Journal of Economic Behavior & Organization 93 (2013) 186–195

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Claims and confounds in economic experiments☆

Daniel John Zizzo*,[1]

School of Economics and CBESS, Norwich Research Park, University of East Anglia, Norwich NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 18 August 2011
Received in revised form 8 May 2013
Accepted 15 May 2013
Available online 24 May 2013

JEL classification:
B41
C90

Keywords:
Confounds
Claims
Experimental design
Methodology
Internal validity
External validity

ABSTRACT

We present a distinctiveness, relevance and plausibility (DRP) method for systematically evaluating potential experimental confounds. A claim is a statement being inferred on the basis of experimental data analysis. A potential confound is a statement providing a prima facie reason why the claim is not justified (other than internal weakness). In evaluating whether a potential confound is problematic, we can start by asking whether the potential confound is *distinctive* from the claim; we can then ask whether it is *relevant* for the claim; and we can conclude by asking whether it is *plausible* in the light of the evidence.

1. Introduction

Assume that an experimental paper makes a claim A. What happens if there is a potential experimental confound B that may affect behavior in the experiment and hence the validity of the claim? Examples are confusion effects (Ferraro and Vossler, 2010), experimenter demand effects (Zizzo, 2010), framing effects (Cookson, 2000), house money effects (Harrison, 2007), demographic effects (Casari et al., 2007), wealth effects (Armantier, 2006), incentives size effects (Slonim and Roth, 1998), task order effects (Hogarth and Einhorn, 1992), sample selection effects (Harrison et al., 2009), risk aversion effects (Vieider, 2011), behavioral noise effects (Hey, 2005), lack of credibility of experimental instructions due to the use of deception (Hertwig and Ortmann, 2001), or lack of control for social preferences explanations (Gächter et al., 2012) or for expectations about the coplayer (Ashraf et al., 2006).

The aim of this paper is to provide a method to identify when potential confounds are a problem for the claims made in an experimental paper, and, if they are a problem, what are valid ways of addressing it. Experimental economics textbooks and, more generally, research methods textbooks have a general discussion of standard responses to the problems of experimental confounds (e.g., Davis and Holt, 1993; Friedman and Sunder, 1994; Jackson, 2008). While there is plenty of econometric

* Tel.: +44 1603 593668; fax: +44 1603 456259.
E-mail address: d.zizzo@uea.ac.uk

[1] I wish to thank Franz Dietrich, two reviewers and participants to presentations in Erfurt, Heidelberg, Jena, Munich and San Sebastian for their feedback. The usual disclaimer applies.

0167-2681/$ – see front matter © 2013 The Author. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jebo.2013.05.006

guidance in the literature,[2] what is missing is a general methodological analysis of how to handle experimental confounds that goes beyond what can be found in the introductory textbooks, and that may be useful especially for junior researchers designing or evaluating experiments.

We try to ask the question of *when* experimental economists should be worried about potential confounds. We are interested in providing an evaluation methodology of interest especially to junior researchers designing and running experiments and, when an experiment has been run and a paper completed, of interest to editors, referees and indeed the wider readership in evaluating an experimental paper, or to the authors of such paper in determining what to do next.

Our method centers around the notions of *distinctiveness, relevance and plausibility* (DRP) of a potential confound. We discuss the use of these concepts in the light of a number of practical examples and show how our DRP method can be employed to systematically evaluate potential confounds criticisms. We also review, in the light of the DRP method, other arguments that have been made or can be made to defend experimental designs against potential confound criticisms.

Section 2 provides the general conceptual framework and background, Sections 3, 4 and 5 respectively introduce the notions of distinctiveness, relevance and plausibility of potential confounds. Section 6 considers the implications of the DRP method as a whole, summarizes the method with a table and discusses what the method can and cannot be used for. Section 7 employs the DRP method to evaluate other arguments on potential confounds. Section 8 concludes.

2. The conceptual framework

Define a *claim* as a statement being inferred in a paper P on the basis of experimental data analysis E. We consider one such claim, which we label as A, and we define a claim D which is different from A.

One reason A may not be justified is that it may not be supported by E *as stated in the paper*, which we can label as E|P. If (E|P ⇒ A) is not justified, that is if the experimental data analysis as stated in the paper does not imply A, then the experimental data analysis does not support the claim A, and so the claim A is in itself not justified.[3] We can label this situation as one of *internal weakness*.[4]

Define a *potential confound* as a statement providing a prima facie reason why A is not justified other than internal weakness.[5] Let us consider one such potential confound, which we label as B; examples have been provided in the introduction. That is, it is possible that E|P ⇒ A, but nevertheless B ⇒ D.

We now need to identify a parsimonious and yet comprehensive procedure by which we can test for whether a potential confound should be considered a problem. The first question we can ask is whether A and B are essentially the same. That is to say, the statement (B ⇒ D) is criticized on the ground of the relationship between A and B. If B is equivalent to A, then B is not a problem for A: that is, B ⇔ A. If B is truly different (distinctive) from A, then based on this test we cannot rule out that (B ⇒ D). This is a *distinctiveness* test of the potential confound.

Assume that B is distinctive from A. The second question we can ask is whether B negates A in principle. That is to say, the statement (B ⇒ D) is criticized on the ground of the relationship between A and D. If the confound has passed the distinctiveness test, we know that D is different from A, but this does not mean that D necessarily contradicts A. The potential confound B may imply D, but may not be relevant for whether A holds. If B is irrelevant, D can be true but at the same time A can also be true. This is a *relevance* test of the potential confound.

Assume that B is both distinctive from A and relevant for A. The third and final question we can ask is whether, based on all available evidence, it is plausible to believe that (B ⇒ D) is true in practice. This is no longer a logical test in the sense of distinctiveness and relevance, but rather a consideration of whether the potential confound is plausible from an empirical viewpoint. This is a *plausibility* test of the potential confound.

The next three sections provide examples of the use of the distinctiveness, relevance and plausibility tests. We have simplified our analysis to the existence of a single potential confound B, but our analysis easily extends to considering multiple potential confounds.[6] If a potential confound is distinctive, relevant and plausible, then it is a problem that needs to be addressed and Section 6 provides a discussion.

We believe the procedure to be comprehensive in the sense that evaluating B means to evaluate the relationship (B ⇒ D), and this can either mean to question each part of this relationship with A (i.e. B with A, B with D), which is what the distinctiveness and relevance tests do, or to question the validity of the whole of the relationship, which is what the plausibility test does. Obviously different evaluation procedures might be possible and this paper does not preclude research on further

[2] Experimental designs can be run to ameliorate the problem; random samples can be used; various sources of confounding can be controlled for using covariates or using suitable econometric tools such as instrumental variables. See Davidson and MacKinnon (1993) for a standard econometric analysis and Greenland et al. (1999) for an example of analysis from a statistician's perspective on confounding.

[3] An example is if a claim is made that a specific frame induces more cooperation in a public good contribution game when in fact, in looking at the regression analysis on contribution the data analysis section of the paper is centered upon, the sign is statistically significant but negative, i.e. in the opposite direction of the claim.

[4] We label it as internal since the lack of justification is based on the data analysis in paper P which is supposed to underpin the claim A made also in paper P.

[5] For examples of definitions of confounds, see Mill (1843/2009), Patten (2007) and Jackson (2008).

[6] In Section 7 we shall consider an argument that will require us to consider two potential confounds at the same time.

classifications; one advantage of the one adopted here is that, in classifying all issues under three possible dimensions, it is suitably parsimonious. In Section 7 we consider additional potential dimensions, and show that they are not needed.

We note, in passing, that the nature of the claim A will depend on the type of experiment. Davis and Holt (1993) classify experiments depending on whether they test behavioral hypotheses from theories, whether they are theory stress tests (i.e., to check for robustness) and whether they aim to identify empirical regularities. Experiments providing advice to policy makers may also exist (Friedman and Sunder, 1994; Schram, 2005). There are, in addition, different views of how experiments relate to theoretical models and the real world (e.g., Jones, 2008; Guala, 2005; Schram, 2005; Sugden, 2005), and this in turn may affect what claims are made.

3. Distinctiveness

The potential confound B needs to be distinctive from the claim A in order for it to be a potential problem. A typical case where this is not met is where the potential confound *is* the actual object of investigation.

Example 1. Requate and Waichman (2011) verify if Cournot market experiments outcomes differ depending on whether the experimental instructions contain (a) a profit table, (b) a profit calculator without a best-response option or (c) a profit calculator with a best-response option. They make the claim A that, while the first two options bring about the same outcomes, option (c) "tends to increase aggregate output to the Cournot level and decrease the incidence of tacit collusion" (p. 36, abstract). The potential confound B here is that the method of provision of profit information affects the outcomes of Cournot market experiments. Obviously A and B are essentially equivalent in this case. As identifying the effect of different methods of provision of profit information is precisely what the claim A is about, the distinctiveness test is failed and there is no problem for A.

4. Relevance

As noted earlier, the relevance test is about whether the potential confound B is relevant for the truthfulness of claim A. This depends on whether the statement that B implies D ($B \Rightarrow D$) has any bearing on whether A is true, which in turns depends on the relationship between A and D.

Orthogonality. Assume that A and D are entirely unrelated, in the sense that D is not a claim of an action or outcome that goes systematically either in the direction of or in the opposite direction of the experimental actions or outcomes implied by the experimental claim. In this case we can say that B is *orthogonal* to A. If so, then B is not relevant since ($B \Rightarrow D$) does not affect whether A is justified.

Example 2. Huck et al. (2004) found that, in a Cournot market experiment setting, a market frame led to more competitive outcomes than if a neutral frame was employed.[7] We may ask whether this potential confound is relevant for the claims made in Huck et al. (1999) regarding the performance of different learning theories in Cournot market experiments. While Huck et al. (1999) only used a market frame, it is not obvious that the significance of the market frame has any bearing for claims about the performance of learning theories relative to another. In this sense the potential confound is orthogonal to the claims made in Huck et al. (1999).

Contrary relevance. Assume that D is a claim of an action or outcome that goes in the opposite direction of the action or outcome stated with claim A. This means that, as $B \Rightarrow D$, B works in the opposite direction of the prediction stated in claim A. If so and if there is still evidence for A notwithstanding the fact that B implies D, B is irrelevant. The only case in which B is relevant is if claim A acknowledges limited evidential support for the action or outcome. We do not then know whether such limited evidence is due to ($B \Rightarrow D$) or not. In this case we can speak of contrary relevance of B.

Example 3. Millner et al. (1990) ran a continuous-time market experiment in which they tested market contestability theory, according to which the threat of hit-and-run entry is enough to keep market outcomes competitive even with a single market incumbent. Their claim is that the evidence is against contestability theory.[8] A potential confound B here is boredom. The claim D can be made that, due to boredom, subjects are less likely to stay out of the market and do nothing. This would lead subjects to engage in more hit-and-run entry and get outcomes more aligned with contestability theory. However, as D implies behavior more aligned with contestability theory and A is a claim about having found evidential support for the opposite conclusion, the potential confound is not a problem.[9] If anything, it is the more striking to find evidence against contestability theory precisely because there should be a bias toward contestability theory if subjects are bored from doing nothing.

Same direction relevance. Assume that D is a claim of an action or outcome that goes in the same direction as the action or outcome stated with claim A. For example, both A and D are claims about the same action being taken by subjects. However, A identifies a reason why the action or outcome takes place, and, assuming that B has passed the distinctiveness test, this

[7] The market frame used terms such as 'firms', 'markets' and 'price'; the neutral frame did not.

[8] As stated in the abstract (p. 584), "the average efficiency of markets contestable by two firms... was significantly less than that associated with the sustainable equilibrium", where the 'sustainable equilibrium' is as defined by contestability theory.

[9] A reviewer wondered whether B can be considered a confound with respect to the magnitude of the effect. If the claim A *were* about a specific magnitude of the effect, then the answer would be yes, but this is not what Millner et al. (1990) are claiming.

reason is not B. That being so, we cannot say whether A is true or not based on the evidence, as the evidence could both be interpreted as claimed by A or as claimed by B. In this case we can speak of same direction relevance of B.

Example 4. Benjamin et al. (2010) present an experiment on social identity and preferences. They begin their conclusions section with the claim A that "our results suggest that social identity affects fundamental economic preferences" (p. 1925). They started each experimental session with a 'background questionnaire' the goal of which was to make ethnicity or gender salient for the subjects. The potential confound B here is that the background questionnaire reinforces the effect of social identity in a way which would not be observed in the real world. The action D induced by the potential confound B in this case is to increase the behavior conforming to a specific social identity (based on ethnicity or gender). There is then a same direction relevance problem, since it is not possible to differentiate whether the evidence backs a 'fundamental' effect of social identity on economic preferences or, more modestly, the effect of salience which may be especially strong in an experimental environment.

Magnifying glass relevance. Consider a magnifying glass. It is an artificial tool that helps the observer study a real world object by artificially increasing its visual size. Now consider a potential confound B with same direction relevance. B may identify an experimental design feature by which the equivalent of a key real world feature is implemented or facilitated, if artificially. In this case the experimental design feature may be desirable insofar as it strengthens the real world significance of the claim A being made. In this case we can speak of magnifying glass relevance of B.

Example 5. Chaudhuri et al. (2009) looked at the role of intergenerational advice in the context of weakest link coordination games. The areas of application are suggested to be macroeconomic equilibrium traps in developing countries, speculative attacks, bank runs and political revolutions. It is implied that the results of the experiments apply to these contexts; in the conclusion, a claim A is explicitly made in terms of message from the experiment for "policy makers (like central bankers) to be able to coordinate a move from a sub-optimal (underemployment) equilibrium" (p. 118). The potential confound here is that there is an experimenter demand effect B.[10] It may increase the degree to which subjects pay attention to advice given by other subjects. Because of this, B has same direction relevance for A.

Assume, however, that the claim made by Chaudhuri et al. (2009) is rephrased to apply not to policy settings but rather to organizational settings. In other words, we consider the weakest link game as applying to the context of an organization (as, e.g., in Brandts et al., 2010). In this context, intergenerational advice can be reinterpreted as being advice that new generations of employees receive from existing employers. The experimenter demand effect would then artificially mirror in the laboratory the social pressure that would be present in the real world organizational setting. If we rephrase A this way, we can then argue for the magnifying glass relevance of B.[11]

5. Plausibility

Assume that the potential confound B is distinctive and relevant. It does not necessarily follow that A is not justified. As discussed in Section 2, B may be empirically implausible, in which case we would still be in a position to make the claim A.

Direct evidence. B may have been tested in the relevant domain. This would directly answer to the question of whether B is plausible.

Example 1 (*again*). As discussed above, Requate and Waichman (2011) test if Cournot market experiments outcomes depend on whether the experimental instructions contain (a) a profit table, (b) a profit calculator without a best-response option or (c) a profit calculator with a best-response option. As a result, if B is about how payoffs are provided in Cournot market experiments, it would be possible to use their evidence in favor or against the plausibility of B.

Example 6. A number of experimental papers employ dictator games to make claims about the effect of a number of factors on social preferences (e.g., Andreoni and Miller, 2002, or Haisley and Weber, 2010) or norms (Guala and Mittone, 2010).[12] The problem here is that the dictator game is a highly artificial task where, for no apparent reason, subjects are asked to consider giving significant amounts of money to strangers (e.g., Bardsley, 2005; Smith, 2010). The potential confound B here is (a) that there is an experimental demand to give and (b) that the cues offered by the experimental environment help subjects make sense of how much they should give.

List (2007), Bardsley (2005) and Zizzo and Fleming (2011) discuss direct evidence showing how experimenter demand shapes dictator game behavior. List's (2007) and Bardsley's (2005) evidence is based on a simple change of the range of possible actions to include taking rather than just giving: this can reverse the agents' apparent generosity. Zizzo and Fleming (2011) find that behavior in a dictator game is connected to a standard questionnaire measure of sensitivity to social pressure (Stöber, 2001). They also find that, when a dictator game and a symmetrical back to back destruction game in which the first mover can simply destroy money of the second mover are played, there is a *positive* relationship between giving and destroying, which is what B predicts.

[10] As defined in Zizzo (2010), experimenter demand effects refer to changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior in the experiment.

[11] Note that, since in the real world there is not the equivalent of constant experimenter-induced priming of social identity, the magnifying glass relevance argument cannot be used in relation to Example 4 and Benjamin et al.'s (2010) claim of a fundamental effect of social identity on preferences.

[12] For example, the claim A is made in Andreoni and Miller (2002, p. 737) that "subjects exhibit a significant degree of rationally altruistic behavior". Guala and Mittone (2010) reinterpret the dictator game as being useful to investigate social norms rather than preferences. However, it is not clear to what extent the so-measured social norms are an artifact or an inflated by-product of experimenter demand.

Table 1
A distinctiveness, relevance and plausibility (DRP) method.

Test	Key points	Does B pass the test?
Distinctiveness	A is equivalent to B	
	A is distinctive from B	Yes
Relevance	Orthogonality of A and B	
	B works against A *but* supportive evidence for A	
	Contrary relevance	Yes
	Same direction relevance: no magnifying glass relevance	Yes
	Magnifying glass relevance	
Plausibility	Direct evidence supportive of B	Yes
	Indirect evidence supportive of B	Yes
	Global plausibility in favor of B	Yes

Options if problem in terms of distinctiveness *and* relevance *and* plausibility	
Actions	Follow up
Drop A	
Run additional statistical analysis if appropriate data available	Iterate process
Do additional appropriate experimental treatments if feasible	Iterate process
Modify A	Iterate process

This table provides a check-list for the use of the DRP method. The method considers the relationship between a claim A and a potential confound B in the light of the three conditions of distinctiveness, relevance and plausibility. The top part of the table considers these conditions. The bottom part reviews what to do if a problem in terms of distinctiveness, relevance and plausibility emerges.

Indirect evidence. There may be behavioral patterns that A can explain but B cannot explain. This is indirect, if not necessarily conclusive, evidence against the plausibility of B.

Example 7. Nikiforakis (2008) makes the claim A that, in a social dilemma setting, "in the presence of counter-punishment opportunities cooperators are less willing to punish free riders" (p. 91, from the abstract). A potential confound B here is that, as a rule of thumb, some subjects may simply mechanically anchor the counter-punishment points to the punishment points values observed. The implied action D is that, because of this experimental rule of thumb, cooperators may then be less willing to punish free-riders. Indirect evidence against B is however presented in Nikiforakis (2008); for example, there is more counter-punishment when subjects are repeatedly re-matched with the same coplayers than when they are not. This pattern of behavior cannot be explained by B.

Global plausibility test. In many cases there will not be directly or indirectly applicable evidence. However, it should still be possible, based on the available evidence at a given point in time, to reach a judgment about the plausibility of B relative to claim A.

Example 6 (*again*). A feature of dictator game experiments is the extraordinary degree of sensitivity to the smallest changes in the design relative to other games, such as in deservingness (Hoffman et al., 1994), the opportunity to play an unattractive lottery (Oberholzer-Gee and Eichenberger, 2008) or visual face like stimuli (Rigdon et al., 2009). Depending on the experimental details, the fraction of givers varies widely, between the around 5 percent of the subjects of Cherry et al. (2002) and the over 95 percent of Branas-Garza (2006). Now assume that a claim is made about the wide significance of a framing effect based on dictator game results. This claim does not pass a global plausibility test. The reason is that the hyper-sensitivity to small changes in the experimental design is not observed in other settings.

6. A DRP method

Distinctiveness, relevance and plausibility – or DRP for short – together provide a method to analyze whether a potential confound B is a problem for a claim A. Assume that B is distinctive from A, relevant for A, and plausible. Also assume that a realistic case for magnifying glass relevance cannot be made. If these conditions apply, the claim A can be judged to be problematic.

There are then two options. If the researcher still wants to make the claim A and if it is feasible, he or she can re-design the experiment or, if the experiment has already been run, do additional treatments or at least additional statistical analysis; B can then be tested again for DRP. Alternatively, the researcher may choose to drop or appropriately modify A. If A is appropriately modified, the evaluation process on whether B is a problem of A can be iterated in terms again of DRP of B. Table 1 provides the key points of a DRP method as discussed in this paper.

It is important to understand what the DRP method tries (and tries not) to achieve. It provides a method for researchers designing and running experiments. When an experiment has been run and a paper completed, it can be of interest to editors, referees and indeed the wider readership in evaluating an experimental paper, or to the authors of such a paper in determining what to do next in the light of a potential confound criticism. As we shall see in the next section, this conceptual framework also enables us to evaluate other arguments that are sometimes used, explicitly and implicitly, as defenses against criticisms of experimental confounds. What the DRP analysis does not provide (nor is it meant to) is a theory of

scientific discovery.[13] The judgments employed in a DRP method of a potential confound are judgments operating *at a given point in time*. They are not meant to preclude research which may change, even radically, those judgments in future DRP analyses. For example, B may appear irrelevant to A but, in the light of new research, this may change; or what may be seen as plausible today may, in the light of new research, be shown to be implausible, or vice versa. That better science can be made in the future when greater knowledge is available is of course just to be expected in a healthy progressive scientific research paradigm, and does not preclude the usefulness of making DRP evaluations in the here and now based on what is currently known.

Another potential problem with a DRP analysis is that the number of potential confounds can in principle be very large if not logically infinite (see Smith, 1994). While this is true in principle, in practice the number of *plausible* (and relevant and distinctive) confounds will typically be a finite list (see Guala, 2002, for a similar argument). Also, if appropriate qualifications are made to A, it is conceivable that a paper might be publishable even if B is a potential problem.

Given the word constraints of experimental papers, it is obviously not practical for all potential confounds to be explicitly dealt with in the discussion section; many potential confounds will be evidently implausible or irrelevant. Nevertheless, an appropriate discussion of relevant potential confounds is appropriate at least for full length experimental papers; this occurs in a number of cases but not (adequately) in others.

A DRP analysis can also be worthwhile to evaluate papers, or possibly re-evaluate them in a different light than those from the authors.

Example 5 (*again*). Example 5 above is a good illustration of this point. In this example we reviewed Chaudhuri et al.'s (2009) experiment on intergenerational advice in the context of weakest link coordination games. We noted how, insofar as claim A (about intergenerational advice) is supposed to apply to policy makers like central bankers, it is problematic. This is because advice in the lab may be especially effective due to experimenter demand, and as a result there is a same direction relevance problem. Instead, if the claim A is modified to apply to an organizational context, the experimenter demand effect would artificially mirror in the laboratory the social pressure that would be present in the real world organizational setting (a case of magnifying glass relevance). A DRP analysis can therefore be used to put Chaudhuri et al.'s (2009) results in a different light.

Example 6 (*again*). A DRP analysis of experimenter demand effects as a potential confound B can be used to identify settings under which this potential confound is not a problem.

B does not pass the distinctiveness test if experimenter demand is what is being investigated using dictator games (as in Zizzo and Fleming, 2011).

B does not pass the relevance test if one may claim magnifying glass relevance. Assume that A relates explicitly to charitable giving, and that the dictator game manipulation is of the kind that a charity could implement to increase funding. One may also assume that the experimenter demand is equivalent to the real world social pressure that the charity would be able to implement, for example by the means of requesting donations through phone calls (Shang and Croson, 2009).[14]

B does not pass the plausibility test if a dictator game is used purely to refer to a sub-game of a more complex game with a natural interpretation. Consider a standard Berg et al. (1995) sequential trust game. Once the trustee receives the investment by the truster, the sub-game is equivalent to a dictator game with the trustee as the dictator. That said, there is a natural interpretation for the whole interaction, which is that of a trust game.

7. Evaluating other arguments

We are now in a position to employ the DRP framework to evaluate four other arguments that can be made to defend experimental claims. We label these arguments the norm defense, the resources defense, the confounds trade-off defense and the claims trade-off defense.[15]

7.1. The norm defense

The norm defense states that a particular experimental design choice can be defended on the grounds that this is 'what everyone else does' (or at least what a previous paper has done). A reference to a scientific norm, as established in a previous paper or papers, is employed as a defense. In a normal scientific paradigm, this is healthy to some extent: replications and extensions of previous experimental designs enable better interpretability and are part of a collective research process in which, in the words of Smith (2010, p. 4), "we want to reduce error and to understand its sources". That being said, 'this is what everyone else does' is fundamentally a non-answer. It does not answer the question of why B is not a problem in a specific context, and, by careful selection of sources, it may omit key evidence showing that B *is* indeed a problem. No doubt, if B does not pass the relevance test, comparability with the previous literature suggests to do 'what everyone else does'.

[13] For one philosophy of science perspective on scientific discovery compatible with our analysis, see Mayo (1996).

[14] Fong and Luttmer (2011) and Reinstein and Riener (2012) provide examples of dictator game experiments receiving some justification in this way.

[15] Another way of looking at these arguments is as extra dimensions (additional to distinctiveness, relevance and plausibility) that may seem to be helpful in evaluating confounds. This section will make clear however that, insofar as they have any cogency, it is because they rely on distinctiveness, relevance or plausibility, and therefore these extra dimensions are not needed.

However, it is not clear on what grounds one would defend 'what everyone else does' if the potential confound is distinctive, relevant and plausible (and does not have magnifying glass relevance).

Example 8. Armantier (2006) employed an ultimatum game with random matching repeated 60 times. There were only two sessions per treatment. Armantier makes claims regarding wealth effects affecting behavior. For example, a claim is that "with time... rich players become more greedy, and this behavior is tolerated by poor subjects" (p. 425).

A potential confound here is boredom. This would have been due the large number of identical rounds in an experiment that could last some 90 min.

Another potential confound here is the lack of independent observations given the combination of random matching and only two sessions per treatment. After the first round, observations within the same sessions share a common history.

A response by Armantier is that "the number of rounds played" is "not unprecedented" and that it "is not uncommon to conduct few sessions" (p. 399). Yet the fact that a claim has a problem because of a potential confound does not mean that other papers do not have the same problem, such as (in this example) boredom or lack of non-independence of observations.[16]

7.2. Resources defense

The resources defense states that, although B is plausible, addressing it can be left for future research.[17] It is true that no single paper can do everything. It is also true that this is an especially serious problem for more innovative experimental designs. However, it is not obvious why, if B is distinctive, relevant and plausible, it should not be taken into account in some way. If a magnifying glass relevance argument cannot be made, or additional statistical analysis or experimental treatments are unfeasible, the appropriate response is to modify the claim A rather than retain it as it is. Put it differently, limitations of resources should not be used to justify a claim where there is a credible potential confound. At most, they can be used to justify how to respond to such a potential confound: for example, by modifying the claim rather than by running additional experimental treatments.

Example 9. Sutter et al. (2007, 2008) ran an experiment on group vs. individual decision making in the context of mobile telephony license auctions of the kind run successfully in the UK. The motivation is that large companies, rather than individuals, put bids for this auction. They made the claim A that their experiment provides "conclusive evidence for several noteworthy differences in the bidding behavior of individuals and small teams", and this is seen to "provide important implications... for real-world auctions" (Sutter et al., 2008, p. 390). In their experiment a large company is modeled as a democratic team jointly deciding a bid. The potential confound B is that in the real world one would typically have a team advising an executive officer in making a bid, rather than a democratic team. This different organizational structure could lead to behavior that may or may not be along the lines of the 'noteworthy differences' and 'important implications' identified by Sutter et al. In a footnote of Sutter et al. (2007, p. 12), they use a resources defense by stating that this potential confound is a 'plausible conjecture that could be tested in the future by running experiments where individual decision-makers receive advice from others'. If it is plausible, it is not however clear why this should not modify the claim being made.[18]

Example 10. Experimental psychologists often argue with experimental economists on whether deception is justified in running experiments (for good overviews, see Hertwig and Ortmann, 2001, 2008). One argument used by some psychologists to defend the use of deception has been that it is necessary to run experiments that would otherwise be unfeasible or require significant additional resources (Cook and Yamagishi, 2008).

I would argue that most experimental economists would intuitively think this argument weak. This can be argued on the grounds that deception is a distinctive, relevant and plausible potential confound. It is seen as plausible because it implies that incentives are no longer salient in the sense of Smith (1982), both in relation to the experiment being run and potentially (due to loss of reputation) in relation to others run at the same laboratory.

7.3. Confounds trade-off defense

A confounds trade-off defense argues that the reason why there is a potential confound is to avoid another potential confound.

Assume that there are two potential confounds, B and B′. There may be a trade-off between them. Choices in experimental design that work toward neutralizing B, may make B′ a more significant problem. The reverse may also be true. Trade-offs of this kind are pervasive in experimental design.

[16] Armantier also argues against the criticism of non-independence of observations based on some collected data. In our framework, this can be recast, as saying that in his view this potential confound does not pass the plausibility test.

[17] Anecdotal evidence suggests that this is a not uncommon response by authors when revise and resubmits are obtained from journals with a request for further experimental treatments. Example 9 below is one such case, as the footnote mentioned there was introduced in response to a reviewer's comment (Kocher, personal communication).

[18] It is interesting to note that Sutter et al. (2008) dropped the footnote, implicitly (and perhaps optimistically) suggesting that the potential confound was not plausible after all.

The pervasiveness of trade-offs of this kind does not however justify a confounds trade-off defense:

(i) the defense *does not identify why the experimenter has chosen a specific point in the trade-off*, e.g. by giving priority to B′ in place of B. Identifying a specific point in the trade-off requires going back to considerations of plausibility and relevance of the two potential confounds. It may also require the experimenter to think whether he or she is more willing to modify the claim A to recognize the potential confound B; or whether he or she is more willing to modify A to recognize the potential confound B′;
(ii) the defense *does not justify why the trade-off should be accepted in the first place*. Assume that the following conditions are met, no matter how the trade-off is solved: both B and B′ are distinctive and relevant; neither of them can be set aside on the grounds of magnifying glass relevance; both of them are plausible. If so, A cannot be rescued on the grounds of a confounds trade-off defense, since any resolution of the trade-off would not suffice to justify A.

Example 11. A well-known problem in preparing experimental instructions is deciding whether to frame the instructions in an abstract or concrete way. There is typically a continuum of possible solutions. At one extreme, for example, one could frame a market price making decision in terms of asking the subject to impersonate a business manager making price choices for a company. At the other extreme, one could avoid talking of companies, prices or quantities and keep the problem in strictly abstract terms (e.g., variables *x*, *y* and *z*).

On the one hand, asking the subject to play an 'as if' role and using concrete language may be more prone to experimenter demand effects. For example Holt (1992) notes the need to avoid loaded language. On the other hand, some concrete context can help subjects' understanding (Holt, 1992; Cooper and Kagel, 2003, 2009). Clearly, in choosing how to frame experimental instructions, there will typically be a trade-off between the facilitation of understanding and the danger of experimenter demand effects. Equally clearly, we would require instructions to provide a sufficient degree of confidence *both* in terms of appropriate level of understanding *and* avoidance of experimenter demand effects. In order to achieve such confidence, we would test the distinctiveness, relevance and plausibility of both potential confounds in the light of the claim A being made.

7.4. Claims trade-off defense

A claims trade-off defense argues that the reason why there is a potential confound is because there are two claims being made in the experiments, which we can label as A and A′. The defense is that it is not possible to address B while trying to achieve both A and A′.

This is a weak argument. It only shows that the experiment is being over-ambitious in what it is trying to achieve. If B is plausible, distinctive and problematically relevant, a better strategy would be to achieve just A or A′ but do so in a way that addresses B. Alternatively, one can revise A and/or A′ to explicitly acknowledge the limitation from the potential confound B.[19]

Example 12. Fehr and Tyran (2001) aim to show indirect money illusion: namely, how "a small amount of money illusion" (p. 1239) at the individual level may cause big changes in aggregate market price setting *behavior* after negative nominal shocks (label this as A). They also make the claim that stated *expectations* of subjects (about pricing decisions of other subjects) are "very sticky" in a way that is related to "the nature of money illusion in our experiment" (p. 1259; label this claim as A′). To prove both A and A′, they collect not only price decisions but also expectations about the mean price chosen by the other subjects and indeed their confidence about these expectations. A potential confound B is that, by asking subjects explicitly to think about expectations, one is collecting evidence for A′ (in terms of stated expectations) but is also enhancing the extent to which subjects take expectations of other subjects in their decision making on prices, thus creating a problem for A.[20] In other words, by asking for expectations one makes it potentially more likely that a small amount of money illusion has the claimed effect. It is not a satisfactory response to say that there is a desire to make both claims even if there is a trade-off between the two. The issue instead is with the distinctiveness, relevance and plausibility of B.

8. Conclusions

We have presented a distinctiveness, relevance and plausibility (DRP) method for evaluating potential confounds. The method is summarized in Table 1 above. It is comprehensive insofar as, if there is a claim (A) and a potential confound (B), we can logically start by asking whether B is actually different from A (distinctiveness test); we can then ask whether B actually contradicts A in principle (relevance test); and we can conclude by asking whether B does contradict A in practice, in the light of the evidence available at a given point of time (plausibility test). While we have considered four potential alternative dimensions in Section 7, the three identified here provide a systematic, parsimonious and comprehensive toolkit to test the significance of potential cofounds.

[19] Still alternative solutions, if feasible to address the problem, would be to run additional experimental treatments or at least additional statistical analysis, and see whether this provides evidence showing that B is implausible.

[20] For example, Rustrom and Wilcox (2009) and Gächter and Renner (2010) show how belief elicitation can distort behavior in a repeated matching pennies game and in a public good contribution setup respectively.

Our analysis aims to fill what is currently a serious gap in methodological thinking about experimental design. The gap is particularly striking given the explosive growth of experimental research in academic journals. It is hoped that using the DRP method can help foster more disciplined experimental designs and more systematic and transparent evaluations of experimental designs and of experimental claims at any given point in time. Obviously further methodological work will be useful to analyze the implications of a DRP method in the context of specific potential confounds, including ones (such as demographic effects) that we have not touched on in this paper. This paper will have achieved its minimal goal if it encourages further pragmatic research on how to design and evaluate experiments.

References

Andreoni, J., Miller, J., 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70, 737–756.
Armantier, O., 2006. Do wealth differences affect fairness considerations? International Economic Review 47, 391–429.
Ashraf, N., Bohnet, I., Piankov, N., 2006. Decomposing trust and trustworthiness. Experimental Economics 9, 193–208.
Bardsley, N., 2005. Experimental economics and the artificiality of alteration. Journal of Economic Methodology 12, 239–251.
Benjamin, D., Choi, J.C., Strickland, J., 2010. Social identity and preferences. American Economic Review 100, 1913–1928.
Berg, J., Dickaut, J.W., McCabe, K.A., 1995. Trust, reciprocity and social history. Games and Economic Behavior 10, 122–142.
Branas-Garza, P., 2006. Poverty in dictator games: awakening solidarity. Journal of Economic Behavior and Organization 60, 306–320.
Brandts, J., Cooper, D.J., Fatas, E., 2010. Stand By Me: Help, Heterogeneity And Commitment In Experimental Coordination Games. University Autonoma de Barcelona, Florida State University and University of Valencia working paper.
Casari, M., Ham, J.C., Kagel, J.H., 2007. Selection bias, demographic effects, and ability effects in common value auction experiments. American Economic Review 97, 1278–1304.
Chaudhuri, A., Schotter, A., Sopher, B., 2009. Talking ourselves to efficiency: coordination in inter-generational minimum effort games with private, almost common and common knowledge of advice. Economic Journal 119, 91–122.
Cherry, T.L., Frykblom, P., Shogren, J.F., 2002. Hardnose the dictator. American Economic Review 92, 1218–1221.
Cook, K.S., Yamagishi, T., 2008. A defense of deception on scientific grounds. Social Psychology Quarterly 71, 215–221.
Cookson, R., 2000. Framing effects in public good experiments. Experimental Economics 3, 55–79.
Cooper, D.J., Kagel, J.H., 2003. The impact of meaningful context on strategic play in signaling games. Journal of Economic Behavior and Organization 50, 311–337.
Cooper, D.J., Kagel, J.H., 2009. The role of context and team play in cross-game learning. Journal of the European Economic Association 7, 1101–1139.
Davidson, R., MacKinnon, J., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York and Oxford.
Davis, D.D., Holt, C.H., 1993. Experimental Economics. Princeton University Press, Princeton.
Fehr, E., Tyran, J.-R., 2001. Does money illusion matter? American Economic Review 91, 1241–1262.
Ferraro, P.J., Vossler, C.A., 2010. The source and significance of confusion in public goods experiments. BE Journal of Economic Analysis and Policy (Contributions) 10, Article 53.
Fong, C.M., Luttmer, E.F.P., 2011. Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment. Journal of Public Economics 95, 372–394.
Friedman, D., Sunder, S., 1994. Experimental Methods: A Primer for Economists. Cambridge University Press, Cambridge.
Gächter, S., Nosenzo, D., Sefton, M., 2012. Peer effects in pro-social behavior: social norms or social preferences? Journal of the European Economic Association (in press).
Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public good experiments. Experimental Economics 13, 364–377.
Greenland, S., Robins, J.M., Pearl, J., 1999. Confounding and collapsibility in causal inference. Statistical Science 14, 29–46.
Guala, F., 2002. On the scope of experiments in economics: comments on Siakantaris. Cambridge Journal of Economics 26, 261–267.
Guala, F., 2005. The Methodology of Experimental Economics. Cambridge University Press, Cambridge.
Guala, F., Mittone, L., 2010. Paradigmatic experiments: the dictator game. Journal of Socio-Economics 39, 578–584.
Haisley, E.C., Weber, R.A., 2010. Self-serving interpretations of ambiguity in other-regarding preferences. Games and Economic Behavior 68, 614–625.
Harrison, G.W., 2007. House money effects in public good experiments: comment. Experimental Economics 10, 429–437.
Harrison, G.W., Lau, M.I., Rutström, E.E., 2009. Risk attitudes, randomization to treatment, and self-selection into experiments. Journal of Economic Behavior and Organization 70, 498–507.
Hertwig, R., Ortmann, A., 2001. Experimental practices in economics: a methodological challenge for psychologists? Behavioral and Brain Sciences 24, 383–451.
Hertwig, R., Ortmann, A., 2008. Deception in experiments: revisiting the arguments in defense. Ethics and Behavior 18, 59–92.
Hey, J.D., 2005. Why we should not be silent about noise. Experimental Economics 8, 325–345.
Hoffman, E., McCabe, K., Shachat, K., Smith, V., 1994. Preferences, property rights, and anonymity in bargaining games. Games and Economic Behavior 7, 346–380.
Hogarth, R.M., Einhorn, H.J., 1992. Order effects in belief updating: the belief-adjustment model. Cognitive Psychology 24, 1–55.
Holt, C.A., 1992. Industrial organization: a survey of laboratory research. In: Kakel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 349–443, http://dx.doi.org/10.1016/0010-0285(92)90002-J.
Huck, S., Normann, H.-T., Oechssler, J., 1999. Learning in Cournot oligopoly – An experiment. Economic Journal 109, C80–C95.
Huck, S., Normann, H.-T., Oechssler, J., 2004. Two are few and four are many: number effects in experimental oligopolies. Journal of Economic Behavior and Organization 53, 435–446.
Jackson, S.L., 2008. Research Methods: A Modular Approach. Wadsworth, Belmont, CA.
Jones, M.K., 2008. On the autonomy of experiments in economics. Journal of Economic Methodology 15, 391–407.
List, J.A., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115, 482–493.
Mayo, D.G., 1996. Error and the Growth of Experimental Knowledge. Chicago University Press, Chicago.
Mill, J.S., 1843/2009. A System of Logic. Ratiocinative and Inductive. Harper & Brothers/Project Gutenberg EBook, New York.
Millner, E.L., Pratt, M.D., Reilly, R.J., 1990. Contestability in real-time experimental flow markets. Rand Journal of Economics 21, 584–599.
Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? Journal of Public Economics 92, 91–112.
Oberholzer-Gee, F., Eichenberger, R., 2008. Fairness in extended dictator game experiments. BE Journal of Economic Analysis and Policy (Contributions) 8, 16.
Patten, M.L., 2007. Understanding Research Methods: An Overview of the Essentials. Pyrczak Publishing, Glendale, CA.
Reinstein, D., Riener, G., 2012. Reputation and influence in charitable giving: an experiment. Theory and Decision 72, 221–243.
Requate, T., Waichman, I., 2011. A profit table or a profit calculator? A note on the design of Cournot oligopoly experiments. Experimental Economics 14, 36–46.
Rigdon, M., Ishii, K., Watabe, M., Kitayama, S., 2009. Minimal social cues in the dictator game. Journal of Economic Psychology 30, 358–367.
Rustrom, E.E., Wilcox, N., 2009. Stated beliefs versus inferred beliefs: a methodological inquiry and experimental test. Games and Economic Behavior 67, 616–632.

Schram, A., 2005. Artificiality: The tension between internal and external validity in economic experiments. Journal of Economic Methodology 12, 225–237.
Shang, J., Croson, R., 2009. A field experiment in charitable contribution: the impact of social information on the voluntary provision of public goods. Economic Journal 119, 1422–1439.
Slonim, R., Roth, A.E., 1998. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. Econometrica 66, 569–596.
Smith, V.L., 1982. Microeconomic systems as an experimental science. American Economic Review 72, 923–955.
Smith, V.L., 1994. Economics in the laboratory. Journal of Economic Perspectives 8, 113–131.
Smith, V.L., 2010. Theory and experiments: what are the questions? Journal of Economic Behavior and Organization 73, 3–15.
Stöber, J., 2001. The social desirability scale-17 (SDS17): convergent validity, discriminant validity, and relationship with age. European Journal of Psychological Assessment 17, 222–232.
Sugden, R., 2005. Experiments as exhibits and experiments as test. Journal of Economic Methodology 12, 291–302.
Sutter, M., Kocher, M.G., Strauβ, S., 2007. Individuals and teams in auctions. In: Social Science Research Network and University of Innsbruck Discussion Paper 2007-23, October.
Sutter, M., Kocher, M.G., Strauβ, S., 2008. Individuals and teams in auctions. Oxford Economic Papers 61, 380–394.
Vieider, F.M., 2011. Moderate stake variations for risk and uncertainty, gains and losses: methodological implications for comparative studies. Economics Letters (in press).
Zizzo, D.J., 2010. Experimenter demand effects in economic experiments. Experimental Economics 13, 75–98.
Zizzo, D.J., Fleming, P., 2011. Can experimental measures of sensitivity to social pressure predict public good contribution? Economics Letters 111, 239–242.