Synthese (2021) 199:9957–9978
https://doi.org/10.1007/s11229-021-03232-2

ORIGINAL RESEARCH

Resolving empirical controversies with mechanistic evidence

Mariusz Maziarz[1,2]

Received: 14 November 2020 / Accepted: 22 May 2021 / Published online: 7 June 2021

Abstract

The results of econometric modeling are fragile in the sense that minor changes in estimation techniques or sample can lead to statistical models that support inconsistent causal hypotheses. The fragility of econometric results undermines making conclusive inferences from the empirical literature. I argue that the program of evidential pluralism, which originated in the context of medicine and encapsulates to the normative reading of the Russo-Williamson Thesis that causal claims need the support of both difference-making and mechanistic evidence, offers a ground for resolving empirical disagreements. I analyze a recent econometric controversy regarding the tax elasticity of cigarette consumption and smoking intensity. Both studies apply plausible estimation techniques but report inconsistent results. I show that mechanistic evidence allows for discriminating econometric models representing genuine causal relations from accidental dependencies in data. Furthermore, I discuss the differences between biological and social mechanisms and mechanistic evidence across the disciplines. I show that economists mainly rely on mathematical models to represent possible mechanisms (i.e., mechanisms that could produce a phenomenon of interest). Still, claiming the actuality of the represented mechanisms requires establishing that crucial assumptions of these models are descriptively adequate. I exemplify my approach to assessing the quality of mechanistic evidence in economics with an analysis of two models of rational addiction.

Keywords Evidential pluralism · Russo-Williamson thesis · Inconsistent results · Statistical malleability · Mechanistic evidence

This article is part of the topical collection "Evidential Diversity in the Social Sciences", edited by Yafeng Shan and Jon Williamson.

✉ Mariusz Maziarz
mariusz.maziarz@uj.edu.pl

[1] Interdisciplinary Centre for Ethics, Jagiellonian University, Grodzka 52, Kraków, Poland

[2] Institute of Philosophy, Jagiellonian University, Grodzka 52, Kraków, Poland

1 Introduction

According to Russo and Williamson (2007), establishing causality requires both difference-making and mechanistic evidence. The normative reading of the Russo-Williamson Thesis inspired the program of evidence-based medicine plus (EBM+) (Parkkinen et al., 2018). The supporters of evidential pluralism advise broadening the evidentiary base accepted by the proponents of evidence-based medicine, which focuses primarily on difference-making (correlational) evidence and assesses the quality of evidence based on the risk of confounding (Borgerson, 2009; La Caze, 2009). While the debate on whether the approach of EBM+ is superior to EBM in regard to assessing evidence for causal claims in medicine is still ongoing (e.g., Canali, 2019; Howick, 2011; Williamson, 2019), evidential pluralism has been considered as relevant for social sciences (e.g., Johnson et al., 2019; Kuorikoski & Marchionni, 2016; Rhemtulla et al., 2017; Runhardt, 2020).

Economics emerges as the social science that is most often considered the field that could benefit from evidential pluralism. The arguments of philosophers supporting evidential pluralism as a plausible approach to causal inference in economics are either descriptive or normative. In the former case, current methods of causal inference are interpreted as agreeing with the program of evidential pluralism. For example, Russo (2009a, 2009b) interpreted some common econometric approaches to modeling (such as structural equation models) as being in agreement with the variational epistemology. According to Russo's reconstruction, later developed with (Moneta & Russo, 2014), econometric models deliver difference-making evidence for correlations estimated with statistical techniques. Still, mechanistic evidence comes in as background knowledge, statistical assumptions, and economic theory that, put together, allow for the causal interpretation of statistical models. Behavioral economics has also been interpreted in accordance with evidential pluralism. Çağlar Dede (2019) analyzed the evidentiary base for Incentivized Smoking Cessation Policies (these include, but are not limited to, supporting quitting smokers with small financial rewards) and argued that considering the results of different types of research being evidence of different types improves the quality of behavioral public policies.

Such arguments are opposed with counterexamples showing that economists put forward causal conclusions based on a single type of evidence. In our reply to Henschen's (2018) support for a version of the manipulationist definition, Mróz and I (Maziarz & Mróz, 2020) showed that at least in some areas of macroeconometrics, causal inferences are put forth exclusively on the ground of difference-making evidence while other economists support their causal conclusions with mechanistic evidence only. In *Philosophy of Causality in Economics: Causal Inferences and Policy Proposals* (Maziarz, 2020), I delivered additional examples of studies drawing causal conclusions from either difference-making or mechanistic evidence and interpreted the practice in line with causal pluralism (i.e., pluralism about concepts of causality) on the ground that different types of evidence give support for different kinds of causal claims (see Reiss, 2009). In a similar vein, Claveau (2012) argued that one out of three causal claims widely accepted

by economists is supported only with mechanistic evidence with no evidence of variational nature. Shan and Williamson (2020) rebutted that there is some correlational evidence for the claim regarding the influence of unemployment benefits eligibility criteria on unemployment rate. Moreover, the claim itself is not accepted consensually.

Even if some examples are questionable, the literature shows that evidential pluralism is at odds with the full range of this field's research practices. In some cases, economists do not support their causal claims with both types of evidence, finding difference-making evidence from econometrics or mechanistic modeling sufficient. These research practices possibly stem from the view that different types of economic policymaking require different types of evidence. For instance, prediction, which is one of the goals of econometrics, does not require mechanistic evidence (see Reiss, 2007). While these examples undermine taking evidential pluralism as a view descriptively adequate to how causal inferences are drawn in the discipline, they do not weaken evidential pluralism considered a normative position. Recently, Shan and Williamson (2020) made a general point that social sciences can benefit from applying the program of evidential pluralism to causal inference.

While I study an empirical controversy in econometrics (and, hence, the article is partially descriptive), my primary purpose is normative. I point out one of the reasons for why evidential pluralism should be imported from medicine to social sciences. I argue that evidential pluralism allows for resolving some controversies when inconsistent causal hypotheses emerge from plausible econometric models. Hence, the case study is employed to draw a lesson from the exchange of arguments among economists and learn how mechanistic evidence can help choose an accurate statistical model in case of conflicting difference-making evidence. My argument stems from the suggestion of Moneta and Russo (2014, p. 70) that "a major gain in adopting the framework [(…) is that it] allows a better and more fruitful analysis of conflict resolutions, i.e. when scientists disagree about causal relations." Moneta and Russo (2014) have exemplified their point with the historical debate regarding monetarism. This may be problematic because such sound controversies heavily depend on the multitude of existing studies and political commitments of involved parties. The reliance on background knowledge and bias may undermine informing the philosophical debate. For this reason, I focus on the recent econometric controversy regarding the influence of concise tax hikes on smokers' behavior (Abrevaya & Puzzello, 2012; Adda & Cornaglia, 2006, 2013). I argue that using mechanistic evidence allows for discriminating econometric models representing genuine causal relations from accidental dependencies in data or lack thereof. Still, assessing the quality of mechanistic evidence is crucial and I suggest how the question of whether mechanistic models represent actual (and not possible) mechanisms can be addressed.

The structure of the paper is as follows. In Sect. 2, I point out that econometric modeling is malleable in the sense that plausible methodological decisions allow for obtaining inconsistent results despite using the same or similar datasets. The fragility of econometric results undermines making inferences from empirical literature. In Sect. 3, I analyze the reasons for why Adda and Cornaglia (2006) and Abrevaya and Puzzello (2012) obtain inconsistent estimates of the effect of concise tax hikes

on cigarette consumption and the levels of cotinine (a biomarker for nicotine) in saliva samples. The analysis shows that both econometric models are plausible and, hence, additional evidence is needed to resolve the controversy. In Sect. 4, I distinguish between economic and biological mechanisms and mechanistic evidence across the disciplines. In Sect. 5, I assess the quality of mechanistic evidence for the inconsistent causal hypotheses put forward by Adda and Cornaglia (2006) and Abrevaya and Puzzello (2012) and conclude that this evidence allows for choosing the econometric result that is more likely to represent the genuine causal dependency. All in all, the case study shows that the approach of evidential pluralism is useful when methodologically-sound statistical models are in disagreement.

2 The malleability of econometric modeling

Scientific disagreement, i.e., "situations in which scientists hold different opinions or theories, or in which they hold different views about scientific results, or even about the processes to obtain such results" (Parkkinen et al., 2017, p. 76), has fallen within the research interests of philosophers (Matheson & Frances 2018; Martini et al., 2013) for some time. Still, the situations when econometricians obtain divergent or even inconsistent causal hypotheses from similar (or even the same) datasets are poorly understood and raise heated methodological debates. A prime example of such a disagreement is the Reinhart-Rogoff controversy. In response to growing public indebtedness after the 2007–2008 financial crisis, Reinhart and Rogoff (2010a) studied a dataset including historic levels of debt and GDP growth for a broad group of countries in a data-driven way. Even though the team refrained from putting forward an explicitly causal conclusion based on the statistical dependency between debt and growth, their policy advice to lower public debt points out that they have interpreted their result in terms of causality. Herndon et al. (2014) chose alternative statistical procedures and obtained results undermining Reinhart and Rogoff's (2010a, 2010b) threshold hypothesis (see Maziarz, 2017). This controversy is just one example of a long list of topics where econometric results are in disagreement. Doucouliagos and Stanley (2013) listed 65 such topics in empirical econometrics. According to Goldfarb's (1997) estimate, approximately 10% of articles published in the American Economic Review, the top economic journal, instantiate the 'emerging inconsistent result'/'emerging recalcitrant result' (ERR) phenomenon

The reasons for why inconsistent results populate econometrics literature are poorly understood. The existing explanations for the phenomenon involve inspiration by conflicting theories (Doucouliagos & Stanley, 2013), p-hacking (De Long & Lang 1992), more or different data, and different or fancier techniques (Goldfarb, 1997). Whether it is a conflict of interest, the institutional setting of academic econometrics, or econometricians' personal views on methodology and economic theory that motivates econometricians to choose alternative statistical techniques, the problem of inconsistent results lies in the malleability of econometric modeling. The dependency of econometric results on different but plausible methodological decisions has been well documented by Morgan and Magnus (1997), who conducted an experiment in applied econometrics. They asked

participants to estimate a few models and measurements on the ground of the same dataset. The differentiation of results astonished them: "[e]ven with the simplest first task of measurement, we had underestimated the ability of eight participating teams to produce different versions of the variables, different models, and different measurement procedures" (p. 464).

Moosa (2019) argued recently that econometric results are fragile in the sense that two studies can report very different results even if they differ by just a few methodological decisions. He highlighted that, often, the inconsistent results are of similar quality. In a similar vein, Clive Granger (1999), a laureate of the Nobel Memorial Prize in Economic Sciences for his developments of econometric methodology, explicitly admitted that the results of statistical modeling depend on methodological decisions to which econometricians commit: "[b]ecause of […] the many decisions that make up the actual economy to the empirical model, and the many possible ways to form an approximation, it is obvious why various models can occur" (p. 5). What is crucial and undermines drawing policy-oriented conclusions from difference-making evidence alone is that these alternative methodological decisions are plausible. Empirical controversies in econometrics arise because, even though *groups of* econometricians may prefer one technique over another, one cannot argue that one or another technique is *objectively* superior because these decisions are subjective and arbitrary, as long as they can be justified from the perspective of econometric methodology. Reinhart and Rogoff (2010b) admitted exactly this before their *American Economic Review* (Reinhart & Rogoff, 2010a) paper became controversial: "[t]hose who have done data work know that mapping vague concepts like 'high debt' or 'overvalued exchange rates' into workable definitions requires arbitrary judgments about where to draw lines; there is no other way to interpret the facts [i.e., data] and inform the discussion."

Another problem related to mapping theoretical concepts into observable variables is the discrepancy between theoretical and statistical models. Spanos (1995) argued that the data generating processes (mechanisms in his terminology) could drastically differ from what is postulated by economic theory. For example, a vast majority of economics relies on the ceteris paribus clause. Econometrics is still concerned with modeling observational data (instead of data generated by an experimental design isolating away external influences). For this reason, the statistical assumptions of econometric models are often not fulfilled what makes such evidence inconclusive. The problem is also strengthened by the 'textbook approach' to statistical modeling, where only one model specification inspired by economic theory is estimated, and the obtained results are interpreted as its test. Spanos disagreed with the standard view that data snooping introduces bias and opted for using the Probabilistic Reduction approach, which is exploratory in nature, to modeling observational data. However, considering that many economists estimate statistical models on the same observational data sets, the two approaches may, in fact, lead to the same effects if economic theory is sufficiently pluralistic to allow for hypothesizing all model specifications that would be attempted in an exploratory analysis (see Denton, 1988). In that case, if economists are either required by journal editors or self-motivated to publish novel results, reversals in findings are likely to emerge (Goldfarb, 1995, 1997).

In sum, the malleability of statistical techniques makes empirical results in economics fragile: using an alternative but plausible estimation strategy or slightly modified sample, econometricians can obtain models that support inconsistent causal hypotheses. In such situations, evidential pluralism comes in handy and, as I show in Sect. 4, constrains the menu of empirical results to only those for which plausible mechanisms exist. Below, I discuss a recent empirical controversy regarding tax elasticities of cigarette consumption and smoking intensity, i.e., the influence of changes in concise tax on cigarette consumption and smoking intensity, and analyze methodological reasons for why Adda and Cornaglia (2006, 2013) and Abrevaya and Puzzello (2012) obtained econometric models suggesting inconsistent causal hypotheses regarding the influence cigarette taxes on nicotine intake.

These studies exemplify the reasons for why inconsistent causal hypotheses can be supported with statistical models of observational data and how mechanistic evidence can help resolve the controversy. The two case studies are representative of the reasons why data researchers obtain inconsistent results. The disagreement between Adda and Cornaglia (2006, 2013) and Abrevaya and Puzzello (2012) results from different but plausible methodological decisions. Hence, one is unable to resolve the controversy by arguing that one of the two inconsistent results is an effect of using inappropriate (from the perspective of statistical methodology) statistical techniques what implies that an additional source of evidence is required for the resolution. Furthermore, the discussion of economic models representing addicts' decision process shows that mechanistic evidence can also be divided, i.e., economists can disagree in regard to which of a few hypothesized mechanisms actually produces the phenomenon. It allows for drawing a lesson that the quality of mechanistic evidence needs to be carefully assessed. Considering that the sources of mechanistic evidence in economics and medicine are different, the case study is a good ground for discussing approaches to evaluating the quality of mechanistic evidence stemming from mathematical models. Still another reason for analyzing the case study of the controversy regarding the influence of excise tax on smoking is that taxation changes not only cigarette purchases but also smoking intensity what exemplifies the view of Kelly et al., (2014) that social and biological mechanisms of diseases are often intertwined. The question regarding the influence of concise tax on smoking intensity falls within the research interests of not only econometricians but also epidemiologists (e.g., Caraballo et al., 1998; Falba et al., 2005; Shiffman & Scholl, 2018), and therefore, I hope, the lesson from the case study can be useful not only for economists but also convince medical researchers to endorse the guidance of evidential pluralism and the movement of EBM+(see Parkkinen et al., 2018). The problem of inconsistent causal hypotheses supported with statistical models is not specific to economics but troubles all disciplines using observational data, with epidemiology being the prime example (Broadbent, 2013; Tatsioni et al., 2007).

3 Estimating the influence of excise tax on smoking intensity

Until the 2000s, econometric evidence consistently showed that raises in cigarette taxes and, subsequently, their prices lead to lower consumption. This correlation has been supported by the theoretical, axiomatic model of rational addiction (see Becker

et al., 1991). According to the model, smokers are rational agents that optimize their utility by choosing the number of cigarettes consumed based on their income, prices of all other goods, and cigarettes' price. Both econometric studies and theoretical models included a simplifying assumption that cigarettes are a homogenous good.

The assumption that smoking and cigarettes are not differentiated is at odds with epidemiological literature. Smokers are modeled as choosing cigarettes including more nicotine and tar, or smoking more intensively in response to tax hikes (see Caraballo et al., 1998). Adda and Cornaglia (2006) decided to exchange the exogenous variable denoting the number of cigarettes consumed with the level of cotinine, one of nicotine's metabolites in blood. Specifically, Adda and Cornaglia (2006, p. 1016) estimated the following two panel-data models[1] that represent the influence of exogenous (explanatory) variables of cotinine levels and cigarette consumption. The first regression estimates the impact of the level of taxation, age, sex, race, education attrition, household size, and the number of smokers in the family on the level of cotinine measured in the salvia sample from ith smoker from state s at time t. The second regression estimates the influence of the same factors on cigarette consumption:

$$\log cot_{ist} = \beta_0 + \beta_1 \log tax_{st} + \beta_2 X_{ist} + \beta_s + \beta_t + v_{ist}$$

$$\log c_{ist} = \alpha_0 + \alpha_1 \log tax_{st} + \alpha_2 X_{ist} + \alpha_s + \alpha_t + u_{ist}$$

where $\log cot_{ist}$: logarithm of cotinine level in ith smoker from state s at time $t\log c_{ist}$: logarithm of cigarettes consumed by ith smoker at time; β_n; α_n: estimated parameters; $\log tax_{st}$: logarithm of taxation in state s at time t; X_{ist}: variables describing smokers' characteristics (age, sex, race, education level, household size, other smoking members in the family); α_s; β_s: state dummies; β_t; α_t: time dummies; v_{ist}; u_{ist}: error terms.

The comparison of the parameters denoting partial correlation between the logarithm of taxation level and cotinine level β_1 and taxation and cigarettes consumption α_1 allows for informing if tax hikes result in more intense smoking and/or lower consumption of cigarettes. These regressions have been estimated on data from National Health and Nutrition Examination Survey (in particular, NHANES III and NHANES 1999–2000 that cover, respectively, the periods between 1988 to 1994 and 1999 to 2000 and deliver data for approximately 20,000 individuals) merged by the econometricians with a database reporting the levels of excise tax on cigarettes for each state (adjusted for inflation). Even though more recent data had been available, Adda and Cornaglia (2006, p. 1016) decided not to use them because "[t]he later waves contain less information on individual characteristics, which restricts our analysis in some cases." Otherwise, they would have to lower the number of explanatory variables (X_{ist}) what could potentially lead to model

[1] Adda & Cornaglia (2006) also estimated a third regression, with smoking intensity being the explained variable, but, considering that the controversy has focused on the former two, I also limit the discussion to these two statistical models.

misspecification and obtaining a model susceptible to the common cause fallacy, where individual characteristics instead of cigarette taxes cause smoking intensity. For instance, income and belonging to minorities are known confounders (Gilkes et al., 2017; Jung et al., 2013).

Unfortunately, NHANES does not collect data on income, and Adda and Cornaglia (2006, p. 1020) needed to use a proxy. They decided to employ educational attainment and house size, which are highly correlated with income. Adda and Cornaglia (2006, p. 2019) used an estimation strategy developed originally by Manski (1994). The estimated parameter values show that tax hikes negatively influence cigarettes consumption ($\alpha_1 = -0.2$) but have no statistically significant effect on cotinine ($\beta_1 \approx 0$). This suggests that smokers adjust their smoking intensity to compensate for raised prices by extracting more nicotine. On this ground, Adda and Cornaglia (2006, p. 1014) admitted that "[b]y ignoring smoking intensity, these [previous] models are misspecified." Their result suggests that the efficacy of cigarette tax increases as a policy tool aimed at reducing nicotine consumption is limited.

Abrevaya and Puzzello (2012) conducted a conceptual replication of Adda and Cornaglia (2006). The replication study uses a broader sample and an alternative estimation strategy, and reports inconsistent results with the original study. Abrevaya & Puzello (2012, p. 1756) reported that "[f]or the unrestricted sample of smokers, the estimated cigarette and cotinine elasticities remain statistically insignificant (they are both positive) despite the much larger sample size. For the restricted sample of long-time smokers, the estimated elasticities are both negative (-0.3271 for cigarettes, −0.2091 for cotinine) with the cigarette elasticity significant at a ten percent level". This might be surprising, considering that the unrestricted sample used in the replication study is significantly larger and includes 3658 observations from 26 states compared to 2685 observations from 13 states in the original study (for the period of 1988–1994 in both cases). Abrevaya & Puzello (2012) pointed out that demand elasticities are estimated exclusively based on an even lower number of countries because some states have been 'observed' only once in the four years. Therefore, the original sample includes only seven and the extended sample 13 states with changes in the value of excise tax. Furthermore, these changes were moderate, ranging from 2 to 23 percentage points. Abrevaya and Puzello (2012) pointed out that the limited variability of taxation across the states suggests that the NHANES dataset used by Adda and Cornaglia (2006) is not suitable for estimating tax elasticities of cigarette consumption and cotinine levels.

The differences between the two statistical models can be considered simply as the estimation of inconsistent values of tax elasticity of smoking (cigarette consumption): Adda and Cornaglia (2006) rejected the null hypothesis (i.e., $\alpha_1 < 0$) while Abrevaya and Puzzello (2012) reported the coefficient to insignificantly differ from null (i.e., $\alpha_1 = 0$). Given that both studies agree that the tax elasticity of smoking intensity (cotinine levels) insignificantly differs from zero ($\beta_1 = 0$), they hypothesize different behavior of smokers and, in effect, put forward inconsistent assessment of the efficacy of concise tax hikes. Adda and Cornaglia (2006) claim that smokers change smoking intensity to extract more nicotine from cigarettes in response to tax hikes. In contrast, Abrevaya and Puzzello (2012) argue that the NHANES dataset is not suitable for estimating tax elasticities and find the evidence inconclusive.

The inconsistency results solely from two differences between these studies: the use of wider sample and an alternative estimation technique. Even though the two teams of econometricians consider their methodological choices as superior and hence their outcome as warranted, one can easily find arguments in support of each technique. Obviously, supporting the view that both alternative methodological decisions regarding sample and estimation strategy are equally good faces the problem of choosing a perspective for such claims since the 'view from nowhere' is hardly attainable. For this reason, my claim is weaker. I argue the two studies are methodologically sound, i.e., that questionable research practices (such as p-hacking) are absent from these studies. None of the two teams broke commonly accepted methodological rules. Furthermore, I claim that the two methodological differences accounted for the difference in outcomes.

Changes in samples influence smokers' proportions that 'bias'[2] the overall elasticity estimate in certain ways. The extended sample is systematically different from the original sample and includes instances with a high number of cigarettes smoked per person daily and a lower level of concise taxes. The extended dataset includes more instances that 'bias' the elasticity estimates towards the null. In contrast, the original dataset includes primarily instances that allow for obtaining elasticity estimate significantly lower than zero. As Adda and Cornaglia (2013) put it, "[a] difference between the two datasets is the inclusion of tobacco states. These states are characterized by higher cigarette consumption, lower taxes, and little variation in excise taxes over the period considered" (p. 3103). Nesson (2017), who used a sample of size two times larger than Adda and Cornaglia (2013) reported the heterogeneity of smokers' responses to tax hikes and, in particular, that tax elasticity of smoking intensity is highest in light and moderate smokers. Considering that the extended sample includes a greater proportion of heavy smokers, the tax elasticities estimated on the extended sample can be expected to be lower in size (i.e., closer to null) than estimates based on the original sample.

The exclusion of some observations in the original study resembles the Reinhart-Rogoff controversy (see Maziarz, 2017). Still, this case is different regarding the motives for the exclusion, which can be rationally justified. The exclusion of the tobacco states can be supported by the argument that these states experience lower variation in taxation. The lower the variation, the less precise elasticity estimates are, and therefore the exclusion improves estimate precision. In contrast, the supporters of the replication study could rebut that such estimates are only representative for a chosen handful of states.

Furthermore, there is also a more practical reason for the exclusion of data by the original study. Namely, National Center for Health Statistics (NCHS), which conducts the NHANES study, codes state variables as missing data for some instances to assert confidentiality. Adda and Cornaglia (2006) have used a dataset delivered by NCHS and excluded instances with missing data, while Abrevaya & Puzello (2012) obtained the full dataset. Even though Adda & Cornaglia could have used a data

[2] The notion of 'bias', which means changing the estimate in some direction seems to be most suitable despite the connotations with the use of inappropriate or questionable techniques.

imputation technique to deal with missing observations for the state variable (α_s; β_s), this would also be problematic because of the multitude of such methods that lead to different results (Gold & Bentler, 2000). However, the externality and independence of the reason for data exclusion can save Adda and Cornaglia (2006) from accusations of p-hacking.

The other methodological decision that influenced elasticity estimates and accounts for the inconsistency is the choice of estimation strategy. Adda and Cornaglia (2006, p. 1016) used two strategies to deal with selection bias. First, they restricted the sample to heavy smokers, which are less likely to quit smoking. Second, they used the approach developed by Manski (1994), which relies on bounding the OLS coefficients using the worst-case bounds. The latter method uses estimates of cessation probability to modify the parameter estimates. It accounts for the possibility that the less heavy smokers are more likely to quit in response to tax hikes. In the opinion of the econometricians conducting the replication study, the use of clustered standard errors is the "appropriate" (Abrevaya & Puzzello, 2012, p. 1751) estimation strategy due to the small number of states. This suggests that Abrevaya and Puzzello (2012) consider the sample restrictions and bounding OLS coefficients to be inappropriate. The use of weighting in regression analysis leads to different results: elasticity estimates lack significance and, in the case of tax elasticity of cigarette consumption, are of sign different from the original study. Abrevaya & Puzello (2012, p. 1758) cited the suggestion of DuMouchel and Duncan (1983) that such changes are a sign of misspecification. However, since the use of weighting changes the influence of particular observations on the overall outcome, it is also plausible with the effects of excise tax hikes on cigarette consumption and cotinine levels being heterogeneous across states.

In response to Abrevaya and Puzzello (2012) criticism, Adda and Cornaglia (2013) delivered another estimation of the tax elasticities of cigarette consumption and cotinine levels. They used an alternative dataset from Coronary Artery Development in Young Adults (CARDIA) study that extends the NHANES sample and covers years between 1988 and 2006. Adda and Cornaglia (2010) first used the extended sample for estimating the influence of excise taxes on passive smoking. State-year observations increased from 60 to 147, which improves the accuracy of elasticity estimates. Moreover, the analysis of the extended sample allowed Adda and Cornaglia (2013) to gather methodological insight and to explain why their original results are at odds with the regressions estimated by Abrevaya and Puzzello (2012). According to Adda and Cornaglia (2013, p. 3102), dynamic selection has the potential to bias OLS regressions of the influence of smoking intensity on changes in cigarette prices and excise tax: the OLS estimates alone are biased towards the null.

Additionally, they rebutted the accusation of Abrevaya and Puzzello (2012), who referred to the work of DuMouchel and Duncan (1983) and pointed out that introducing weights to a regression can bias the results. Furthermore, Adda and Cornaglia (2013) pointed out that their weighting, resulting from demographic characteristics, improves the estimate's accuracy (oversampling of certain groups and accounting for nonparticipation). The reason is that the sample weights are correlated with the endogenous variable what undermines the assumption of DuMouchel and Duncan (1983). In such situations, the use of weights is justified (Maddala,

1983). Additionally, Adda and Cornaglia (2013) used one-period lagged tax because taxes are reported at the end of the fiscal year (end of June) while smoking at the end of the calendar year. This contrasts with the previous practice of using contemporaneous taxes and possibly improves estimate accuracy. The analysis of the extended sample confirmed their original results that "smoking intensity responds to price changes over this period" (Adda & Cornaglia, 2013, p. 3102).

However, the arguments presented by Adda and Cornaglia (2013) in their response do not allow for taking the results of Abrevaya and Puzzello (2012) as fallacious or spurious. In particular, clustered standard errors are known to lead to estimating confidence intervals and *p*-values that are too small in comparison to the actual accuracy of point estimates (MacKinnon & Webb 2017). Given that the NHANES dataset includes observations from states with different levels of excise tax, standard errors are likely to be clustered. Generally speaking, there are two approaches to inference from clustered standard errors data: appropriate model specification and using cluster-robust estimation technique (Rand, 1971). Adda and Cornaglia (2006) and Abrevaya and Puzzello (2012) used these alternative but plausible solutions and none of the results can be chosen as superior from the perspective of statistical/econometric methodology. Goldfarb (1997) listed "more/different data" and "different or 'fancier' techniques" as the first two reasons for the emergence of inconsistent results. These reasons, i.e., using the wider sample and different estimation technique by the replication study made the econometricians obtain inconsistent tax elasticities of cigarette consumption. Considering that discriminating the accurate from false result is impossible on the ground of econometric methodology, one cannot tell the accurate from the spurious work based on data alone.

4 Mechanistic evidence in economics

In such situations, evidential pluralism comes in handy and offers an additional ground, external to difference-making evidence, to resolve controversies and make useful inferences. Fortunately, Russo and Williamson (2007) have left the notion of mechanism undefined what allows for applying this program beyond medicine despite biological and social mechanisms are very distinct phenomena. In later work, the minimal definition of mechanism from the complex-systems theories (see Machamer et al. 2000) has been endorsed (Russo & Williamson, 2011). The understanding of causal mechanisms as complex systems is exceedingly broad and capable of describing both biological and social mechanisms. According to the complex-systems theories, mechanisms consist of "entities and activities organized in such a way that they are responsible for the phenomenon" (Illari & Williamson, 2012, p. 12; see Glennan, 2017). This interpretative flexibility allows for applying this notion of mechanism to biological and social phenomena (Kelly et al. 2014) and, hence, using evidential pluralism as a guidance for causal inference not only in medicine but also in economics. Considering the fragility of econometric results that lack robustness to minor changes in statistical techniques and sample, economics might be the field where the application of the advice stemming from the normative reading of the Russo-Williamson Thesis (RWT) (Russo & Williamson, 2007) is most

fruitful. However, resolving econometric controversies with mechanistic evidence requires understanding the differences between economic and biological mechanisms and in evidence for mechanisms across the disciplines.

Marchionni (2017) argued that the general notion of mechanism is in agreement with causal mechanisms in the domain of economics and supported the view that economic mechanisms are "complexes of rational agents, usually classified into social categories, whose actions and interactions generate causal relationships between aggregate-level variables." (p. 427). Even though some contemporary economics models include groups of agents driven by limited (bounded) rationality or heuristics (Rabin, 2013), most entities described by economic models are represented as rational agents that face a maximization problem under budgetary constraints. The ontological difference regarding the entities constituting mechanisms is not the only difference that distinguishes economics (or, more broadly, social sciences) from medicine. The two disciplines differ in the types of evidence used for mechanistic inferences. Medical researchers use a broad spectrum of empirical methods such as in vitro and animal models research, biomedical imaging, analyzing tissue samples, etc. (Clarke et al. 2014). Except for pharmacodynamics and epidemiological modeling, most models used in biomedical sciences are of physical nature. For example, Bolker (2009) distinguished between exemplary models (i.e., samples of the same organisms) and surrogate models (i.e., other organisms). In contrast, economics relies mainly on studying mathematical models despite some laboratory and field experiments that also deliver knowledge of mechanisms. Structural equation modeling can be considered still another source of mechanistic evidence (Mouchart & Russo, 2011; Mouchart et al., 2010, 2020; Wunsch et al., 2014) if one identifies data generating process represented by a structural model with mechanism. Such modeling relies on economists' background knowledge and statistical analysis (recursive decomposition, in particular). Considering that the efforts aimed at developing methods for assessing the quality of mechanistic evidence have primarily been concerned with medicine (Parkkinen et al., 2018), applying evidential pluralism to social sciences requires further research in this area.

The heavy reliance on mathematical modeling raises the problem that some mechanistic models in economics can represent possible but not actual mechanisms, i.e., mechanisms that could produce the phenomenon but may not, in fact, operate in the target. This relates to the debate concerned with the distinction between how-actually and how-possibly explanations. Many classical economics models (with Schelling's (1969) checkerboard model being the prime example) have been interpreted as delivering how-possibly explanations only (e.g., Aydinonat, 2007). Such simplifying, minimal models do not provide an understanding of their targets (Fumagali, 2015). Hence, using a model of mechanism as mechanistic evidence requires establishing that the model represents the actual mechanism. One way to proceed with this task is to consider empirical support for the model (Verreault-Julien, 2019). Glennan (2005) offered a useful framework for distinguishing between models of actual and possible mechanisms that encapsulates to analyzing mechanical and behavioral adequacy, i.e., assessing if models accurately describe entities and their interactions and reproduce aggregate behavior of the phenomenon. All mathematical models include simplifications and idealizations (Mäki, 2005) but, to assess if a

given model accurately represents the actual mechanism responsible for producing the phenomenon of interest, it is useful to distinguish between two types of assumptions. Some assumptions (e.g., perfect divisibility of goods) assert that the model is mathematically tractable and do not distort results despite being empirically inadequate. Others influence either mechanical adequacy (by postulating entities or their actions at odds with our current understanding of the mechanism) or behavioral adequacy (by distorting the aggregate effects of the mechanism's work). Such 'crucial assumptions' should be empirically adequate to assert that the mechanism represented by the model is the actual mechanism (Solow, 1956; Maziarz, 2020, p. 125 et seq.). Robustness analysis can help identify those idealizations that are harmless and those assumptions that can distort modeled mechanism (Kuorikoski et al., 2010).

The assessment of the quality of mechanistic evidence is crucial for the successful application of evidential pluralism to social sciences, given that many domains of economic theory are split. In fact, the inconsistencies in difference-making and mechanistic evidence in economics are related (Doucouliagos & Stanley, 2013). Doucouliagos and Stanley (2013) suspect that econometricians are biased in their research to confirm the theory that they accept a priori. However, the opposite direction of the influence, where econometric results inspire mathematical modeling, cannot be excluded based on their evidence. With no exception, the mathematical models concerned with addiction are also split and represent inconsistent mechanisms. Inconsistent mechanistic evidence can help resolve empirical controversies only if one is able to distinguish models representing genuine causal mechanisms from those that characterize another possible mechanism or misrepresent the actual one. Therefore, a careful assessment of the quality of addiction models is needed to resolve the econometric controversy regarding tax elasticities of smoking intensity (see Sect. 5). Otherwise, one faces both an empirical controversy regarding the accuracy of statistical models and a theoretical controversy concerning the question of which of the possible mechanisms is the actual one.

5 Models of rational addiction

Two types of rational addiction models are relevant for the econometric controversy regarding the tax elasticity of cigarette consumption. The classical models and its derivatives represent addicts as utility maximizers choosing between addictive goods and other products. In contrast, Adda and Cornaglia (2006) put forward a model including smoking intensity as an additional factor that can be modified to maximize utility. The two models belong to the rational addiction theory that emerges from the neoclassical tradition relying on modeling economic agents as utility maximizers under a budgetary constraint. This approach remains "the consensus among economists" (Rogeberg, 2020, p. 187) but has been criticized for misrepresenting agents' behavior and decision processes (Dosi & Roventini, 2016). Rogeberg (2004) observed that modeling addicts' behavior as driven by utility maximization is particularly questionable considering the nature of addiction and argued that some economists, despite paying lip service to the realism of such models, participate in a modeling game, while others interpret such models instrumentally and use it for

delivering 'as-if' predictions. The controversy regarding the adequacy of rational addiction theory and the disagreement among models belonging to this framework shows that economists tend to disagree not only about statistical evidence but also (or, possibly, more strongly) about mechanistic evidence and, hence, only evaluation of evidence for mechanisms postulated by theoretical models can break the impasse.

Overall, the studies comparing the predictions of rational addiction theory to empirical data deliver conflicting evidence (Rogeberg, 2020, p. 185). Some econometric studies that test what Glennan (2005) labels 'behavioral adequacy' show that market data are in agreement with rational addiction models. For instance, Bask and Melkersson (2004) estimated a rational addiction model for Swedish alcohol and cigarettes markets and observed that the empirical demand for alcohol agrees with model predictions but smokers behave differently than the model's rational agents. Baltagi and Griffin (2001) obtained similar results for a panel of countries that confirm the accuracy of the rational addiction model for cigarette consumption data. In contrast, Laux (2000) showed that smokers either have preferences unstable in time or are driven by bounded rationality what contrasts with the mechanism represented by the rational addiction model. The results of experimental economics suggest that the predictions of rational addiction theory differ systematically from the actual decisions regarding consumption (e.g., Fehr & Zych 1998). The mechanical adequacy of the rational addiction theory has received support from Smith and Tasnádi (2007), who reviewed biomedical literature discussing how opiates influence the working of the mammalian brain and concluded that the results support the seemingly unrealistic assumptions of the rational addiction theory, such as adjacent complementarity (reinforcement) and the role of conscious decisions in shaping consumption. But their conclusion is in stark contrast with the common-sense view that addiction is an irrational process when one loses control over their own actions and decisions (Rogeberg, 2020).

There are some alternative frameworks for modeling addiction that reject the mainstream neoclassical assumption that economic agents are rational. One plausible framework is offered by the self-control literature dating back to Akerlof's (1991) analysis of procrastination that opposes the assumption regarding intertemporal rationality. According to this approach, individuals choose a series of actions that are not optimal from the intertemporal perspective, i.e., "without fully appreciating how these actions will affect future perceptions and behavior" (Akerlof, 1991, p. 1). The reasons for the intertemporal irrationality are that the sum of minor errors becomes comparatively large due to aggregation and agents fail to accurately discount future costs and benefits, and predict future effects. This conclusion is also applicable to analyzing the behavior of addicts: decisions of substance abusers are maximizing but the optimization problem does not include future costs of addiction. Another framework for modeling addictive behavior emerges from the theory of melioration (Herrnstein & Prelec, 1991). It predicts that sequential decisions distributed over a period of time are suboptimal given the agent's preferences. The theory of melioration is based on evidence stemming from psychological experiments that agent's preferences revealed in a sequence of choices differ from their actual intertemporal preferences. Regardless of their plausibility, the alternatives to rational addiction theory remain unorthodox approaches and, for this reason,

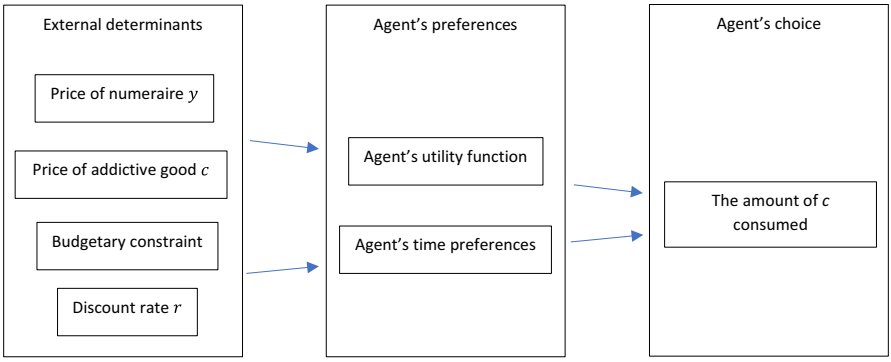

Fig. 1 The mechanism of addict's decision process represented by the classical model of rational addiction

my further analysis focuses on the mainstream neoclassical framework. As I have admitted in Sect. 4, assessing the quality of mechanistic evidence remains crucial for the accuracy of conclusions, especially when a few plausible mechanisms have been described what is the case in regard to the addiction modeling literature in economics.

The classical model of rational addiction has been developed by Becker and Murphy (1988). It is based on the model of changing consumer preferences put forward by Stigler and Becker (1977). According to this model, addicts face intertemporal optimization problem and choose between numeraire (composite good) y that stands for the consumption of all other goods and an addictive good c under a budgetary constraint. In contrast to the standard consumer choice model, the utility brought by the addictive good is altered by its stock (sum of all previous consumption) S. The assumption allows accounting because consuming more c makes consumers more addicted and lowers the utility from future consumption. What follows, the utility at t is given by the following function: $U(t) = u(y(t), c(t), S(t))$. The price of y is constant in time (Becker & Murphy, 1988 p. 677) and, therefore, the ratio of utility to prices of the two goods depends solely on the price of the addictive good c. Similarly to other models of consumer choice, the model predicts that changes in prices of c lead to changes in the proportion of utility delivered by c in comparison to y. In effect, raising prices of c (e.g., in response to changes in taxation) lead to lower consumption of c.

This model can be considered a mechanistic model, similar to other neoclassical models (Maziarz, 2020, pp. 122–126). It represents the abstracted mechanism of addict's decision process, where numeraire y, addictive good c, and the agent are entities, which characteristics determine how they interact. The effect of these actions and interactions is the agent's decision regarding (solely) the amount of addictive good c consumed in each period t, see Fig. 1.

In contrast to the classical model of rational addiction, Adda and Cornaglia (2006, p. 1015) construct the model of compensatory behavior that diverges from the classical model assuming that the addictive good (cigarettes) is homogenous.

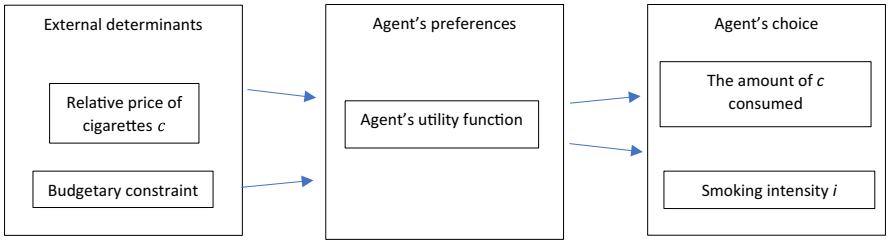

Fig. 2 Adda and Cornaglia's (2006) model including smoking intensity

Instead, the economists include smoking intensity (i) as a second factor optimized by the addict. According to the model, smokers maximize their utility depending positively on the consumption of nicotine n, and numeraire q, and negatively on smoking intensity i. Nicotine intake depends on the number of cigarettes c and smoking intensity i:

$$n = c * i$$

Smokers' choice is subject to the budgetary constraint, where disposable income y is spent on cigarettes (which price is p times larger than q) and the composite good:

$$y = pc + q$$

The model predicts that the effects of higher prices on intensity is ambiguous since more intense smoking is less pleasant. In effect, those who can afford it, consume more cigarettes instead of smoking more intensively. Hence, according to the model, smokers' response to tax hikes is heterogeneous: some smokers lower their cigarette consumption and raise smoking intensity, while others do not change their behavior. This prediction agrees with the model of rational addiction (Becker & Murphy, 1988) that younger and poorer addicts react more strongly to price changes regarding cigarette consumption. However, the reasons (explanation) for these predictions remain in stark contrast: Adda and Cornaglia (2006) model shows that people with lower disposable income are more likely to lower their cigarette consumption because of the compensatory behavior, see Fig. 2. Instead, the classical model relies on the mechanism of discounting future consumption and predicts that such addicts reduce their nicotine intake in response to tax hikes.

Both models, at this point, can be considered as models of possible mechanisms. The actuality of the represented mechanism can be established on the basis of assessing the empirical adequacy of models' crucial assumptions (see Sect. 4). Both models employ the notion of numeraire (composite good) that stands for all goods consumed by addicts other than cigarettes. Obviously, this assumption is at odds with our common-sense understanding of consumer behavior that can choose among a multitude of different goods. However, it does not influence the predictions of the model in a meaningful way. For the question at hand, it does not matter if the addict spends their part of the income not used for buying cigarettes on clothes or bread. The division of income between cigarettes and other goods matters for

model predictions. Other assumptions shared by the two models or (more broadly) the models of consumer choice can also be considered as not being crucial for the behavioral adequacy of the model.

Instead, it is essential to note that the classical model of rational addiction that represents addicts' choice of the consumption of an addictive good (e.g., cigarettes) goes with an unspoken assumption that the addictive good is homogenous. This assumption is not empirically adequate. Even if the addicts could not control for smoking intensity and the amount of nicotine extracted from each cigarette, they could consume tobacco products that include more nicotine and tar in response to tax hikes. However, epidemiological literature delivers evidence that smokers can indeed modify the intensity of smoking and the amount of nicotine extracted from each cigarette (Ashley et al., 2011; Patterson et al., 2003; Russell et al., 1980). This abstraction by omission of the classical model of rational addiction can be considered as a crucial assumption because it omits one of the two ways in which smokers react to changes in cigarette taxation. The epidemiological evidence suggests that this tacit assumption (excluding smoking intensity from the model) is at odds with our current understanding of smokers' behavior, i.e., it is not empirically adequate. What follows, the classical model fails at representing the actual mechanism by not being mechanically adequate.

The mechanistic evidence suggests that smokers choose not only the number of cigarettes smoked but also smoking intensity, and hence the econometric results of Adda and Cornaglia (2006) are additionally supported. The Adda and Cornaglia (2006) mathematical model can be shown to represent the actual mechanism of the addict's utility maximization while the classical model leaves out an essential aspect of the mechanism. Given this, the program of evidential pluralism suggests that the positive dependence between concise tax and smoking intensity can be interpreted as causal. All in all, the mechanistic evidence resolves the empirical controversy regarding estimating tax elasticity of smoking intensity: the result of Adda & Cornaglia is fragile but genuinely causal. In contrast, the lack of the coefficient's significance (Abrevaya & Puzzello, 2012) emerges from the fragility of econometric results, i.e., is an artifact produced by the applied statistical techniques.

6 Concluding remarks

The example of the controversy regarding estimating tax elasticity of cigarette consumption and smoking intensity instantiates a general problem with causal inference based on statistical models. Econometric results are fragile: even minor changes in estimation technique and sample can overturn previous conclusions. Evidential pluralism, which encapsulates to the normative reading of the Russo-Williamson Thesis that causal claims require both difference-making and mechanistic evidence offers an additional constraint for distinguishing between genuinely causal and purely correlational results. Still, further philosophical work is needed to support a stronger claim and argue that the constraint of evidential pluralism also applies to such situations when associations found in data are robust in the sense that no disagreement among statistical models is present.

The main obstacle with applying the program of evidential pluralism to social sciences is that mechanistic evidence can also be divided. The two theoretical models of rational addiction and alternative frameworks considered in Sect. 5 are a good example. In such cases, a careful assessment of the quality of the evidence is needed. Previous efforts have been primarily concerned with assessing the quality of mechanistic evidence in medicine. Considering the differences in mechanistic evidence across the disciplines, applying evidential pluralism to other fields might require developing new approaches to assessing the quality of mechanistic evidence in specific fields. For example, mechanistic evidence in economics comes primarily from mathematical models. Such models represent possible mechanisms, but establishing their actuality, i.e., asserting that the represented mechanism operates in model's target requires additional empirical support. I have suggested that empirical adequacy of crucial assumptions warrants actuality of represented mechanisms but further research is needed to fully understand the limitations and sources of mechanistic evidence in economics. All in all, I believe that relatively little has been said about sources of mechanistic evidence in economics and evaluating their quality, and this creates a promising area for future research.

The controversy regarding tax elasticities of cigarette consumption and smoking intensity shows that the message of evidential pluralism can be especially relevant for those disciplines and fields where interventional studies are infeasible. These include specific areas of economics and, in particular, macroeconomics, but also noninterventional clinical studies in medicine, where the program of evidential pluralism has been started. The reason is that observational studies can report conflicting results because of the malleability of statistical techniques. In such cases, considering the plausibility of mechanisms supposedly producing these conflicting results can tell the result describing a genuine causal dependency from spurious correlation.

Acknowledgements The author acknowledges thoughtful comments received from the two anonymous reviewers for Synthese and his supervisors: Tomasz Rzepiński and Tomasz Żuradzki.

Funding The work of Mariusz Maziarz has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498). Mariusz Maziarz received scholarship from the Foundation for Polish Science (FNP).

Declarations

Conflict of interest The author does not report any conflict of interest in regard to this work.

References

Abrevaya, J., & Puzzello, L. (2012). Taxes, cigarette consumption, and smoking intensity: Comment. *American Economic Review, 102*(4), 1751–1763.

Adda, J., & Cornaglia, F. (2006). Taxes, cigarette consumption, and smoking intensity. *American Economic Review, 96*(4), 1013–1028.

Adda, J., & Cornaglia, F. (2010). The effect of bans and taxes on passive smoking. *American Economic Journal: Applied Economics, 2*(1), 1–32.

Adda, J., & Cornaglia, F. (2013). Taxes, cigarette consumption, and smoking intensity: Reply. *American Economic Review, 103*(7), 3102–3114.

Akerlof, G. A. (1991). Procrastination and obedience. *The American Economic Review, 81*(2), 1–19.

Ashley, M., Saunders, P., Mullard, G., Prasad, K., Mariner, D., Williamson, J., & Richter, A. (2011). Smoking intensity before and after introduction of the public place smoking ban in Scotland. *Regulatory Toxicology and Pharmacology, 61*(3), S60–S65.

Aydinonat, N. E. (2007). Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. *Journal of Economic Methodology, 14*(4), 429–454.

Baltagi, B. H., & Griffin, J. M. (2001). The econometrics of rational addiction: the case of cigarettes. *Journal of Business & Economic Statistics, 19*(4), 449–454.

Bask, M., & Melkersson, M. (2004). Rationally addicted to drinking and smoking? *Applied Economics, 36*(4), 373–381.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy, 96*(4), 675–700.

Becker, G. S., Grossman, M., & Murphy, K. M. (1991). Rational addiction and the effect of price on consumption. *The American Economic Review, 81*(2), 237–241.

Bolker, J. A. (2009). Exemplary and surrogate models: Two modes of representation in biology. *Perspectives in Biology and Medicine, 52*(4), 485–499.

Borgerson, K. (2009). Valuing evidence: Bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine, 52*(2), 218–233.

Broadbent, A. (2013). Stable Causal Inference. In: Broadbent, A. *Philosophy of Epidemiology*. Berlin: Springer.

Çağlar Dede, O. (2019). Behavioral policies and inequities: the case of incentivized smoking cessation policies. *Journal of Economic Methodology, 26*(3), 272–289.

Canali, S. (2019). Evaluating evidential pluralism in epidemiology: mechanistic evidence in exposome research. *History and Philosophy of the Life Sciences, 41*(1), 4.

Caraballo, R. S., Giovino, G. A., Pechacek, T. F., Mowery, P. D., Richter, P. A., Strauss, W. J., & Maurer, K. R. (1998). Racial and ethnic differences in serum cotinine levels of cigarette smokers: Third National Health and Nutrition Examination Survey, 1988–1991. *Journal of American Medical Association, 280*(2), 135–139.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339–360.

Claveau, F. (2012). The Russo-Williamson Theses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 806–813.

Denton, F. (1988). The significance of significance: rhetorical aspects of statistical hypothesis testing in economics. In A. Klamer, D. McCloskey, & R. Solow (Eds.), *The Consequences of Economic Rhetoric* (pp. 163–168). Cambridge University Press.

Dosi, G., & Roventini, A. (2016). The irresistible fetish of utility theory: from "pleasure and pain" to rationalising torture. *Intereconomics, 51*(5), 286–287.

Doucouliagos, C., & Stanley, T. D. (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys, 27*(2), 316–339.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association, 78*(383), 535–543.

Falba, T., Teng, H. M., Sindelar, J. L., & Gallo, W. T. (2005). The effect of involuntary job loss on smoking intensity and relapse. *Addiction, 100*(9), 1330–1339.

Fehr, E., & Zych, P. K. (1998). Do addicts behave rationally? *Scandinavian Journal of Economics, 100*(3), 643–661.

Matheson, J., & Frances, B. (2018). Disagreement. E. N. Zalta (ed.) *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/disagreement/.
Fumagalli, R. (2015). No learning from minimal models. *Philosophy of Science, 82*(5), 798–809.
Gilkes, A., Hull, S., Durbaba, S., Schofield, P., Ashworth, M., Mathur, R., & White, P. (2017). Ethnic differences in smoking intensity and COPD risk: an observational study in primary care. *NPJ Primary Care Respiratory Medicine, 27*(1), 1–6.
Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 443–464.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.
Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling, 7*(3), 319–355.
Goldfarb, R. S. (1995). The economist-as-audience needs a methodology of plausible inference. *Journal of Economic Methodology, 2*(2), 201–222.
Goldfarb, R. S. (1997). Now you see it, now you don't: emerging contrary results in economics. *Journal of Economic Methodology, 4*(2), 221–244.
Henschen, T. (2018). What is macroeconomic causality? *Journal of Economic Methodology, 25*(1), 1–20.
Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics, 38*(2), 257–279.
Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives, 5*(3), 137–156.
Howick, J. (2011). Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making. *Philosophy of Science, 78*(5), 926–940.
Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science, 2*(1), 119–135.
Johnson, R. B., Russo, F., & Schoonenboom, J. (2019). Causation in mixed methods research: The meeting of philosophy, science, and practice. *Journal of Mixed Methods Research, 13*(2), 143–162.
Jung, Y., Oh, J., Huh, S., & Kawachi, I. (2013). The effects of employment conditions on smoking status and smoking intensity: The analysis of Korean Labor & Income Panel 8th–10th wave. *PloS One*, 8(2), e57109.
Kelly, M. P., Kelly, R. S., & Russo, F. (2014). The integration of social, behavioral, and biological mechanisms in models of pathogenesis. *Perspectives in Biology and Medicine, 57*(3), 308–328.
Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science, 83*(2), 227–247.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science, 61*(3), 541–567.
La Caze, A. (2009). Evidence-based medicine must be. *Journal of Medicine and Philosophy, 34*(5), 509–527.
Laux, F. L. (2000). Addiction as a market failure: using rational addiction results to justify tobacco regulation. *Journal of Health Economics, 19*(4), 421–437.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
MacKinnon, J. G., & Webb, M. D. (2017). Pitfalls when estimating treatment effects using clustered data (No. 1387). *Queen's Economics Department Working Paper*.
Maddala, G. (1983). *Econometrics*. McGraw-Hill.
Mäki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology, 12*(2), 303–315.
Manski, Ch. (1994). The selection problem. In Ch. Sims (Ed.), *Advances in econometrics* (pp. 143–170). Cambridge University Press.
Marchionni, C. (2017). Mechanisms in economics. In S. Glennan & Ph. Illari (Eds.), *The routledge handbook of mechanisms and mechanical philosophy* (pp. 423–434). Routledge.
Martini, C., Sprenger, J., & Colyvan, M. (2013). Resolving disagreement through mutual respect. *Erkenntnis, 78*(4), 881–898.
Maziarz, M. (2017). The Reinhart-Rogoff controversy as an instance of the 'emerging contrary result' phenomenon. *Journal of Economic Methodology, 24*(3), 213–225.
Maziarz, M. (2020). *The philosophy of causality in economics: Causal inferences and policy proposals*. Routledge.

Maziarz, M., & Mróz, R. (2020). Response to Henschen: causal pluralism in macroeconomics. *Journal of Economic Methodology, 27*(2), 164–178.
Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology, 21*(1), 54–76.
Moosa, I. A. (2019). The fragility of results and bias in empirical research: an exploratory exposition. *Journal of Economic Methodology, 26*(4), 347–360.
Morgan, M. S., & Magnus, J. R. (1997). The experiment in applied econometrics. *Journal of Applied Econometrics, 12*(5), 459–661.
Mouchart, M., & Russo, F. (2011). Causal explanation: recursive decompositions and mechanisms. *Causality in the sciences,* Oxford: Oxford University Press, 317–337.
Mouchart, M., Orsi, R., & Wunsch, G. (2020). Causality in econometric modeling. From theory to structural causal modeling. *Quaderni - Working Paper* DSE N° 1143. http://dx.doi.org/https://doi.org/10.2139/ssrn.3542299
Mouchart, M., Russo, F., & Wunsch, G. (2010). Inferring causal relations by modelling structures. *Statistica, 70*(4), 411–432.
Nesson, E. (2017). Heterogeneity in Smokers' Responses to Tobacco Control Policies. *Health Economics, 26*(2), 206–225.
Parkkinen, V. P., Russo, F., & Wallmann, C. (2017). Scientific disagreement and evidential pluralism: Lessons from the studies on hypercholesterolemia. *Humana. Mente: Journal of Philosophical Studies, 32*, 75–116.
Parkkinen, V. P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., & Williamson, J. (2018). *Evaluating evidence of mechanisms in medicine: principles and procedures*. Springer.
Patterson, F., Benowitz, N., Shields, P., Kaufmann, V., Jepson, C., Wileyto, P., & Lerman, C. (2003). Individual differences in nicotine intake per cigarette. *Cancer Epidemiology and Prevention Biomarkers, 12*(5), 468–471.
Rabin, M. (2013). Incorporating limited rationality into economics. *Journal of Economic Literature, 51*(2), 528–543.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.
Reinhart, C. M., & Rogoff, K. S. (2010a). Growth in a Time of Debt. *American Economic Review, 100*(2), 573–578.
Reinhart, C., & Rogoff, K. (2010b). *Debt and growth revisited*. Retrieved from: https://mpra.ub.uni-muenchen.de/id/eprint/24376 Access: October 18th 2020.
Reiss, J. (2007). Do we need mechanisms in the social sciences? *Philosophy of the Social Sciences, 37*(2), 163–184.
Reiss, J. (2009). Causation in the social sciences: Evidence, inference, and purpose. *Philosophy of the Social Sciences, 39*(1), 20–40.
Rhemtulla, M., Wijsen, L. D., & Van Bork, R. (2017). Toward a causal interpretation of the common factor model. *Desputatio, 9*(47), 581–601.
Rogeberg, O. (2004). Taking absurd theories seriously: economics and the case of rational addiction theories. *Philosophy of Science, 71*(3), 263–285.
Rogeberg, O. (2020). The theory of rational addiction. *Addiction, 115*(1), 184–187.
Runhardt RW (2020) Philosophy of Causation in the Age of Science. Hermann, J. et al. (eds.) Philosophy in the Age of Science?: Inquiries Into Philosophical Progress, Method, and Societal Relevance. Washington, DC: Rowmann & Littlefield, 159–169.
Russell, M. A. H., Sutton, S. R., Feyerabend, C., & Saloojee, Y. (1980). Smokers' response to shortened cigarettes: Dose reduction without dilution of tobacco smoke. *Clinical Pharmacology & Therapeutics, 27*(2), 210–218.
Russo F (2009). What Do Social Scientists Do?. Russo, F. (ed.) Causality and Causal Modelling in the Social Sciences: Measuring Variations, Berlin: Springer, 15–33.
Russo, F. (2009a). The rationale of variation in methodological and evidential pluralismApplied. Retrieved from: http://philsci-archive.pitt.edu/id/eprint/4992 Access: September 8th, 2020.
Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*(2), 157–170.
Russo F, & Williamson J (2011) Epistemic causality and evidence-based medicine. History and Philosophy of the Life Sciences, 563–581.
Schelling, Th. (1969). Models of segregation. *American Economic Review: Papers and Proceedings., 59*(2), 488–493.

Shan Y & Williamson J (2020) Applying Evidential Pluralism to the social sciences. European Journal for Philosophy of Science. Under Review.

Shiffman, S., & Scholl, S. (2018). Increases in cigarette consumption and decreases in smoking intensity when nondaily smokers are provided with free cigarettes. *Nicotine and Tobacco Research, 20*(10), 1237–1242.

Smith, T. G., & Tasnádi, A. (2007). A theory of natural addiction. *Games and Economic Behavior, 59*(2), 316–344.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The Quarterly Journal of Economics, 70*(1), 65–94.

Spanos, A. (1995). On theory testing in Econometrics: Modeling with nonexperimental data. *Journal of Econometrics, 67*(1), 189–226.

Stigler, G. J., & Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review, 67*(2), 76–90.

Tatsioni, A., Bonitsis, N. G., & Ioannidis, J. P. (2007). Persistence of contradicted claims in the literature. *Journal of American Medical Association, 298*(21), 2517–2526.

Verreault-Julien, P. (2019). How could models possibly provide how-possibly explanations? *Studies in History and Philosophy of Science Part A, 73*, 22–33.

Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science, 32*(1), 33–61.

Wunsch, G., Mouchart, M., & Russo, F. (2014). Functions and mechanisms in structural-modelling explanations. *Journal for General Philosophy of Science, 45*(1), 187–208.