Bergh, Andreas; Wichardt, Philipp C.

Working Paper

On Credibility and Causality in Economics: A Critical Appraisal

CESifo Working Paper, No. 11224

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bergh, Andreas; Wichardt, Philipp C. (2024) : On Credibility and Causality in Economics: A Critical Appraisal, CESifo Working Paper, No. 11224, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/302709

CESifo WORKING PAPERS

11224
2024

Original Version: July 2024
This Version: September 2024

On Credibility and Causality in Economics: A Critical Appraisal

Andreas Bergh, Philipp C. Wichardt

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11224*

On Credibility and Causality in Economics: A Critical Appraisal

Abstract

Establishing causal relationships is a core aspect of empirical economics. Borrowing ideas from the medical sciences, we propose tentative guidelines for reliable causal inferences that cover aspects related to both the study itself and its fit with the existing background knowledge. We argue that the current paradigm in economics (often connected to the credibility revolution) tends to put too much emphasis on internal aspects related solely to the study itself. To illustrate and substantiate this view, we discuss various excellent studies from different fields of economics, which all express causal and highly policy relevant claims. From an applied point of view, the conclusion drawn is that policy implications based on single studies are inherently uncertain, even when the respective studies are state of the art.

JEL-Codes: B410, C900, D900.

Keywords: causality, empirical economics, methodology, credibility.

Andreas Bergh
Department of Economics
Lund University / Sweden & The Research Institute of Industrial Economics (IFN)
andreas.bergh@ifn.se

*Philipp C. Wichardt**
Department of Economics
University of Rostock / Germany
philipp.wichardt@uni-rostock.de

* corresponding author

This version: September 5, 2024
We are grateful to Hossein Asgharian, Jordi Brandts, Pol Campos-Mercarde, Jens Dietrichson, Lina Maria Ellegård, Robert Östling, Simon Reese, Sandra Schaefer, Björn Tyrefors, Roel van Veldhuisen, Erik Wengström and Joakim Westerlund for helpful comments and discussions. Wichardt thanks the Arne Ryde foundation for financial support. Bergh thanks Länsförsäkringars forskningsfond and Jan Wallanders och Tom Hedelius stiftelse. The usual disclaimer applies.

1 Introduction

Answering questions about causal relationships and the projectability of past observational patterns to future situations are core themes in any applied science. This is doubtlessly true also of economics as a social science. In economics, questions about causal processes are raised, for example, in connection with questions such as: Does privatisation affect service quality (e.g. Knutsson and Tyrefors, 2022)? Do monetary incentives crowd out moral or social incentives (e.g. Bowles, 2008; Sandel, 2012)? Does poverty impact mental health (e.g. Ridley et al., 2020a,b)? In fact, the 2021 Nobel Prize in Economics was awarded to David Card, Joshua Angrist and Guido Imbens for their methodological contributions to this endeavour as described in "Answering Causal Questions Using Observational Data" (Committee for the Prize in Economic Sciences in Memory of Alfred Nobel, 2021). A core aspect of their work is the improvement of the credibility of empirical work, i.e. the reliability of (causal) inferences drawn from data (see also Angrist and Pischke, 2010; or Imbens, 2010).

Within economics as well as outside, the corresponding discussion became known as the "Credibility Revolution" (e.g. Angrist and Pischke, 2010; The Economist, 2021; Siddiqi et al., 2022; or Boesche, 2022). Part of this development in behavioural economics was that nowadays larger data sets, more field studies and preregistration of expected effects have become the standard (see also Bryan et al., 2021). And when it comes to establishing causal claims, the reliance on randomised control trials (RCTs) has become even stronger. While these developments have increased the reliability of (experimental) empirical studies in behavioural economics and elsewhere, the increased focus on formal quality criteria of single studies is not entirely without risk. In particular, the knowledge of any science – be it causal or otherwise – never rests in the results of a single study but always in consensus and the synthesis of many studies (see, for example, Kuhn, 1997; Oreskes, 2019), a synthesis the details of which are exposed to constant change and modification (e.g. Neurath, 1921). If the confidence put in single results increases, so does the risk of neglecting relevant knowledge gained elsewhere.

The present paper is motivated by the belief that in order to establish a consensus on reliable causal inferences, there is a need for a pragmatic methodological discussion within economics about what is necessary to differentiate causality from correlation. Accordingly, in Section 2, we present what we consider to be a useful tentative basis for guidelines for establishing reliable causal claims. This proposal is made without entering much into the deeper philosophical debate about causality (for example, Cartwright, 2007b, argues for a pluralistic understanding of causality). Instead, we draw on the famous Bradford Hill criteria (Hill, 1965; see also Howick et al., 2009; or Siddiqi et al., 2022). These, we believe, provide a good basic intuition of what is necessary to convincingly argue for causality instead of mere association; an intuition which is basically reflected also in modern variations of the original criteria (e.g. Howick et al., 2009; Schünemann et al., 2011). We only link (some of) these

criteria to more fundamental philosophical arguments inasmuch as we consider necessary to support our subsequent argument.

In a nutshell, the core point made is that more than single empirical studies are required to make (reasonably[1]) reliable causal claims – and this holds irrespective of the quality of these studies (including RCTs; see, for example, Cartwright, 2007a; or Deaton and Cartwright, 2018). Instead, existing background knowledge from, for example, various similar studies has to support such claims to be reliable (see also Leeson, 2020). And although requiring "support by background knowledge" is not as clear-cut and formal a criterion as one might wish for – given that economists are used to, for example, economic theory providing formal proofs and definitive (theoretical) statements – we consider exactly this partly subjective reference to existing background knowledge to be an important aspect in the assessment of causal claims derived from empirical research.[2]

To illustrate our point, consider a simple daily life example. You've hit the neighbours window with a ball and the window shattered the moment it was hit. Now the neighbour claims it was the ball that caused the shattering and, arguably, few people would claim otherwise. However, the point we want to emphasise is that it is not the clarity of the observation of the event itself nor the perfect coincidence in the timing (the ball hitting, the window shattering) alone that support the causal claim made by your neighbour.[3] Various pieces of background knowledge are implicitly used as well: that (common) glass is fragile, that glass often breaks when hit hard, that balls (as yours) are sufficiently strong to break glass, etc. All this is known either from experience or from credible hearsay. And all this enters into the judgement of the causal claim made by your neighbour. Put simply, this is essentially not different in science, except that "credible hearsay" requires references from respected, peer-reviewed journals (if it has not yet become "common knowledge in the community"). The criteria discussed in Section 2, therefore, cover both types of aspects, those related to the study itself, i.e. the equivalent of observing your ball hitting the window, and those addressing the integration of the results into a broader scientific context or "hearsay".

Note how the emphasis on broader external criteria for the reliability of a causal claim differs from the discussion about the credibility generated by randomised study design regarding experiments or quasi-experiments and the special status of RCTs (e.g. Cartwright, 2007a, 2007b; Deaton, 2010; Heckman and Urzua, 2010; Imbens, 2010; Deaton and Cartwright, 2018; Boesche, 2022). For example, Imbens (2010, p. 400), referring to what he calls the

[1] "Reasonably" is added here (and dropped later) to acknowledge that the question how to judge "reliable" may not have an objective answer.

[2] See Wichardt (2014) for a related argument regarding theoretical arguments in economics.

[3] If there is doubt that observing the situation alone – the physical contact and the simultaneity in time – is not enough, consider a doctor trying to keep a patient alive using cardiac massage (physical contact). If he fails, few people would argue that his touching the patient caused his death (despite the simultaneity of death and contact). Note that referring to the heart failure starting before the massage and claiming that this would have caused the death anyhow could be met by the question how we know that and how we know that nothing similar was going on for the window. The answer: From background knowledge.

"natural experiments revolution", writes "By emphasizing internal validity and study design, this literature has shown the importance of looking for clear exogenous sources of variation in potential causes." We have no intention of arguing against the benefits of randomisation in the design of single studies or the general importance of careful study design (and analysis). Instead, what we want to draw attention to is the additional relevance of existing background knowledge for the reliability of causal inferences. In that sense, we see our pragmatic proposal as both compatible with the "natural experiments revolution" and ideas expressed by its critics (e.g. Deaton, 2010; Deaton and Cartwright, 2018). For instance, giving the example of smoking having a causal effect on lung cancer, Imbens (2010, p. 401) states "[...] history abounds with examples where causality has ultimately found general acceptance without any experimental evidence.". Relatedly, Deaton (2010, p. 426) in connection with expressing his reservation regarding the reliance on RCTs writes "I shall argue that the analysis of projects needs to be refocused toward the investigation of potentially generalizable mechanisms that explain why and in which contexts projects can be expected to work.", a statement we interpret as implicitly asking for the integration of external background knowledge. Similarly, the proposal to see causality as a pluralistic concept (Cartwright, 2007b) appears to be compatible with the reliability of (practical) causal claims to derive also from external (pluralistic) background knowledge.

In order to illustrate our argument, we discuss five intricate and well published empirical studies from economics making causal claims in order to illustrate our point in Section 3. All papers discussed address highly policy relevant topics so that the claims made about causality are by no means innocuous and eventually inconsequential. For the purposes of the discussion, we briefly summarise the key arguments made and argue that if also more external criteria were considered, the causal claims made would become less clear than is suggested by the authors. It should be clear, though, that the intention of doing so is not to criticise the authors for making their claims (nor the reviewers or editors for letting them pass - or possibly even asking for them). Rather, our aim is to suggest that the respective causal claims, which need not be wrong, would have been made more cautiously if the general debate about what it means to establish causality and the relevance of external criteria had been more prominent. Emphasising this aspect appears desirable to avoid possible subsequent mishaps regarding potential policy implications. Note that, for example, the currently popular idea of nudging (Thaler and Sunstein, 2008) as a policy tool is based on much confidence in empirical scientific knowledge.[4]

Apart from adding a word of caution to various causal claims made, the emphasis on the dual nature of supportive information for causality – quality of the specific study and back-

[4]The idea of nudging is to modify decision environments based on (empirical) scientific knowledge about behavioural biased so as to improve the outcomes for the decision makers (cf. Thaler and Sunstein, 2008); for a critical discussion see, for example, Hausman and Welch (2010), Guala and Mittone (2015), Infante et al. (2016), or Kemper and Wichardt (2024).

ground knowledge – implicitly entails also a broader methodological comment. In particular, it expresses a word of caution regarding a potential overemphasis of quality criteria for single empirical studies as a result of the credibility revolution. Naturally, making stronger quality demands for empirical studies is also a matter of resources. Accordingly, stronger quality demands are likely to result in fewer (and supposedly more seemingly promising) studies. Hence, stronger demands may run the risk of fostering more uniformity instead of heterogeneity as, for example, asked for by Bryan et al. (2021). This is especially true of a science which strongly relies on few top journals. Yet, economics is a rather young social science with an arguably still growing understanding of the generally very complex processes driving social (economic) interaction. In view of this, we believe that too strong criteria for single studies may run the risk of loosing valuable background information. Metaphorically speaking, doing so may improve the clarity about a specific point in an intricate painting about which, however, in general still little is known – and it may do so at the expense of a better idea of what the overall picture actually is about.

The discussion in economics is full of topics which are highly relevant for the organisation of society. Wealth and life-satisfaction, privatisation of public services, driving forces of populist voting, long-term risks of using monetary incentives, the impact of trade on polarisation, or poverty and (mental) health are discussed among the examples. Others include, for example, gender and educational achievement (e.g. Delaney and Devereux, 2021), or determinants of trade (e.g. Söderlund, 2023; both references referring to causality). Reliable knowledge about relevant causal relations is desirable in either case. Yet, all of these topics are rather complex and findings are rarely fully aligned – and may even vary depending on the research group studying them (Menkveld et al., 2024). Hence, it is commonly the broader surveys that provide reliable pictures of probable causal relations. Yet, many of the empirical studies, for example, summarised in the book "Behavioral Game Theory" (Camerer, 2003), most likely would not conform to today's quality criteria in the community. Still, taken together their quantity and diversity provide highly valuable information about many general tendencies in behaviour which are relevant for economics.

To wit, the bottom line of the present argument can be seen as (a) asking for a more explicit discussion about guidelines for establishing causal claims, and (b) a cautious warning that focusing too much on refinements of single items of such guidelines comes at a cost which, at some point, may outweigh its benefit.

2 Causality - Some Pragmatic Ideas

Addressing the reliability of causal inferences from observational data, we start with a proposal for practical guidelines for how to reliably differentiate causality from correlation. In doing so, we sidestep both primarily statistical questions and deeper philosophical issues of the topic; the interested reader is referred to Holland's (1986) paper "Statistics and Causal

Inference" and Cartwright (1989) or Woodward (2005) for two enlightening philosophical discussions (among many[5]). Instead, we acknowledge the fact that scientific studies almost naturally have limitations and that rarely all causal processes possibly leading to an effect will be known and considered, and ask what is necessary to nevertheless make reliable causal claims as best we can.

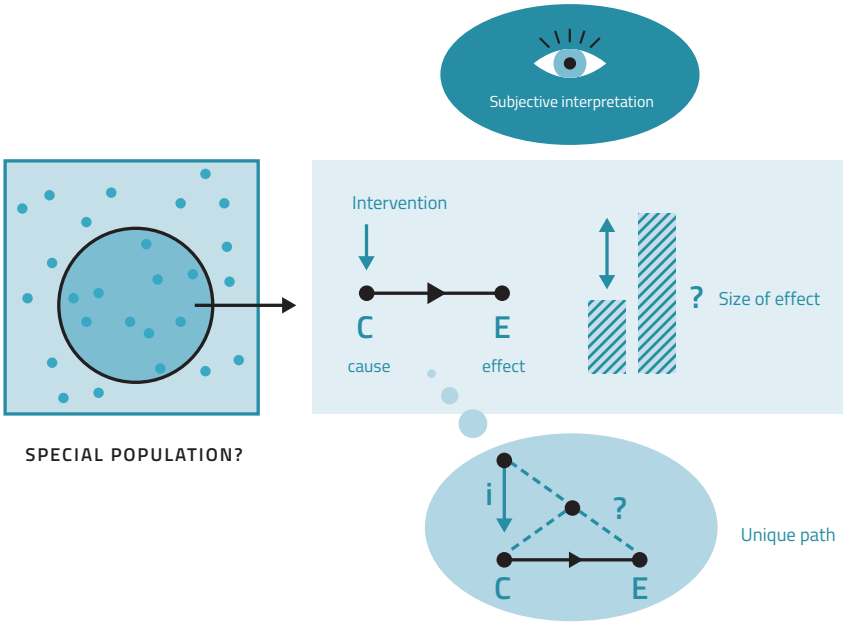

Figure 1: Illustration of some possible pitfalls in causal inference. A sample of some population is exposed to a certain intervention (i) to modify a putative cause (C) and measure the consequence on the effect (E). Is the sample special? Are there other routes along which i could affect E? Is the size of the effect reasonably strong with respect to sample structure and size? Could the subjective view of the researcher(s) bias their interpretation? What gives confidence regarding specific claims based on an observation is a priori knowledge derived from sources *external* to the study at hand.

In order to do so, we outline some ideas based on what is known as the Bradford Hill Criteria in the medical sciences (Hill, 1965; see Table 1). The criteria were introduced to clarify how to reliably separate causal empirical relations from mere associations; see Figure 1 for an illustration of common sources of doubt. While the initial proposal of the criteria was primarily inspired by questions arising in contexts where experimental data are scarce, the criteria still provide a reasonable basis for a pragmatic discussion of reliable causal inferences from any empirical data (see Bryan et al., 2021, for a discussion of some of the more general

[5]See, for example, for Ross and Woodward (2023) for a general discussion and further references, or Maziarz (2020) with a discussion focusing on economics. See Dague and Lahey (2019) for an instructive summary of methods. See Cartwright (2007b) for an argument for a more pluralistic view of causation which, in our view, further supports the idea that from a practical perspective reliability of causal claims is strengthened by meeting also external credibility criteria.

problems).[6] Moreover, although there is neither a "biological gradient" in the social sciences (the dose-response curve is rather an incentive-response curve) nor do the social sciences possess nearly as precise knowledge about interpersonal processes as the medical sciences have about physiological ones (not least because bacteria and viruses have less complex social habits), we believe that the criteria translate naturally to economics. Yet, it should be noted that no one of the criteria is considered necessary or sufficient for causation (nor is any combination), as strict formal rules were not what Hill had in mind – and neither do we.

Strength	Size of overall effect (possibly relative effect)
Consistency	Repetition by different people, in different places, circumstances, times...
Specificity	Specficity of putative causal factors and effects
Temporality	Temporal order (effect after putative cause) and no common causes
(Biol.)Gradient	Size of specific association "dose-response curve"
Plausibility	Plausibility of causation given our present knowledge
Coherence	Consistency with generally known facts
Experiment	Confirmation of association in experiments
Analogy	Existence of analogous associations (depending on circumstance)

Table 1: Bradfod Hill Criteria for Reliability of Causal Inferences (cf. Hill 1965; own summary)

Put simply, the aim of the Bradford Hill Criteria, when introduced in 1965, was to establish some form of meaningful practical guideline about when to judge some association in the data as causal (Hill, 1965). While some variants have been discussed over the years (e.g. Howick et al., 2009),[7] the original criteria are still used as a guideline today (e.g. Siddiqi et al., 2022). In the following, we refer to the original as we believe that this makes the core ideas most transparent. The main distinction we make – in criteria referring to internal and external aspects of a certain study – can be made analogously for more recent variants (e.g. Howick et al., 2009; Schünemann et al., 2011).[8]

[6] Bryan et al. (2021) discuss problems of lacking heterogeneity in empirical data arguing that a lack of replicability may often derive from different background characteristics – which, following our argument, cannot be known without further background knowledge as asked for among the Bradford Hill Criteria.

[7] Howick et al. (2009) provide a more specific mention of mechanistic evidence and nice graphical illustrations of the concept; they drop specificity and experiment which is included in what they call direct evidence.

[8] For example, Howick et al. (2009) consider 'Direct, Mechanistic and Parallel', where 'Direct' summarises aspects which are internal to the study, 'Parallel' aspects which are external and 'Mechanistic' can be seen as capturing a bit of both. The authors also relate their criteria to the Bradford Hill original ones (p.187).

Roughly speaking, the Bradford Hill Criteria can be assigned to two broad categories; see Table 1 for reference. One category would cover aspects referring to the study at hand, namely *strength*, *specificity*, *temporality*, *gradient*, and *experiment*. Following Hill (1965), *strength* refers to the size of the effect (primarily the absolute size, but depending on circumstances also the relative effect), with larger effect sizes given more reason to assume causality. *Specificity* in turn concerns the putative causal factors considered such as a certain group, area, etc., again with higher specificity of the association strengthening believes in its causality. The temporal order with the effect following the putative cause while accounting for potential common causes such as self-selection into certain conditions or environments is accounted for in *temporality*. Furthermore, the strength of the response to a variation in the putative cause – what might be referred to as "incentive-response curve" in economics – is covered by *gradient*, with stronger reactions strengthening claims at causality. And, finally, *experiment* emphasising that controlled variations in conditions allowing for stronger support of causal claims. We believe these criteria, which all refer to a specific empirical result, are the aspects most focused on in connection with the credibility revolution. In the sequel, we refer to this category as referring to *internal* aspects of the study (see Remark 1 for a comment on the difference to internal validity).

The other category, then, would include the items linking the association under consideration to related knowledge; this would cover *consistency*, *plausibility*, *coherence* and *analogy*. Again following Hill (1965), *consistency* asks whether a certain association has been observed repeatedly "by different persons, in different places, circumstances and times" (Hill, 1965, p.296).[9] As Hill (1965, p.297) writes "I would myself put a good deal of weight upon similar results reached in quite different ways [...]". *Plausibility* in turn concerns the congruence of the considered causality with the current (general) knowledge in the respective scientific community, i.e. is what is suspected plausible given what we know. Relatedly, *coherence* addresses the question if the supposed causal relation conflicts with known more specific facts about considered association – in the original case "of the natural history and biology of the disease" (Hill, 1965, p.298), in the economic setting rather about the theoretical and empirical knowledge about the respective association; again stronger coherence making causality more likely. Finally, *analogy* concerns the fit of the suspected causation with associations which in some meaningful sense (again implicitly referring to the background knowledge of the scientific community) are similar; better fit providing stronger support for causality. In the sequel, we refer to this category as *external* aspects of the study.

Note that, for example, Imbens (2010, p. 405) also mentions the relevance of similarities of populations, goals, and closeness in time between a study and its validity for external use

[9] In a similar vein, the relevance of repetition of findings in different circumstances is also emphasised by Dague and Lehay (2019) as strengthening external validity of empirical results. Their discussion focuses more on experimental and statistical methods, though, an less on the interplay of inferences drawn from a certain study and the general background knowledge this relates to.

in a certain context. All of these facets are close to some criteria entailed in the external aspects discussed above. Yet, for the present purposes it is important to distinguish between our discussion of reliable inference at causality and the validity debate (see Remark 1 below).

Remark 1 (Relation to Internal/External Validity)**.** *For the purposes of the present discussion, we consider criteria for reliable inferences at causation related to internal and external aspects of a study. Yet, we do not directly relate these to internal and external validity. This is due to a difference in focus. In the validity discussion, causality is primarily connected to internal validity (see, for example, Vazire et al., 2022) and usually does not depend on, for example, effect size or gradient. External validity, in turn, is commonly associated with the generalisability of results (e.g. Vazire et al., 2022), and is often seen as being partly in conflict with internal validity when it comes to a specific study (e.g. Guala, 2005; the idea roughly being that a cleaner lab design, needed to test a putative mechanism, implies a "longer way" to generalise findings to a "messy" outside world, in which more things might interfere and exact mechanisms are harder to identify – and vice versa).*

In the present discussion, by contrast, both internal and external aspects of an empirical study are considered inasmuch as they support the reliability of a causal claim derived from the data – irrespective of the specificity or generality of the claim. Accordingly, an empirical study being experimental is only seen as one *argument in favour of a causal interpretation (cf. Table 1). And an experimental study with higher internal validity would be a stronger such argument. But as also emphasised by Hill (1965; see also below) – and this is crucial here – more arguments are needed for a causal claim to be truly reliable (as in the introductory example about your neighbour's window) and no combination of arguments can a priori be seen as necessary and sufficient. Naturally, a more specific causal claim might be easier to motivate by reference to internal aspects of a study, while a more general causal claim derived from a single study might call for more support through external aspects and background knowledge (see Section 3). Yet, the details always depend on circumstances (including existing background knowledge) and there never is a tension or tradeoff between internal and external arguments (as for validity). They merely add to one-another.*

Following the Bradford Hill criteria, the credibility of a causal claim derives from both the internal and the external aspects related to the corresponding empirical study. As we illustrate in the sequel, it is in particular causal claims that are less specific which call for a stronger reference to external aspects in the overall argument.

Moreover, referring to all criteria and the questions entailed in them, Hill (1965, p.299) writes "No formal tests of significance can answer those questions. Such tests can and should remind us of the effects that the play of chance can create, and they will instruct us in the likely magnitude of those effects. Beyond that, they contribute nothing to the 'proof' of our hypothesis." Thus, although economics is used to formal proofs and strict empirical standards, and certainly for good reasons, establishing causal claims, according

to Hill, always requires some form of informal, eventually subjective judgment as well. As Rohrer (2018, abstract) put it when discussing correlation and causation, "Drawing valid causal inferences on the basis of observational data is not a mechanistic procedure but rather always depends on assumptions that require domain knowledge and that can be more or less plausible. However, this caveat holds not only for research based on observational data but for all empirical research endeavors.".[10] We believe it is worthwhile and, in fact, favourable to acknowledge this fact rather than ignoring it.

Finally, as randomised control trials (RCTs) are often viewed as the gold standard of experimentation (and internal validity),[11] it is noteworthy that the particular relevance of background knowledge, i.e. external aspects, has also been emphasised in the more specific debate about the role of RCTs. For example, Deaton and Cartwright (2018, p. 2) write "We argue that any special status for RCTs is unwarranted. [...] When little prior knowledge is available, no method is likely to yield well-supported conclusions." They later add that "You cannot know how to use trial results without first understanding how the results from RCTs relate to the knowledge you already possess about the world, and much of this knowledge is obtained by other methods." (Deaton and Cartwright, 2018, p. 3). Regarding the particular relevance of external credibility in connection with RCTs, this becomes most transparent in the last sentence of the abstract of their paper "RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not 'what works', but 'why things work'."

In a similar vein albeit focusing on quasi-experiments, Boesche (2022) highlights the implicit use of causal homogeneity in the extrapolation of effects and writes (p.10) "While it is true that a single quasi-experiment is insufficient to justify such assumptions [of causal homogeneity], a multitude of consistent quasi-experimental results may provide ampliative evidence for causal homogeneity across the whole population."; causal homogeneity here referring to the assumption that sufficiently large subpopulations are affected the same way by some treatment. He later continues (p. 11) "[...] each successful replication of a causal estimate for a different subpopulation offers some inductive evidence that some assumption of causal homogeneity is justified.". The latter statement, we see in close connection to *consistency* and *analogy* as gathered within the criteria referring to external aspects of the study. Thus, the quotes from the more philosophical debate provided in the above paragraphs strengthen our point that external consistency is not just a nice add-on for causal claims and indicate that its relevance derives from deeper considerations.

Put differently, the more practical point we want to emphasise in view of the discussion

[10] Rohrer (2018) provides an excellent discussion of potential pitfalls in the search for causal relations using directed acyclic graphs.

[11] See, for example, Dague and Lahey (2019) who discuss different aspects of causal inference methods in applied microeconomics with a focus on inferences derived from single studies.

around the credibility revolution in economics is that, from our reading of the literature, aspects concerning the external support of a study are usually less explicitly discussed in economics. While there commonly is an extensive emphasis on and debate of criteria regarding internal aspects (sample size, preregistration, treatment variations,...), intuition gained from related research – let alone research conducted in other sciences – is often frowned upon and is hardly admitted to counterbalance potential lacks in internal aspects. As we argue in the next section, this partial neglect of the external criteria in the discussion may foster causal claims of potentially unwarranted strength, despite the internal soundness of the respective studies being very high. And, in a similar vein, the lighthearted rejection of studies with potential deficiencies regarding internal criteria without giving much thought and credit to their support through external ones comes at a cost – of both ideas and a more comprehensive background knowledge.

As Cartwright (2019, preface) puts it "many of our most useful principals (including high-level 'laws' of physics) are [...] (in the words of Pierre Duhem) 'symbolic representations' that we use to model, predict, and navigate the world;". With this in mind, our argument could somewhat sloppily be rephrased as a question: Given that the world is hardly ever exactly the same, wouldn't it be reassuring to know that the principals we use to navigate it have proved successful under more than one very specific set of circumstances before we try to convince others of their truth or usefulness? The likewise sloppy summary of the intuition of the Bradford-Hill criteria probably could be a simple: Yes, it would.

3 Five Case Studies

In this section, we present and discuss a variety of cases of well received studies in order to illustrate the relevance of external criteria for causal inferences. All but one of these studies present specific data used to draw causal inferences. The remaining one, discussed last, is a review. All studies are chosen because they are well thought through, have passed excellent peer review and address highly relevant questions. Moreover, they all suggest causal inferences which – apart from the review – rely mainly on the reference to internal aspects of the respective study. Lacking reasonable support from external criteria and mostly even emphasising disagreements in the literature, we argue, they all provide reason to ask for some more caution regarding the reliability of the causal claims made.

The review discussed in Section 3.6 indicates that causal statements made in economics are not generally based on internal aspects of a study – even by readers within economics. In this example, the authors explicitly refer to single studies as *evidence* for causality, i.e. they implicitly acknowledge the need for external support. In our view, it is this distinction between treating studies as evidence for causality vs. establishing causal claims based on single studies that is crucial. In this distinction, the evidence for causality may well be of different value, depending on the internal aspects of the respective study. However, as with

inductive inferences in general, a reliable inference at causality seems difficult to make based on a single point of evidence.

With this in mind, the following discussion of different studies with arguably strong support with respect to internal criteria is presented to support the view that, if causal inferences are sought after, a more explicit discourse within economics of how to understand causality and how to judge the reliability of causal claims seems warranted; see Table 2 in Section 3.5 for an overview.

3.1 Lottery Wins and Psychological Well-Being

As a first example, we consider a paper by Lindqvist et al. (2020) which studies the connection between wealth and individual well-being. In order to avoid typical endogeneity problems, Lindqvist et al. use data from a Swedish sample of people who had gained considerable wealth through lottery wins in the past to assess the influence of wealth on well-being. Given the distribution of lottery wins over the past (5-22 years before the questionnaire), they assess effects of such windfall-gains on, for instance, life-satisfaction, happiness or mental health over time.

The paper is chosen for the purposes of the present discussion, as Lindqvist et al. provide an excellent analysis and focus on establishing a causal relation. The choice of instrument (lottery wins), for example, is motivated by stating (2020, p. 2704) "To credibly estimate the causal effects of wealth, it is necessary to isolate a source of variation in wealth that is plausibly unrelated to other determinants of well-being." Moreover, the authors write (p. 2705) "... our study compares favourably both in terms of statistical power and the credibility of our causal inference." and the caption of a figure illustrating main effects reads (p. 2712) "Causal impact of wealth on primary outcomes and domain-specific measures of life-satisfaction.". In connection with their statement "There is clear evidence that wealth improves people's evaluations of their lives as a whole." (p. 2704, referring to their data) it seems justified to interpret exactly this statement as causal in meaning.

All in all, the results presented by Lindqvist et al. are very clear and nicely relate to the broader discussion about the impact of wealth on well-being. However, despite the relevance of the findings and the fact that the suggested causal influence of wealth on well-being may well exist, we believe that a reliable causal claim would necessitate more external support for the results. As mentioned in Section 2, part of what gives credence to causal claims derived from empirical data is their integration into what is already known (via consistency, plausibility, coherence, and analogy); see Bryan et al. (2021) for a related discussion about the need for heterogeneity. This integration, we believe, is too weak in the present case to (reliably) justify the causal claims made.

In order to illustrate this point, we offer a possible alternative explanation for why lottery winners in the sample of Lindqvist et al. look more favourable on their lives. For example,

Lottery winners may feel that life has given them a valuable present, which improves their life in various ways.[12] Wether this is due to this 'present' being wealth or whether an unexpected recovery of a close and beloved relative from a serious illness might have had the same effect, to us, is difficult to judge.

Simply put, it is not obvious that causal effects of lottery wealth generalize to wealth in general. Lindqvist et al. refer mostly to general wealth effects, but lottery wins are a combination of wealth and chance matters. And while the alternative interpretation need not be any better than the original – and likely is not – it is provided to exemplify where more confirmatory external background knowledge would be needed to judge more reliably.

More specifically, the broader claim – loosening the specificity of the association (referring to essentially unexpected lottery gains, during a certain time period, for a sample from the Swedish population,...) – increases the need for stronger integration of criteria related to external aspects of the study (such as coherence and consistency). Once the step is taken from an unexpected gain to general wealth, or from experiences made in a certain time period to a general claim without any reference to time, or from a certain subject group to people in general, questions arise. Does the chance aspect matter? Was the time period considered special? Could the effect depend on the choice of a subject group? The external validity one would grant the claims made, i.e. in how far results can be extended to other contexts, then strongly depends on the support these claims would receive from external aspects of the study, i.e. in how far existing knowledge supports them. For instance, how much is known from similar studies, made in similar circumstances, by different people, etc. (like in the case of the ball and the window).

In fact, to some degree, later passages of the paper may indeed be interpreted as Lindqvist et al. being aware of possible alternative interpretations. For example, they write (p. 2722/3) "Overall, our study advances understanding of the broader question of why wealth and well-being often go hand in hand by providing credible and precise estimates of the long-run causal impacts of large changes in wealth in a sample of Swedish lottery players." While the statement implicitly takes causality as given (which seems contentious), it does refer not to wealth per se but to 'large changes in wealth'. Similarly the statement (p. 2723, italics in original) "We find that lottery wealth causes sustained increases in *Overall LS*." (p. 2723, italics in original; LS = Life Satisfaction) leaves room for 'lottery wealth' being interpreted with respect to the change rather than the wealth itself.

Nevertheless, perhaps because it is common to do so in economics, the focus of the study is more on suggesting clear conclusions than alternative interpretations and on differentiation from others than on emphasising similarities. Following the previous discussion and the criteria introduced above, however, reliable causal (broader) inferences would necessitate more references to analogous findings and arguments suggesting the independence of the

[12] We should note here that Lindqvist et al. (2020) find clear positive effects only for lump-sum prizes not for repeated monthly payments which they speculate may have to do with non-linearities of wealth effects.

effect of chance aspect in wealth acquisition, i.e. a stronger reliance on external aspects. Seen from that angle, we would argue that the study by Lindqvist et al. is rather an excellent data point of evidence for causality than establishing a reliable causal claim.

3.2 Privatising Ambulance Services

As a second example, we consider a paper by Knutsson and Tyrefors (2022) in which they study the quality and efficiency of public and private firms using evidence from a partial privatisation of ambulance services in Sweden. For their study, Knutsson and Tyrefors exploit the fact that around Stockholm the assignment of patients to ambulances of an either public or private provider can be treated as effectively random. While they find that "private ambulances reduce costs and perform better on contracted measures such as response time" they "perform worse on noncontracted measures such as mortality" (both cited from the abstract). These findings are obviously highly relevant economically and socially, as is also emphasised by the authors, for instance, in the introduction to the paper.

In view of a causal interpretation of their results Knutsson and Tyrefors write (p. 2215/6) "..., we find that private ambulances cause increased mortality.". Interestingly, with respect to the present discussion about internal and external credibility, they argue for the reliability of their causal claims by saying (p. 2217) "Our confidence that we have estimated causal effects is supported by a credible study design." – a passage clearly emphasising only internal aspects of the respective study. Thus, the paper can be interpreted as providing highly relevant and interesting results from an excellent study which are interpreted causally based on internal criteria.

Moreover, when proposing an explanation for their results (p. 2257), they say "Causal estimates are complemented by descriptive evidence suggesting that private firms have higher turnover, require more hours from their staff, rely more on overtime, and provide less on-the-job training." (p. 2257). In connection with a statement from the introduction (p. 2214) saying "This article furthers the literature by providing a credible empirical evaluation of the performance of public and private for-profit providers in an acute health care environment." the description of the results arguably entails a form of policy relevant valuation based on an allegedly causal relation between ambulances being private and probability of death.

While again the internal strength of the study is very high, the authors themselves state that "The evidence on outsourcing in health care is concentrated in nursing home markets and mixed (...)." (p. 2214), i.e. external support of the results is scant. Thus, while Knutsson and Tyrefors add high quality evidence to an important ongoing debate, we would argue that given the lack of external support for their claim, the association between private ambulances and mortality need not be causal in the way suggested. For example, in principle it seems possible that privatisation of some ambulance services leads to an increase in performance of public ones (e.g. to prevent further privatisation, or to prove common arguments for privatisation wrong, etc.). Expressed technically, the stable unit treatment

value assumption (SUTVA; Rubin, 1980, 1986) would be violated. This would also show as private ambulances performing worse than public ones, but still privatisation could have a positive effect overall. While this alternative hypothesis to some extent questions internal aspects of the study (privatisation affecting both groups), we would not see that as serious problem let alone as a request for even higher standards. Rather, although a violation of SUTVA would question the reliability of *statistical* inferences, we again would argue that especially more reference to external aspects (e.g. coherence and analogy) is needed for causal claims to be reliable for this important topic.

3.3 EU-Funds and Populism

In an intriguing paper on the effects of fiscal redistribution on voting behaviour, Albanese et al. (2022) study the connection between EU funds and populism. Both the EU's attempts at equalisation between regions and the current increase in populism in many democratic countries are topics of constant public debate, and so are the potential economic driving forces of populism. In their paper, the authors argue that EU financing indeed has a tangible negative effect on populism; the considered outcome variable being electoral outcomes at the municipality level, where the degree of populism for each party is determined using the anti-establishment score developed by Norris and Inglehart (2019).

In fact, the authors suggest that the connection between EU support and populism found in their data is causal, writing (Abstract) "...larger EU financing caused a drop in populism of about 9% of the mean of the dependent variable.". In order to manifest the effect, Albanese et al. use "a spacial discontinuity design (RDD) to establish causality between funds and populism" (p. 2) exploiting a regional border determined by the eligibility of different Italian regions for funding. Moreover, the analysis includes a variety of robustness tests to support its internal validity. Eventually, Albanese et al. conclude by saying (p. 18), "We have shown that financial transfers injected by the EU regional policy toward Italian lagging areas have had the ability to reduce the anti-establishment component of populism" and by emphasising the political relevant of their results.

Regarding the external reliability of the causal claims made, Albanese et al. (p.2) report mixed prior evidence with the two studies deemed closest to their own work showing no correlations. In fact, also Albanese et al. emphasise the differences of their results in comparison to earlier studies. Accordingly, external criteria – at least consistency and analogy – would call for caution regarding reliable causal inferences. Given the political relevance of the topic and the implied policy implications, we believe this to be important to keep in mind.

In fact, the authors themselves note that "Our interpretation is that in the treated units the populist content of vote is lower because economic insecurity is lower, thanks to disbursements from the EU. However, an alternative potential explanation may be at work. If

people living in the Convergence Objective municipalities would be fully aware that funds come from the EU, they could simply react by increasing their pro-European attitude. In equilibrium, if this attitude is correlated with *Populism*, the underlying story and the related policy implications would be somehow different." (p. 15, emphasis in original). And yet another possible interpretation could be that regions receiving no (or less) EU-funds are fertile grounds for populists who capitalise on anger, envy or relative deprivation towards regions that do receive such funding as the estimates presented in the paper do not inform us about the level of populism that would prevail in the absence of EU funds; this again would call SUTVA into question (Rubin, 1980, 1986).

Note that the question whether any of the alternative interpretations is reasonable or even better is beyond the point of the present discussion. What is relevant is that different interpretations are possible and that due to a lack of external support of the causal claims made (similar findings under comparable conditions elsewhere) reliable judgements about underlying causal processes are difficult to make.

3.4 Monetary Incentives for Covid-19 Vaccination in Sweden

The next example is taken from the discussion about economic incentives to induce higher vaccination rates against Covid-19. Obviously, the question was of exceptional policy relevance at the time. In an influential study, Campos-Mercade et al. (2021) were able to show that monetary incentives (a one-time payment of 200 SEK (approximately 20 dollars) for taking the first shot) indeed increase corona vaccine intakes among the targeted population in Sweden. As their results were met with skepticism regarding the long-term effects of economic incentives for behaviour that arguably ought to be morally driven, they published a follow up study addressing this concern (Schneider et al, 2023). This second study is what we focus on.

Naturally, the question of whether economic incentives have the potential to crowd out other moral or social incentives to engage in a certain behaviour is of considerable policy relevance (see also Bowles, 2008; or Sandel, 2012). Hence, it is somewhat surprising that the Schneider et al. (2023) paper is titled "Financial incentives for vaccination do not [sic!] have negative unintended consequences". Moreover, in the abstract, Schneider et al. (2023) write "Our findings inform not only the academic debate on financial incentives for behaviour change but also policy-makers who consider using financial incentives to change behaviour." and add in the conclusion (p. 532) "Our study provides important evidence that will allow policy-makers to make more informed decisions when weighing the costs and benefits of introducing financial incentives to change behaviour.". Furthermore, the respective evidence is described as causal and suggested to describe an effect with a broader relevance; for example, the authors write (p.526) "Here we report findings from a large-scale, pre-registered study that causally measures the unintended consequences of offering financial incentives to

encourage healthy and prosocial behaviour (n=5,019).", i.e., here, as elsewhere, the reference is made to "financial incentives" in general and not a one-time payment (in the special context of Covid vaccination). Thus, taken together, we interpret these statements as suggesting clear causal inferences with a wider policy relevance.

In view of the present discussion, it is noteworthy that the authors themselves write (abstract, 1st sentence) "Financial incentives to encourage healthy and prosocial behaviours often trigger initial behavioural change, but a large academic literature warns against using them.". The apparent lack of academic consensus, which the authors acknowledge, indicates a considerable lack of external coherence. Nevertheless, the authors base their comparably general claims on one study using a one-time financial incentive of a certain size (200 SEK) for one particular behaviour; also the follow-up behaviour considered is arguably naturally connected and may well be seen as part of the incentivised procedure, namely the 2nd shot of the vaccine.[13] Thus, the specific results of the study – providing a one-time monetary incentive for Covid-19 vaccination does not decrease later uptake of follow-up vaccinations - are interesting and were highly relevant at the time, we believe that skepticism is warranted regarding the inferences regarding causality, and especially the wider causal statements made. The reasons for this again relate to a lack of sufficiently strong external support.

In order to exemplify where external credibility would increase the reliability of the causal inferences made, we want to briefly emphasise two aspects. For instance, the very special situation (the urgency entailed in the Covid-19 pandemic) could be argued for as providing distinctive external justification for introducing one-time monetary incentives, which may interfere with underlying causal processes in a way not considered in the study. In view of crowding out of moral incentives, for example, the long-term effect of an additional monetary incentive under such special circumstances may conceivably be different from the effect of making payments for vaccinations common for a longer time (and then dropping them again); where "conceivably" essentially means "based on unspecified but existing 'knowledge' about / experience with human behaviour". A similar claim could be made about the special subject group (in terms of trust; cf. Green et al., 2011; Larson, 2013), although the authors in this regard provide analogous compatible results regarding the effects of one-time incentives from the US.

Moreover, it seems noteworthy that most experimental studies, if they consider learning / adjustment of subjects to the task, allow for more than one round of trials. For example, Falkinger et al. (2000), studying the influence of an economic incentive mechanism on public good provision, allow subjects several trial rounds. In fact, the results Falkinger et al. find suggest that experience with the economic incentive mechanism to begin with does have a negative impact on contribution behaviour once this is taken away (see also Festre and Garrouste, 2015, for a review of long-term effects of economic incentives; see also Bowles, 2008).

[13]The authors also report some related follow up behaviour for which they find no impact of the treatment. Still inferencs made are based on the one-time payment.

Thus, while we believe that the authors make a valuable contribution to an important discussion by conducting a study with excellent view on internal criteria for causal inferences, we once again would argue that more external support (e.g. coherence and consistency) is needed to reliably establish the causal inferences made.

3.5 A Brief Recap

Before moving on to an example that is more in line with the practical guidelines for causal inferences proposed in Section 2, we briefly recap the main aspects from the preceding discussion. As we hope has become clear in the course of the argument, all studies referred to in Sections 3.1 - 3.4 can be argued for as addressing relevant questions, having high internal validity, being informative, and policy-relevant as they stand. At the same time, all make causal inferences despite a lack of external support for the inferences made, which might be less reliable than a first reading of the papers may suggest; see Table 2 for reference.

In order to clarify the importance of criteria referring to external aspects of a study for our discussion, we want to emphasise that our argument is *not* based on the specific aspects we address for every single study. More specifically, the discussion in Sections 3.1 and 3.4 might be interpreted as simply suggesting unwarranted generalisation. Similarly, the argument provided in Sections 3.2 and 3.3 might be viewed as primarily indicating potentially unnoticed confounders. Both would be beside the point, though. Not only would this put far too much weight on our suggestions or the belief that they might be true. It would also miss that the points focused on – as the studies themselves – are just exemplary to illustrate the need for a good fit of any reliable causal claim with what is already known.[14]

Note that the focus on external aspects also implies that none of the comments made in the previous subsections should be read as asking for more treatments, better statistical analysis, et cetera by the same authors. As we hope the discussion in Section 2 has made clear, a central part of external support is derived from consistency (the repetition by different people, in different circumstances,...) and, relatedly, analogy (analogous associations), here referring to the original Bradford Hill criteria proposed by Hill (1965). No reasonable improvement of any single study can provide that. This simply is where reliable causal inferences require a good fit with the broader background knowledge from the scientific community. This, we believe, is particularly relevant for causal claims made in areas with a high

[14]For instance, generalisation, of course, is a step that is almost always needed to obtain socially relevant insights from any study. How this is achieved in a proper way is what the external *validity* debate focuses on (see Remark 1). The point here is how casual claims need to be integrated into existing knowledge to be reliable. A similar point could arguably be made for many studies which are more careful in their wording regarding causality. For example, an excellent paper from the realm of trade and polarisation with intriguing insights is the paper by Autor et. al. (2020). While the authors are much more careful in the exposition of their results, more reference to compatible evidence supporting effects of "trade exposure" (a term used e.g. in the title of the paper) and how it leads to political polarisation, would be desirable as the results derived in the paper come from the so-called China shock, which may or may not generalize to trade exposure in general.

Paper	**Causal Claim**	**Evidence**	**Alternative Interpretation**
Lindqvist et al. (2020), Review of Economic Studies	Wealth increases life satisfaction.	Comparisons of life satisfaction between Swedish lottery winners of different amounts and between lottery winners and non-winners.	Windfall gain / *lottery wealth* increases life satisfaction
Albanese et al. (2022), European Economic Review	EU funds mitigate populism.	A regression discontinuity design that shows that populist parties are larger in Italian regions that receive less EU-transfers	Receiving limited EU-funds when neighbouring regions receive a lot promotes populism.
Knutsson and Tyrefors (2022), Quarterly Journal of Economics	Private ambulances cause increased mortality (but lower response time).	Analysis of a partial privatisation of ambulance services in Sweden, where the assignment of patients to public or private ambulances can be treated as effectively random.	Partial privatisation causes increased effort in public ambulances.
Schneider et al. (2023), Nature	Financial incentives for vaccination do not have negative unintended future consequence	Analysis of a large-scale, pre-registered study; one-time financial incentive of 200 SEK for Covid-19 vaccination; no effect on 2nd dose uptake.	Providing a *one-time* monetary incentive for Covid-19 vaccination did not decrease later uptake of follow-up vaccination in *Swedish population sample.*

Table 2: Overview of discussed studies.

policy relevance.

While we do not want to suggest that economists should not make policy recommendations based on the available knowledge (even if not reliably causal), we do want to suggest being more careful in view of claims about causality. According to the Oxford Advanced Learners Dictionary, 'causality' is "the relationship between something that happens and the reason for it happening.".[15] Thus, if scientists refer to causality, it seems reasonable to

[15]https://www.oxfordlearnersdictionaries.com/definition/english/causality?q=causality (2.5.2024)

assume the uninitiated reader (policymakers, the general public,...) to understand that as saying that reasons for certain happenings have been identified. As reasons are typically sought after to control and change happenings, it seems important for a responsible science to be clear about the reliability of references to causality and causal processes. Otherwise, there would be a risk of misinterpretations with consequences beyond that of single academic studies. As argued in Section 2 and exemplified in the preceding subsections, reliable causal inferences require both (some form of) internal and external support.

3.6 Poverty, Depression and Anxiety

Finally, we briefly consider a review paper by Ridley et al. (2020a,b) titled "Poverty, depression, and anxiety: Causal evidence and mechanisms". The paper is added here as it also considers the potential (bidirectional) causal relation between poverty and mental health. Yet, different from the previous studies, it, being a review, does not present its own data but establishes claims based on the broader basis of various studies.

Regarding the causal claims made, the caption of the illustration provided in the review summary (Ridley et al., 2020a), for example, reads "The causal relationship between poverty and common mental illnesses. This schematic shows the principal mechanisms we identify, *on the basis of theory and empirical evidence*, through which poverty and depressive and anxiety disorders interact." (own emphasis). And the review itself then presents the various effects found in different studies (see, for example, Ridley, 2020b, Figure 4). Thus, while acknowledging that "The most compelling evidence that poverty causes mental illness comes from RCTs that evaluate antipoverty programs." (Ridley et al., 2020b, p. 2/3), the authors do not infer causality form any single one of the studies but take these only as evidence supporting a causal claim derived from a variety of different studies.

In view of the quality of single studies, the previous citation about RCTs indicates that the authors are well aware that there are different degrees of internal credibility and that different studies contribute differently to the reliability of a causal claim. Yet again, it is the combination of various arguments – some more compelling, some less compelling – in favour of a (bidirectional) causal relation expressed in a variety of studies conducted in different contexts by different people with a different focus that is used to make a specific claim, i.e. it is the combination of (study) internal and external arguments that is used to establish causality.

4 Concluding Remarks

Acquiring knowledge about causal relationships is central to any empirically oriented social science, not least in view of potential demands for policy advice. Yet, while the credibility revolution in economics has done much to inform the science about how to improve empirical

studies and their analyses in this regard, a general debate about what it is that supports causal claims (let alone a consensus) is mostly lacking. In the present paper, we have taken a pragmatic approach and proposed to consider two categories of criteria to guide judgments about causality – referred to as internal and external credibility - which are based on the Bradford Hill Criteria from the medical sciences (Hill, 1965). Discussing five prominent examples, we have argued how especially a stronger explicit awareness of the external aspects of credibility could help to avoid overly strong claims regarding causality based on single studies.

Personally, we are convinced that a social science like economics, which has a high policy relevance, should be careful regarding expressed convictions about causal factors in social processes, not least because policymakers are commonly not the ones further differentiating and scrutinising answers given by science. Thus, we believe that a clearer picture of what is required for causal claims to be reliable could, in fact, help to improve not only the credibility of the causal claims themselves but also the credibility of economics. The criteria discussed in the paper are meant to help improve the clarity of this picture, a clarity that may also help to avoid occasional causal (mis)interpretations of results when only correlational inferences are mentioned.

We want to conclude by reemphasising the possible cost connected with primarily high demands on internal aspects of a study, i.e. on the quality of specific research designs and analyses. We believe this to be important as internal credibility provides the seemingly more "clear-cut" criteria. Strength, specificity, temporality, gradient and experiment all appear more or less "objectively" measurable, while the external criteria (consistency, plausibility, coherence, analogy) appear clearly more subjective. And which scientist would not prefer the objective criterion over the subjective one? Who would want to risk the blame for making a contentious subjective judgment that might turn out to be wrong? Yet, despite preferring objective criteria ourselves, we believe that the arguments in favour of adding more subjective criteria when it comes to establishing reliable causal claims are reasonable and many.

To exemplify what might happen if too much relative weight is put on internal criteria and high demands on formal standards, let us consider possible consequences. For one thing, higher standards in terms of, for example, sample size or number of treatment variations in experiments naturally are associated with higher costs. Accordingly, they favour institutions with more resources. Given the existing focus on a few top journals in economics and the general lack of replication studies, this could arguably lead to a further reduction of views and a less pluralistic science. In fact, not only the plurality of views could be at risk. External aspects, as discussed above, require consistency of findings with already known evidence – if only from analogous studies. It is the availability and amount of such evidence that makes reviews so valuable and relevant for establishing causal claims (e.g. Camerer, 2003; Ridley et al., 2020). If such plurality in results is missing, single "results" are likely

to become more mainstream. Yet, there is always a chance for misinterpretation or simple statistical error. Ample analogous evidence would serve as insurance, especially if obtained under varying conditions – even if not always confirming to the highest standards regarding internal criteria.

A further aspect in connection with higher standards concerns the strong demand for preregistration. As indicated in the discussion of Section 3.1, this can actually be seen as a way to add some external support through the backdoor as hypotheses and analysis plans are commonly derived from what is already known. In that sense, preregistration is not only about the internal strength of a study. Yet, this also comes at a cost, namely a loss in exploratory studies. If not much is known, for example about the reactions to a type of treatment by different genders or certain age groups, what should be preregistered? Of course, without it, there is a risk of data mining. Yet, in a science where also aspects which are external to a specific study are constantly considered, such observations should be eliminated easily and within a short time. To us, the risk that some high-standard study produces chance results or overlooks something or sets a new unwarranted standard seems larger, especially so once background knowledge "worthy" to judge against becomes more scarce.

To sum up, the purpose of this paper was two-fold. On the one hand, it can be seen as a call for more discussion about what we as economists will understand by causality purely with an eye on practical questions, albeit philosophically motivated. On the other hand, it suggests considering the potential cost of becoming strict regarding criteria related to internal aspects of a single study. In neither case did we intend to insist on specific solutions. Rather, our aim was to highlight the questions and suggest potential directions for answers. As for causality, we believe that no single contribution can settle the matter.

References

Albanese, G., Barone, G., & de Blasio, G., 2022. Populist Voting and Losers' Discontent: Does Redistribution Matter?. *European Economic Review* 141, 104000.

Angrist, J., & 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, 3-30.

Autor, D., Dorn, D., Hanson, G., & Majlesi, K., 2020. Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure. *American Economic Review* 110, 3139–3183.

Boesche, T., 2022. Reassessing Quasi-Experiments: Policy Evaluation, Induction, and SUTVA. *British Journal for the Philosophy of Science* 73, 1-22.

Bowles, S., 2008. Policies Designed for Self-Interested Citizens May Undermine" the Moral Sentiments": Evidence from Economic Experiments. *Science*, 320(5883), 1605-1609.

Bryan, C., Tipton, E., & Yeager, D., 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour* 5, 980–989.

Camerer, C., 2003. *Behavioral Game Theory.* Princeton University Press, Princeton, New Jersey.

Campos-Mercade, P., Meier, A., Schneider, F., Meier, S., Pope, D., & Wengström, E., 2021. Monetary Incentives Increase COVID-19 Vaccinations. *Science*, 374(6569), 879-882.

Cartwright, N., 1989. *Nature's Capacities and their Measurement.* Clarendon Press, Oxford.

Cartwright, N., 2007a. Are RCTs the Gold Standard?. *BioSocieties* 2, 11-20.

Cartwright, N., 2007b. *Hunting Causes and Using Them: Approaches in Philosophy and Economics.* Cambridge University Press, Cambridge.

Cartwright, N., 2019. *Nature, the Artful Modeler.* Open Court, Chicago, Illinois.

Committee for the Prize in Economic Sciences in Memory of Alfred Nobel, 2021. Answering Causal Questions Using Observational Data. https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf (3.4.2024)

Dague, L., & Lahey, J., 2019. Causal Inference Methods: Lessons form Applied Microeconomics. *Journal of Public Administration Research And Theory* 29, 511-529.

Deaton, A., 2010. Instruments, Randomization, and Learning About Development. *Journal of economic literature* 48, 424-455.

Deaton, A., & Cartwright, N., 2018. Understanding and Misunderstanding Randomiyed Control Trials. *Social Science & Medicine* 210, 2-21.

Delaney, J., & Devereux, P., 2021. The Economics of Gender and Educational Achievement: Stylized Facts and Causal Evidence. *Oxford Research Encyclopedia of Economics and Finance.* Accessed 18 Apr. 2024. https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-663.

Falkinger, J., Fehr, E., Gächter, S., & Winter-Ebmer, R., 2000. A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence. *American Economic Review*, 91(1), 247-264.

Festré, A., & Garrouste, P., 2015. Theory and Evidence in Psychology and Economics about Motivation Crowding Out: A Possible Convergence?. *Journal of Economic Surveys* 29(2), 339-356.

Green, A., Janmaat, G.., & Cheng, H., 2011. Social Cohesion: Covering and Diverging Trends. *National Institute Economic Review* 215, R6-R22.

Goodman, N., 1983. *Fact, Fiction, and Forecast.* Harvard University Press, Cambridge.

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

Guala, F., & Mittone, L., 2015. A Political Justification of Nudging. *Review of Philosophy and Psychology* 6, 385-395.

Hausman, D. M., & Welch, B., 2010. Debate: To Nudge or Not To Nudge. *Journal of Political Philosophy* 18, 123-136.

Heckman, J., & Urzua, S., 2010. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics*, 156, 27-37.

Hill, A.B., 1965. The Environment and Disease: Association or Causation?. *Proceedings of the Royal Society of Medicine* 58, 295-300.

Holland, P., 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 945-960.

Howick, J., Glasziou, P., & Aronson, J.K., 2009. The Evolution of Evidence Hierarchies: What Can Bradford Hill's 'Guidelines for Causation' Contribute?. *Journal of the Royal Society of Medicine* 102(5), 186-194.

Imbens, G.W., 2010. Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic literature* 48, 399-423.

Infante, G., Lecouteux, G., & Sugden, R., 2016. Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics. *Journal of Economic Methodology* 23, 1-25.

Kemper, F., & Wichardt, P., 2024. Welfare Justifications and Responsibility in Political Decision Making - The Case of Nudging. *Critical Policy Studies*, forthcoming.

Knutsson, D., & Tyrefors, B., 2022. "The Quality and Efficiency of Public and Private Firms: Evidence from Ambulance Services". *The Quarterly Journal of Economics* 137(4), 2213–2262.

Kuhn, T., 1997. *The structure of scientific revolutions.* University of Chicago Press, Chicago.

Larsen, C., 2013. *The Rise and Fall of Social Cohesion: The Construction of Social Trust in the US, UK, Sweden and Denmark.* Oxford University Press, UK.

Leeson, P., 2020. Economics is not statistics (and vice versa). *Journal of Institutional Economics* 16, 423-425.

Lindqvist, E., Östling, R., & Cesarini, D., 2020. Long-Run Effects of Lottery Wealth on Psychological Well-Being. *Review of Economic Studies* 87, 2703-2726.

Maziarz, M., 2020. *The Philosophy of Causality in Economics: Causal Inferences and Policy Proposals.* Routledge.

Menkveld, A., et al., 2024. Nonstandard Errors. *Journal of Finance*, forthcoming.

Neurath, O., 1921. Anti-Spengler. Reprinted in Haller, R., Rutte, H., 1981: *Gesammelte Philosophische Und Methodologische Schriften, Otto Neurath.* Hölder-Pichler-Tempsky, Wien.

Norris, P., & Inglehart, R., 2019. *Cultural backlash: Trump, Brexit, and authoritarian populism.* Cambridge University Press.

Oreskes, N., 2019. *Why Trust Science?.* Princeton University Press, New Jersey.

Ridley, M., Rao, G., Schilbach, F., & Patel, V., 2020a. Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms. *Science* 370(6522), 1289 (Review Summary).

Ridley, M., Rao, G., Schilbach, F., & Patel, V., 2020b. Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms. *Science* 370(6522), eaay0214.

Rohrer, J., 2018. Thinking Clearly About Correlations and Causation: Graphical Causal models for Observational Data. *Advances in Methods and Practices in Psychological Science* 1, 27-42.

Ross, L., & Woodward, J., 2023. Causal Approaches to Scientific Explanation. *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*, Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/.

Rubin, D., 1980. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* 75, 591–593.

Rubin, D., 1986. Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association* 81, 961-962.

Sandel, M., 2012. *What Money Can't Buy: The Moral Limits of Markets.* Macmillan.

Schneider, F., Campos-Mercade, P., Meier, S., Pope, D., Wengström, E., & Meier, A., 2023. Financial Incentives for Vaccination Do Not Have Negative Unintended Consequences. *Nature*, 613(7944), 526-533.

Schünemann, H., Hill, S., Guyatt, G., Akl, E., & Ahmend, F., 2011. The GRADE Approach and Bardford Hill's Criteria for Causation. *Journal of Epidemiology & Community Health* 65, 392-395.

Siddiqi, S., Kording, K., Parvizi, J., & Fox, M., 2022. Causal Mapping of Human Brain Function. *Nature Reviews Neuroscience* 23, 361-375.

Söderlund, B., 2023. The Importance of Business Travel for Trade: Evidence from the Liberalization of the Soviet Airspace. *Journal of International Economics* 145, 103812.

Thaler, R., & Sunstein, C., 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Penguin, London.

The Economist, October 12, 2021. The Nobel Prize in Economics Celebrates an Empirical Revolution. https://www.economist.com/finance-and-economics/2021/10/12/the-nobel-prize-in-economics-celebrates-an-empirical-revolution (3.4.2024)

Vazire, S., Schiavone, S., & Bottesini, J, 2022. Credibility beyond replicability: Improving the Four Validities in Psychological science. *Current Directions in Psychological Science* 31, 162-168.

Wichardt, P., 2014. Models and Fictions in (Micro-) Economics. Available at SSRN 2487909.

Woodward, J., 2005. *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press, New York.