subject to completely different conceptual frameworks in the area of study, a standardized measurement for a representative sample of the whole population may rely on untenable positivistic assumptions. This problem emerges most frequently when minority groups and subcultures are studied using surveys. In those cases, the ability to generate survey measures is mistakenly viewed as evidence of their informational value. Let us take, for instance, the case of employment rate, which constitutes a basic economic indicator for many societies. Obviously, measuring the employment rate may not make much sense when economic activities are generally conducted without payment for labor or when bartering systems are extensively used by some subgroup. Analogously, happiness surveys' results are biased depending on the cultural community from which the respondent is *(Powdthavee, 2007, p. 59).* Assuming a nonnegative deterministic bias mainly due to the maintenance of self-esteem and the desirability of happiness, Holländer *(2001, p. 243)* points out that the pressure to overstate one's subjective well-being (SWB) depends on the respective culture and, therefore, the bias distribution may vary across populations from different cultures even if the SWB distribution is the same. He adds that, since the average bias might depend significantly on the respective culture, comparisons of populations from different cultures might be problematic as well.

4. Validity in empirical economics and the role of philosophy

Among practicing economists, validity is often considered as a purely empirical issue and philosophy deemed of little help to deal with it. Two reductionist assumptions underlie this kind of view: one is to see philosophy of economics, or philosophy of social science in general, as a task exclusive to philosophers, rather than as a collective endeavor involving them and social scientists; the other implicit assumption, very popular among mainstream economists with a strong tendency towards scientism and other social scientists prone to radical naturalism, is to acknowledge only a subsidiary role for philosophy. We reject both assumptions and share Rosenberg's *(2015, p. 1–9)* understanding that social scientists have unavoidably to take sides on philosophical issues in order to determine what questions are answerable in their field, as well as what methods are most appropriate to answer them. Philosophers of economics should not take radical naturalism for granted, nor should they submissively assume the role of feeding the economists' illusion of closely resembling physicists, chemists and medical researchers.

Providing philosophical insights into very specific, technical issues related to validity in empirical economics certainly requires constantly keeping track of the economists' specific practices. Economists interested and knowledgeable in philosophy, as well as philosophers with economic background, may be in a position to fulfil this requirement and make valuable contributions. Yet, philosophy must also be vindicated as a global, evaluative and eminently critical reflection. An important question open to philosophical evaluation is whether current empirical economics fulfils the validity requirements established by the current normative framework shaping science, thereby contributing to debunk any overstated self-attribution of scientificity by mainstream economists. The current piecemeal, analytic strategies implemented by economists in their quest for validity should also be under philosophical scrutiny, especially to reflect on their prospects given the interconnected validity challenges arising from the intricacies of the social domain. All in all, the philosophical inquiry into the adequacy of empirical economics methodology to handle the ontological features of the social domain is not only legitimate, but also potentially useful to steer the collective effort in economic research into the most promising directions. Last but not least, philosophy is called to play a role in distinguishing those validity problems in empirical economics related to pseudoscience (hidden violation of scientific norms) from those simply resulting from an un-

scientific inquiry (neither violation of scientific norms nor satisfactory fulfilment of scientific requirements yet).

5. Conclusions

The empirical support of mainstream economics, from the mid-twentieth century to the mid-eighties, was very partial and of limited validity, as it proved unable to satisfactorily accommodate localism and control confounds. Yet, a deductive emphasis, together with a number of ideal assumptions and procedures, made mainstream economists naively think themselves to be immune to those validity challenges largely confronted in other social sciences. Economists avoided the struggle to gain more direct access to psychological variables and at the same time impoverished the institutional ones. As a consequence, they for the most part eluded the validity issues posed by experimental and survey research on the social domain. Over the last decades, this picture has changed due to several methodological enlargements which have paved the way for economics to attain a greater empirical support. Despite the progress that it represents, we have attempted to show that these additions confront economists with some inescapable challenges to the quest for validity affecting empirical social research and, ultimately, with the controversy around the scientific status of economics as a social science.

In the preceding discussion, we have examined the ontological roots of validity problems faced by today's enlarged empirical economics, focusing on awareness, the lack of homogeneous structure in shared psychological properties, and various sorts of extreme forms of holism affecting both individual actions and social interactions. At the same time, we have vindicated a role for philosophy in discussing the methodological challenges related to the above ontological peculiarities. Our hope is that this effort contributes to a more rational orientation of economic research, favoring a more realistic appraisal of the degree to which economics can aspire to be recognized as a science and, therefore, the sense in which it can be considered as such. Some avenues of inquiry to explore in our future research concern the extension of our combined (ontological and methodological) analysis to the whole machinery of empirical economics (including, for instance, econometrics and cases studies), and the identification of clear cases of pseudoscientific practices in use in empirical economics.

REFERENCES

Aaskoven, L. (2016). Fiscal Transparency, Elections and Public Employment: Evidence from the OECD. *Economics & Politics*, 28(3), 317–341.

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

Alt, J. E. and *Lassen, D. D.* (2006). Fiscal transparency, political parties, and debt in OECD countries. *European Economic Review*, 50(6), 1403–1439.

Angrist, J., *Azoulay, P.*, *Ellison, G.*, *Hill, R.* and *Lu, S. F.* (2017). Economic Research Evolves: Fields and Styles. *American Economic Review*, 107(5), 293–97.

Angrist, J. D. and *Pischke, J.-S.* (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.

Backhouse, R. E. (2010). The Puzzle of Modern Economics: Science or Ideology? Cambridge: Cambridge University Press.

Backhouse, R. and *Cherrier, B.* (2017). The age of the applied economist: the transformation of economics since the 1970s. *History of Political Economy*, 49, Supplement), 1–33.

Ball, S., *Eckel, C.*, *Grossman, P. J.* and *Zame, W.* (2001). Status in markets. *Quarterly Journal of Economics*, 116(1), 161–88.

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

Bertrand, M. and *Mullainathan, S.* (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review*, 91(2), 67–72.

Biddle, J. E. and *Hamermesh, D. S.* (2017). Theory and Measurement: Emergence, Consolidation, and Erosion of a Consensus. *History of Political Economy*, 49, Supplement, 34–57.

Blinder, A. S. (1991). Why are prices sticky? Preliminary results from an interview study. *American Economic Review*, 81(2), 89–96.

Boettke, P. J., *Leeson, P. T.* and *Smith, D. J.* (2008). The Evolution of Economics: Where We are and How We Got Here. *The Long Term View*, 7(1), 14–22.

Bond, T. N. and *Lang, K.* (2018). The Sad Truth About Happiness Scales: Empirical Results. *NBER Working Papers,* 24853. Cambridge, MA: National Bureau of Economic Research (NBER).

Bond T. N. and *Lang K.* (2019). The Sad Truth About Happiness Scales. *Journal of Political Economy*, forthcoming.

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

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

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

Bradler, C. (2015). How Creative Are You? – An Experimental Study on Self-Selection in a Competitive Incentive Scheme for Creative Performance. *ZEW Discussion Papers,* 15-021. Mannheim: Centre for European Economic Research (ZEW).

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312.

Card, D., *Chetty, R.*, *Feldstein, M.* and *Saez, E.* (2010). Expanding Access to Administrative Data for Research in the United States. NSF SBE 2020 White Paper. Arlington, VA: National Science Foundation Directorate of Social, Behavioral, and Economic Sciences.

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

Chetty, R. (2012). Time Trends in the Use of Administrative Data for Empirical Research. NBER Summer Institute presentation. Available at the author's website.

Claveau, F. (2011). Evidential variety as a source of credibility for causal inference: beyond sharp designs and structural models. *Journal of Economic Methodology*, 18(3), 233–253.

Cubitt, R. (2005). Experiments and the domain of economic theory. *Journal of Economic Methodology*, 12(2), 197–210.

Curtin, R. T. (2004). Psychology and Macroeconomics. In: *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, ed. by J. S. House, F. T. Juster, R. L. Kahn, H. Schuman and E. Singer. Ann Arbor, MI: University of Michigan Press, 121–155.

Deaton, A. and *Stone, A. A.* (2016). Understanding context effects for a measure of life evaluation: how responses matter. *Oxford Economic Papers*, 68(4), 861–870.

Dohmen, T. and *Falk, A.* (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review*, 101(2), 556–590.

Downward, P. and *Mearman, A.* (2007). Retroduction as mixed-methods triangulation in economic research: reorienting economics into social science. *Cambridge Journal of Economics*, 31(1), 77–99.

Einav, L. and *Levin, J.* (2014). Economics in the age of big data. *Science*, 346(6210), 1243089.

Falk, A., *Becker, A.*, *Dohmen, T.*, *Enke, B.*, *Huffman, D.* and *Sunde, U.* (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4), 1645–1692.

Falk, A., *Becker, A.*, *Dohmen, T.*, *Huffman, D.* and *Sunde, U.* (2016). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *IZA Discussion Papers,* 9674. Bonn: Institute for the Study of Labor (IZA).

Fudenberg, D. (2006). Advancing Beyond Advances in Behavioral Economics. *Journal of Economic Literature*, 44(3), 694–711.

Graham, C. (2008). The Economics of Happiness. In: *The New Palgrave Dictionary of Economics*, 2nd edition, ed. by S. Durlauf and L. Blume. Hampshire: Palgrave MacMillan.

Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: causal explanation and generalisation. *Journal of the Theory of Social Behaviour*, 12(3), 225–250.

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

Guala, F. (2003). Experimental Localism and External Validity. *Philosophy of Science*, 70(5), 1195–1205.

Guala, F. (2005). *The Methodology of Experimental Economics*, Cambridge: Cambridge University Press.

Guala, F. (2008). Experimentation in Economics. 2nd draft, prepared for the *Elsevier Handbook of the Philosophy of Science, 13: Philosophy of Economics*, ed. by U. Mäki. (http://users.unimi.it/guala/Handbook%20Elsevier3.pdf – Access Date: 09.01.2019).

Hamermesh, D. S. (2013). Six Decades of Top Economics Publishing: Who and How? *Journal of Economic Literature*, 51(1), 162–172.

Harrison, G. W. and *List, J. A.* (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.

Hausman, D. (1992). *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hausman, D. M. (2018). Philosophy of Economics. In: *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), ed. by Edward N. Zalta. (https://plato.stanford.edu/archives/fall2018/entries/economics/ – Access Date: 09.01.2019).

Heckman, J. J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics*, 115(1), 45–97.

Heckman, J. J. (2005). The Scientific Model of Causality. *Sociological Methodology*, 35(1), 1–97.

Heckman, J. J. (2008). Econometric Causality. *International Statistical Review*, 76(1), 1–27.

Heukelom, F. (2011). How validity travelled to economic experimenting. *Journal of Economic Methodology*, 18(1), 13–28.

Hoffman, E., *McCabe, K.* and *Smith, V. L.* (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *The American Economic Review*, 86(3), 653–660.

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

Holländer, H. (2001). On the validity of utility statements: standard theory versus Duesenberry's. *Journal of Economic Behavior and Organization,* 45(3), 227–249.

Hutchison, T. (1998). Ultra-deductivism from Nassau Senior to Lionel Robbins and Daniel Hausman. *Journal of Economic Methodology*, 5(1), 43–91.

Imbens, G. W. and *Wooldridge, J. M.* (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), 5–86.

Jensen, N. M., *Mukherjee, B.* and *Bernhard, W. T.* (2014). Introduction: Survey and Experimental Research in International Political Economy. *International Interactions*, 40(3), 287–304.

Jiménez-Buedo, M. (2011). Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*, 18(3), 271–282.

Jimenez-Buedo, M. and *Guala, F.* (2016). Artificiality, Reactivity, and Demand Effects in Experimental Economics. *Philosophy of the Social Sciences*, 46(1), 3–23.

Jiménez-Buedo, M. and *Miller, L. M.* (2010). Why a Trade-Off? The Relationship between the External and Internal Validity of Experiments. *THEORIA. An International Journal for Theory, History and Foundations of Science*, 25(3), 301–321.

Jones, M. K. (2007). A Gricean analysis of understanding in economic experiments. *Journal of Economic Methodology*, 14(2), 167–185.

Juster, F. T. (2004). The Behavioral Study of Economics. In: *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, ed. by J. S. House, F. T. Juster, R. L. Kahn, H. Schuman and E. Singer. Ann Arbor, MI: University of Michigan Press, 119–130.

Kahneman, D. and *Krueger, A. B.* (2006). Developments in the Measurement of Subjective Well-Being. *The Journal of Economic Perspectives*, 20(1), 3–24.

Kalaitzidakis, P., *Mamueas, T. P.* and *Stengos, T.* (2003). Rankings of academic journals and institutions in economics. *Journal of the European Economic Association*, 1(6), 1346–1366.

Kling, C. L., *Phaneuf, D. J.* and *Zhao, J.* (2012). From Exxon to BP: Has Some Number Become Better Than No Number? *Journal of Economic Perspectives*, 26(4), 3–26.

Krueger, A. B. and *Stone, A. A.* (2014). Progress in measuring subjective well-being: Moving toward national indicators and policy evaluations. *Science*, 346(6205), 42–43.

Lawson, T. (2015a). A Conception of Social Ontology. In: *Social Ontology and Modern Economics*, ed. by S. Pratten. London: Routledge, 19–52.

Lawson, T. (2015b). Methods of abstraction and isolation in modern economics. In: *Social Ontology and Modern Economics*, ed. by S. Pratten. London: Routledge, 315–337.

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

Lissitz, R. W. (ed.) (2009). *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC: Information Age Publishing.

List, J. A. (2009). An introduction to field experiments in economics. *Journal of Economic Behavior & Organization*, 70(3), 439–442.

Loewenstein, G. (1999). Experimental Economics from the Vantage-Point of Behavioural Economics. *The Economic Journal*, 109(453), 25–34.

Lucking-Reiley, D. (1999). Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review*, 89(5), 1063–1080.

Mäki, U. (2018). Empirical turn in economics? *The 30th Annual EAEPE (European Association for Evolutionary Political Economy) Conference 2018. Evolutionary foundations at a crossroad: Assessments, outcomes and implications for policy makers*, Nice, France, 06–08 September.

Markus, K. A. and *Borsboom, D.* (2013). Frontiers of Test Validity Theory: Measurement, Causation, and Meaning. New York: Routledge.

Mason, C. F., *Phillips, O. R.* and *Nowell, C.* (1992). Duopoly behavior in asymmetric markets: an experimental evaluation. *The Review of Economics and Statistics*, 74(4), 662–670.

McCloskey, D. N. and *Ziliak, S. T.* (2001). *Measurement and meaning in economics: the essential Deirdre McCloskey*. Cheltenham, UK: Edward Elgar.

Meyer, B. D., Mok, W. K. C. and *Sullivan, J. X.* (2015). Household Surveys in Crisis. *Journal of Economic Perspectives*, 29(4), 199–226.

Morgan, J. and *Patomäki, H.* (2017). Contrast explanation in economics: its context, meaning, and potential. *Cambridge Journal of Economics*, 41(5), 1391–1418.

Morgan, J. N. (1967). Contributions of survey research to economics. In: *Survey Research in the Social Sciences*, ed. by C. Glock. New York: Russell Sage Foundation, 217–268.

Naef, M. and *Schupp, J.* (2009). Measuring Trust: Experiments and Surveys in Contrast and Combination. *IZA Discussion Papers,* 4087. Bonn: Institute for the Study of Labor (IZA).

Newman, J. (2017). Our Obsession with Survey Data is Ruining Economics. Article posted at July 23, 2017 on the *Mises Wire*. (https://mises.org/blog/our-obsession-survey-data-ruining-economics – Access Date: 09.01.2019)

Panhans, M. T. and *Singleton, J. D.* (2017). The empirical economist's Toolkit: from models to methods. *History of Political Economy*, 49, Supplement, 127–157.

Perraton, J. (2011). Explaining growth? The case of the trade-growth relationship. *Journal of Economic Methodology*, 18(3), 283–296.

Porter, M. E., *Ketels, C.* and *Delgado, M.* (2007). The Microeconomic Foundations of Prosperity: Findings from the Business Competitiveness Index. In: *The Global Competitiveness Report 2007–2008*, ed. by M. E. Porter, K. Schwab, and X. Sala-i-Martin. London: Palgrave Macmillan, 51–81.

Powdthavee, N. (2007). Economics of Happiness: A Review of Literature and Applications. *Chulalongkorn Journal of Economics*, 19(1), 51–73.

Powdthavee, N. (2015), Would You Like to Know What Makes People Happy? An Overview of the Datasets on Subjective Well-Being. *The Australian Economic Review*, 48(3), 314–320.

Presser, S. (1984). The Use of Survey Data in Basic Research in the Social Sciences. In: *Surveying Subjective Phenomena*, ed. by C. F. Turner and E. Martin. New York: Russell Sage Foundation, 93–114.

Presser, S. and *McCullogh, S.* (2011). The Growth of Survey Research in the United States: Government-Sponsored Surveys, 1984–2004. *Social Science Research*, 40(4), 1019–1024.

Price, L. (2017). *Psychometric Methods: Theory into Practice*. New York: The Guilford Press.

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

Rodrick, D. (ed.) (2003). *In Search for Prosperity. Analytical Narratives on Economic Growth*. Princeton: Princeton University Press.

Rosenberg, A. (1992). *Economics: Mathematical Politics or Science of Diminishing Returns?* Chicago: University of Chicago Press.

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

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

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

Ross, D. (2011). Estranged parents and a schizophrenic child: choice in economics, psychology and neuroeconomics. *Journal of Economic Methodology*, 18(3), 217–231.

Ross, D. and *Kincaid, H.* (2009). Introduction: The New Philosophy of Economics. In: *The Oxford handbook of philosophy of economics*, ed. by H. Kincaid and D. Ross, Oxford: Oxford University Press, 3–32.

Ruzzene, A. (2014). *Using case studies in the social sciences. Methods, inferences, purposes*. Thesis to obtain the degree of Doctor from the Erasmus University Rotterdam (EUR) by

command of the rector magnificus Prof. Dr. H. A. P. Pols and in accordance with the decision of the Doctorate Board. Rotterdam: EUR.

Santos, A. C. (2009). Behavioral experiments: how and what can we learn about human behavior. *Journal of Economic Methodology*, 16(1), 71–88.

Saris, W. E. and *Gallhofer, I. N.* (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, New Jersey: John Wiley & Sons.

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

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2), 225–237.

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

Shulman, R. G. and *Shapiro, I.* (2009). Reductionism in the Human Sciences: A Philosopher's Game. In: *Philosophy of the social sciences: philosophical theory and scientific practice*, ed. by C. Mantzavinos. Cambridge: Cambridge University Press, 124–129.

Siakantaris, N. (2000). Experimental Economics Under the Microscope. *Cambridge Journal of Economics*, 24(3), 267–281.

Siedler, T. and *Sonnenberg, B.* (2010). Experiments, Surveys and the Use of Representative Samples as Reference Data. *RatSWD Working Papers,* 146. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD).

Simon, H. A. (1997). *An Empirically Based Microeconomics*. Cambridge: Cambridge University Press.

Smith, V. L. (2010). Theory and experiment: What are the questions? *Journal of Economic Behavior & Organization*, 73(1), 3–15.

Suppes, P. (1982). Problems of Causal Analysis in the Social Sciences. *Epistemologia. Rivista Italiana di Filosofia della Scienza*, 5, special issue, 239–250.

Syll, L. P. (2016). Deductivism: the fundamental flaw of mainstream economics. *Real World Economics Review,* (74), 20–41.

Venaik, S. (2007). Abstract of the seminar: *Factors affecting the choice of measurement models in SEM*. Teached at the NUS on April 17. (http://bschool.nus.edu.sg/Departments/BussPolicy/Seminar%20Abstract_%20Profiles/SunilVenaik.abstract.htm – Access Date: 20.07.2011)

Vischer, T., *Dohmen, T.*, *Falk ,A.*, *Huffman, D.*, *Schupp, J.*, *Sunde, U.* and *Wagner, G. G.* (2013). Validating an Ultra-Short Survey Measure of Patience. *Economics Letters*, 120(2), 142–145.

Voigt, S. (2013). How (Not) to measure institutions. *Journal of Institutional Economics*, 9(1), 1–26.

Vugt, M. V. (2001). Self-interest as self-fulfilling prophecy. *Behavioral and Brain Sciences*, 24(3), 429–430.

Wainer, H. and *Braun, H. I.* (eds.) (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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