Title: There is Cause to Randomize

Abstract: While practitioners think highly of randomized studies, some philosophers argue that there is no epistemic reason to randomize. Here I show that their arguments do not entail their conclusion. Moreover, I provide novel reasons for randomizing in the context of interventional studies. The overall discussion provides a unified framework for assessing baseline balance, one that holds for interventional and observational studies alike. The upshot: practitioners' strong preference for randomized studies can be defended in some cases, while still offering a nuanced approach to evidence-appraisal, one where not all non-randomized studies are treated equally.

Funding: This paper is based on research that was funded by ANID (Chile) and Gates Cambridge Trust.

Contact Information: Name: 'Cristian', Surname: 'Larroulet Philippi'. Affiliation: Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK; e-mail: cristianlarroulet@gmail.com. ORCID: 0000-0001-5793-4670.

Acknowledgments: Thanks to Katharina Bernhard, Adrian Erasmus, Julia Staffel, Jacob Stegenga, and the referees for helpful comments and suggestions.

1. Introduction.

Most researchers in statistics, biomedical, and social sciences hold in high esteem the random allocation of subjects to experimental groups for causal inference. But some philosophers of science have challenged the arguments usually provided for randomization, to the point where they consider randomization as, "for the most part, … a waste of effort and resources" (Urbach 1985, 258). Their arguments have been—and remain—influential. Now that important policy movements are downplaying the role of randomized studies for treatment approval (e.g., the *21st Century Cures Act* in the US), the time is ripe for reassessing their arguments.

Some arguments for randomization—e.g., that it "guarantees" strict comparability between experimental groups—are clearly exaggerations. The most trenchant criticisms of randomization, by Peter Urbach (1985; Howson and Urbach 2006) and John Worrall (2002, 2007), provide both good compilations of these claims and sound arguments against them. However, I contend, Urbach and Worrall's conclusion—that there is no reason to randomize—is incorrect.

Importantly, my arguments do not assume the frequentist approach to statistical inference. Like Urbach and Worrall, I focus mainly on the balance between experimental groups. Although the focus is specifically on experimental design, the framework for assessing baseline balance presented below has two further payoffs. It rationalizes practitioners' strong

preference for randomized studies in some cases, while still offering a nuanced approach to evidence appraisal.

I start by restating the problem of causal attribution (2). I then clarify where the substantive disagreement lies between the (more reasonable positions among) practitioners and the critics of randomization (3). Urbach and Worrall's main arguments are discussed in sections (4) and (5), where I show they do not entail that there is no epistemic value in randomizing. Sections (6) and (7) provide novel epistemic and non-epistemic reasons to randomize in the context of interventional studies.

2. The Problem.

Randomized studies are taken to be a solution to the problem of determining the causal impact that one variable (sometimes called "independent") has on another variable ("outcome" or "dependent") within a study population. Here the causal impact of a variable is understood as the difference that it makes on a specific outcome in the study population.[1]

Using a simplified example will prove helpful. In the simplest (but common) case, the independent variable of interest (*T*) is dichotomous. When researchers observe, in a given

[1] The program evaluation literature defines causal effects using Donald Rubin's potential outcomes model (e.g., Angrist and Pischke 2009). Here I follow Deaton and Cartwright's (2018, 3) more neutral presentation.

study population, a difference in the outcome of interest between those for whom T=1 (the "treatment" group) and those for whom T=0 (the "control" group), that difference may be attributed to the causal impact of T and/or to the causal impact of other variable(s). Imagine a researcher interested in the impact that some job training programs (T) have on labor market outcomes (Y), typically wages or employment-status. Labor market outcomes are related to many variables, such as education (schooling), work experience, gender, and family networks. Let us formalize these relations with a simple linear model: $Y_i = \beta T_i + \boldsymbol{X'}_i\boldsymbol{\gamma}$. Subscript i refers to the individual, T to attending (or not) a job training program, and β to the impact of T on Y.[2] Vectors $\boldsymbol{X}$ and $\boldsymbol{\gamma}$ refer, respectively, to all the minimally sufficient factors (apart from T) that affect labor market outcomes and their respective impact on them. Factors in $\boldsymbol{X}$ are called "covariates" (Deaton and Cartwright 2018), "prognostic factors" (Urbach 1985), or "potential confounders" (Fuller 2019).[3]

[2] If T has no impact, $\beta = 0$.

[3] I'm simplifying here. First, I'm assuming that the impact of T on Y is the *same* for all subjects ($\beta_i = \beta$ for all i). I'm also glossing over Fuller's (2019) distinction between potential confounders that are "directly causal" and those that are "associational-causal." Finally, as said, the factors in $\boldsymbol{X}$ are minimally sufficient causes in the study population, unlike Fuller's (2019) "confounders," which are only components of sufficient causes. None of these subtleties affect my arguments below.

Imagine our researcher is looking at labor market outcomes of participants and non-participants of a job training program in an area of India. She observes that the average income of participants is higher than that of non-participants (i.e., $E(Y_i|T=1) > E(Y_i|T=0)$).[4] Is she warranted in attributing the observed difference to the causal impact of T? Not if she suspects that those who participated differ in *other* variables that also affect income (e.g., education). Imagine participants have higher education, but do not differ in other variables of $\boldsymbol{X}$ (age, gender, etc.). If our researcher does not have data on education, she cannot control for that variable. In econometric jargon, she faces an "omitted variable bias" problem. This is just another way of saying that the two groups the researcher is observing are not "balanced" or "matched" with respect to covariates.

3. The Disagreements.

Both Urbach's (1985, Howson and Urbach 2006) and Worrall's (2002, 2007) discussions largely focus on the value of randomly[5] assigning subjects for attributing causation. Thus,

[4] $E(\bullet)$ refers to the expectation in the study population (i.e., the average in the observed data).

[5] Here, "randomly assign" means assigning by a procedure taken to be random by practitioners. In general, there are two categories. First, when the randomization is done at the researchers' offices, it is done by using "random numbers" generated by a software. Second, when the randomization is done in the field, it is done by using physical randomization procedures like

their focus is on what I call "simple randomization," in contrast to "stratified randomization" (discussed below). The problem they raise for simple randomization is that in any given trial, we might just be "unlucky," and get imbalanced groups with respect to some covariate. Their response: instead of randomizing, researchers should control for all suspected covariates. In our running example, the researcher should assemble groups that are balanced regarding each covariate in $\boldsymbol{X}$. If what threatens the causal inference is imbalance in covariates, these authors ask, why prefer to assemble groups by a chancy method versus by deliberately matching the covariates?

Many practitioners, at least in the social sciences, would look at this discussion with surprise. For one, it is not news that a random assignment can produce baseline imbalance, especially in small samples (see Senn 2013; Deaton and Cartwright 2018). More importantly, practitioners take for granted that if we *could* control for *all* covariates there would be no problem of causal inference. Thus no need to randomize. But we are never, or almost never, in possession of a database with *all* covariates. The list of variables included in $\boldsymbol{X}$ is very long—many factors affect the typical outcomes we are interested in. Of course, some of them might be unknown to us. But, even if we know what all the covariates are, there are some of which we are almost never in possession. A stock example: Wages are affected by schooling,

---

drawing balls from urns. Neither Urbach nor Worrall provide a definition of randomization, but this is probably what they have in mind.

work experience, gender, and other variables researchers usually have access to; but also by harder-to-measure and usually-unavailable variables such as subjects' "ability", "motivation," and "family connections" (Blackburn and Neumark 1995). So the proposed solution is not feasible—researchers do not have all suspected covariates available to control for.

What advocates of randomization have in mind, at least in economics and in the program evaluation literature more generally, is stratified (or blocking) randomization. This involves forming groups based on observed covariates and then randomly assigning participants to treatment within these groups.[6] Doing so effectively controls for observed covariates before randomization, which amounts to applying Urbach and Worrall's proposed solution to the extent possible, and then randomizing.

Why have Urbach and Worrall focused largely on the distinction between simple randomization and controlling for all covariates (that is, between the undesirable and the

[6] Bruhn and McKenzie's (2009) survey article shows that most researchers in development economics use stratified (*not* simple) randomization. The recently published *Handbook of Field Experiments* also recommends stratified randomization (Duflo and Banerjee 2017, 76). An alternative practice to stratification, re-randomization, also secures balance in observed covariates. Re-randomizing becomes more practical as the number of observed covariates increases.

unfeasible)? Worrall explicitly recognizes the alternative strategy of stratified randomization, but clarifies he focuses on simple randomization because that is "what most commentators have in mind in assessing the power of the methodology" (2007, 452 fn. 1). This is not, as I have suggested, what most social science *researchers* have as standard practice. Perhaps Worrall's focus on the medical literature played a role here.[7] Urbach seems to think that controlling for all likely covariates is a feasible option. This seems to me the only way to understand Urbach's (striking) remark that the situation in which there are "certain factors …believed to have an influence on the experimental outcome but [for which] we possess no means whatsoever to detect their presence, however crudely" is "a rather unusual situation" (1985, 267). This optimism seems as unwarranted in medicine as in the social sciences (Deaton and Cartwright 2018).

Alas, the previous discussion does not resolve all disagreement between Urbach & Worrall on the one hand, and practitioners on the other. It does, however, help us situate more precisely the specific disagreement regarding experimental design:

[7] See (Worrall 2002, S323 fn. 6). Moreover, stratification and re-randomization are not feasible when the recruitment of subjects overlaps in time with randomization (as is not uncommon in medical trials). Thanks to a referee for this point.

(D) In an interventional study, after having controlled for the observed covariates (e.g., by stratifying), is there any (pro tanto) *specific* reason to randomize (i.e., one that is not already a reason for some other allocation method)?

Against practitioners, Urbach and Worrall argue that, after controlling for observed covariates, there is no special value in randomization for producing baseline balance. (Randomizing can do "no further epistemic good," says Worrall 2007, 463.) Given that they do not see any other reason at stake, faced with (D) they conclude, in Worrall's words, "there's no cause to randomize." Before assessing their arguments, I clarify some aspects of (D).

Note that (D) is about experimental design. It takes as given that we are in the context of an interventional study (i.e., a prospective trial where the researcher has control over the method of subjects' assignment to experimental groups). Moreover, in (D) the focus is on the balance of *unobserved* covariates, since the observed covariates are controlled.

I emphasize that question (D) is different from questions such as:

(D'): Are randomized studies (always, or in general) better for causal inference than observational studies?

(D''): Should we conduct a randomized trial to answer question X?

(D) versus (D'). The quality of a study depends on several factors besides baseline balance (e.g., the representativeness and size of the study population, the measuring instruments, blinding procedures, statistical analyses, etc.) Therefore, categorical rankings ("evidence hierarchies") constructed considering only one dimension—namely, whether the study is randomized or not—are poor tools for assessing the *overall quality* of studies. Better tools—called "quality assessment tools" by Stegenga (2018, 76)—assess studies taking into account all the relevant factors. Which factors are relevant depends on the specific causal question at stake (Deaton and Cartwright 2018). Thus, a positive answer in (D) does not entail a positive answer in (D').

(D) versus (D''). Clearly, to *decide* to conduct an interventional study (regardless whether group assignment is random or not) requires considering not only the several quality-of-evidence aspects just mentioned besides balance, but also non-epistemic (ethical and financial) considerations. Withholding treatment to patients may be unethical, and interventional studies may be costlier than alternatives. So, a positive answer in (D) does not entail a positive answer in (D'').

Crucially, these further epistemic and non-epistemic considerations do not bear on (D). The non-epistemic considerations mentioned are pertinent for the *decision* of conducting an interventional study (versus using data already available for an observational study). But once this decision has been taken, and we ask how to allocate subjects to experimental groups, those considerations no longer make a difference—they apply equally to randomized

and non-randomized interventional studies.[8] The same is true regarding the further epistemic considerations: assigning randomly (versus not) does not affect the other quality-of-evidence factors already mentioned.

Nonetheless, (D) is not just an important question about scientific methodology. Although an answer to (D) neither fully settles the evidence-appraisal question (D'), nor the pragmatic question (D''), it plays an important role in answering them. To see this, note that all current quality assessment tools give a positive value to randomization, varying in the weight given to this factor (see Stegenga 2018). But what justifies this positive value? A claim like "ceteris paribus randomization is better for baseline balance" would justify this positive value. Urbach and Worrall reject this claim, and I will defend it while arguing for a positive answer to question (D). In this sense, this paper's conclusion is not only relevant for the experimental design question. Moreover, thorough consideration of (D) will reward us with a unified framework for assessing the expected baseline balance of both interventional and observational studies.

4. The Non-Comparative Argument.

[8] For example, moral reasons against giving only a placebo to the control group hold (or not) independently of whether its membership is allocated randomly or by another method.

Urbach and Worrall provide two major arguments against any special benefit accruing from randomization—one non-comparative and the other comparative. The non-comparative argument aims at challenging directly what they take to be the best reason for randomization: that randomization makes imbalance in unobserved covariates unlikely.[9] The idea is that any randomly generated variable is not systematically (i.e., in general, in the overall population) correlated to *Y*'s covariates.[10] That is, if we randomly assign members of the population as a whole to two groups, the groups will not differ in *Y*'s covariates. So, the idea goes, it is unlikely for this correlation to appear in our specific study population (see Worrall 2002, S323).

Worrall (2002) acknowledges that the larger the trial, the less likely it is that a *given* unobserved covariate is unbalanced under randomization. This, however, can only support the claim that randomizing (in large studies) makes significant imbalance unlikely *in any given* unobserved covariate. But this is not sufficient to defend randomizing, since the real question for inferring causation in a given trial seems to be whether there is balance in *all* unobserved covariates. And as Urbach (Howson and Urbach 2006, 195-6) and Worrall (2002, S323-4) say, there might be "innumerable" unobserved covariates. So, the probability

[9] Another main reason they discuss relates to selection bias (see section 5).

[10] Variables such as, say, "education," "motivation," "ability," etc., do not share common causes with (nor cause or are caused by) the randomly generated variable.

that by pure bad luck, after randomizing in a given trial, there is imbalance in *some* unobserved covariate or other may, for all we know, be quite high.

Now, pointing out that (even) with random assignment there is a non-negligible possibility of imbalance in some unobserved covariate does not constitute a reason *against* randomizing. For all we know, this possibility is equally (or more likely!) present in any alternative non-random assignment mechanism yet to be proposed. At least, nothing has been said to the contrary. Nevertheless, their argument undercuts the *positive* reason for randomizing under consideration, namely, that imbalance is unlikely if we randomize. So, their conclusion does not entail randomization is on a par with non-randomization (recall, in the context of (D)), but it is pertinent to our question because it undercuts a positive reason for positioning randomization above non-randomization.

Urbach and Worrall are correct to reject the claim that randomization makes the probability that there is imbalance in some unobserved covariate or other small. However, this entails a threat to internal validity (i.e., whether we can attribute causation within the study population) only if we assume that all covariates need to be balanced for warranting a causal attribution. Fortunately, the condition for sound causal attribution in our context is *not* as strict. Look at the difference in means of the experimental groups:

(1) $$E(Y_i|T=1) - E(Y_i|T=0) = E(\beta T_i + \boldsymbol{X'}_i\boldsymbol{\gamma}|T=1) - E(\beta T_i + \boldsymbol{X'}_i\boldsymbol{\gamma}|T=0)$$

$$= \beta + E(\boldsymbol{X'}_i\boldsymbol{\gamma}|T=1) - E(\boldsymbol{X'}_i\boldsymbol{\gamma}|T=0)$$

The second line indicates that the difference in means the researcher observes amounts to *T*'s impact (β) plus the overall difference across groups with respect to the covariates weighted by their effects. Thus the condition that needs to be satisfied for causal attribution is that expression (2) below be equal to (or not significantly different from) 0.[11]

(2) $E(\boldsymbol{X'}_i\gamma|T=1) - E(\boldsymbol{X'}_i\gamma|T=0)$

This condition is substantially *weaker* than requiring each and every covariate to be balanced. What is required here is only that the sum of the covariates' effects on *Y* produces no substantial net difference across groups.[12] It has been argued that this weaker condition, unlike the stronger one, is likely to hold after randomization in large trials (Fuller 2019, 923).

[11] This is the condition needed for attributing the *total* observed difference in outcome to *T*. Attributing the total observed difference is important for quantifying the causal impact of *T*. If we only aim at justifying that *T* has some impact, we just need to be justified in the claim that the imbalance in covariate is smaller than the total observed difference in outcome (see Fuller 2019, 920).

[12] This weaker condition is endorsed by Deaton and Cartwright (2018) and (in its dichotomic version) by Fuller (2019). But it has been accepted in the program evaluation literature since Rubin's work in the 70's.

Now, in the context of (D), some of the variables in $\boldsymbol{X}$ are directly controlled for. Using our example, let us suppose the researcher has data on schooling, age, and gender, but not on other variables (e.g., ability, motivation, etc.). And let us distinguish between the covariates the researcher observes ($\boldsymbol{X}_{ob}$) and those she does not observe ($\boldsymbol{X}_{un}$). The expression (2) can be decomposed as follows:

$$E(\boldsymbol{X'}_{i,ob}\boldsymbol{\gamma}_{ob}|T=1) + E(\boldsymbol{X'}_{i,un}\boldsymbol{\gamma}_{un}|T=1) - \Big(E(\boldsymbol{X'}_{i,ob}\boldsymbol{\gamma}_{ob}|T=0) + E\big(\boldsymbol{X'}_{i,un}\boldsymbol{\gamma}_{un}|T=0\big)\Big)$$

Because of stratification, $\boldsymbol{X'}_{i,ob}\boldsymbol{\gamma}_{ob}$ is the same in both groups (so they are cancelled). In the context of (D), then, condition (2) is simplified into (2'):

(2') $E(\boldsymbol{X'}_{i,un}\boldsymbol{\gamma}_{un}|T=1) - E(\boldsymbol{X'}_{i,un}\boldsymbol{\gamma}_{un}|T=0)$

What is required here is only that the sum of the *unobserved* covariates' effects on Y produces no net difference across groups. In typical randomized studies, some of the main covariates (i.e., those that have the largest γ's) are measured at baseline (thus, they are part of $\boldsymbol{X}_{ob}$). As long as this is the case, the condition needed for causal attribution is more likely to hold in stratified randomization than in simple randomization. This is because after controlling for the main covariates, there are less variables left to produce significant imbalance, and those left have less capacity to do so (due to their smaller coefficients).

What does this entail for our question (D)? The reason for randomizing under scrutiny was that it makes imbalance in any given unobserved covariate unlikely. Urbach and Worrall

challenged the relevance of that result, and argued against the claim that imbalance in each and every unobserved covariate is unlikely. We found, however, that their result is also not completely relevant for the validity of causal inference. Thus, their result does not challenge the virtues of randomization. Instead, what is relevant and correct is that randomizing makes significant imbalance in (2) unlikely (in large trials), and that stratified randomization makes significant imbalance in (2) even more unlikely. Nevertheless, although relevant for the assessment of the virtues of randomization, this is not immediately relevant for (D). *For all we know, the alternative to randomization gives the same result.* In order to answer (D), we need to assess *comparative* arguments—arguments that directly discuss the merits of randomization versus its alternative.

5. The Comparative Argument.

What is striking in the literature is that very little discussion has been given to what exactly is the alternative to randomization when asking (D).[13] If the question is what is the specific value of randomizing, knowing the alternative to randomization seems key. Urbach (1985) is pretty much the only source that mentions some specific alternatives to randomizing. It is worth quoting him at length.

[13] Senn (2013, 1448) shares this concern.

> [T]here is no reason to think that [unobserved covariates] would balance out more effectively between the groups by using a physical randomizing device rather than by employing any other method. In short, it seems reasonable to use any method of distributing subjects to different treatment groups, provided that there is no evidence that the selection procedure will produce unbalanced groups. (1985, 267)

> [I]t should be noted that a randomized design is not the only, or even the best, way of instilling confidence that the trial was free from unconscious experimenter bias. The same result could be got by insisting on the allocation being done by a person without any axe to grind or by one who has no ability or knowledge to exercise a prejudice. For example, in a medical trial the selection of patients might be entrusted to an independent, non-medical person, or it might be performed according to the order in which they present themselves, or by whether their hair is parted on the left or the right, or one could simply permit the subjects to choose their own groups, always ensuring of course that they have not been informed of which treatment is to be applied to which group. (1985, 271)

The alternative to randomization Urbach suggests is using what I will call an *unsuspicious* variable (U, for short). A U is a variable, not randomly generated, about which the researcher has no positive evidence that it will produce baseline imbalance. Urbach argues that it is *not* the case that we should randomize, because any U is as good as randomizing (Worrall follows him here; see 2007, 463). Urbach sometimes claims something stronger—that

randomization is not "even always the best way of constructing the treatment groups in a clinical trial" (Howson and Urbach 2006, 259).

From a Bayesian perspective, clarifies Urbach, what matters for causal inference in this context is that alternative hypotheses are ruled out (1985, 260-270; Howson and Urbach 2006, 255-259). We are interested in the hypothesis that *T* has a positive effect on *Y* in the study population (hypothesis *H)*. When, in our running example, we observe a substantive difference in income across experimental groups (evidence *E*), the degree to which *E* corroborates *H* depends on how well alternative hypotheses explain *E* and how likely those alternative hypotheses are. Thus, if we suspect that, say, education is unbalanced enough to make (2) differ from 0 significantly, *H* does not receive much support from *E*. This is because *E* is also explained by the imbalance in (2) produced by education. Had we controlled for education, then, education could not explain *E*. Controlling for each suspected covariate makes it the case that *E* strongly supports the hypothesis that the treatment affects income, since no other suspected cause could have played a role. This is the Bayesian rationale for having all suspected covariates balanced. "In this approach," concludes Urbach, "there is no advantage to be gained from allotting patients … by some physical stochastic process. Any method of allocation is satisfactory, so long as we have no reason to think that it will have a material influence on the outcome of the experiment" (1985, 270).

Using this reasoning, both authors downplay the relevance for (D) of the argument that randomization controls for what they call "unconscious experimenter bias" or "selection

bias." If the alternative to randomizing is (not a U, but) just physicians choosing who gets treatment and who does not, Urbach and Worrall acknowledge that unconscious desires to benefit patients can lead to unbalanced groups (Urbach 1985, 270-1; Howson and Urbach 2005, 258; Worrall 2002, S324-5). Randomization, they agree, can be a useful way of avoiding this situation. However, any U does the trick. So, the fact that randomization controls for selection bias gives us *no* specific reason to randomize in (D).

Summing up, Urbach and Worrall conclude that there is no advantage to be gained from randomizing—thus, that it is *not* the case that we should randomize—by arguing that any U is as good as randomizing when it comes to attributing causality. There are, I believe, two flaws here. First, the main premise is incorrect—as argued in section (6), it is not true that any U is as good as randomizing for causal inference. Second, even granting the premise the conclusion does not follow. Imagine that any U is as good as randomizing for causal inference in (D), in the sense that both warrant a causal attribution. It does not follow that there are no reasons to randomize: differences in research designs that do not matter for inference might still matter for action. This distinction is well-known to Bayesians. Consider two researchers that perform the exact same experimental protocol except that one uses an optional and the other a non-optional stopping rule. If both obtain the same data, Bayesians insist both researchers should draw the same inference. As Steele (2013) shows, however, this does not entail that the Bayesian is indifferent between *choosing* an optional versus a non-optional stopping rule. Thus, no difference for inference does not entail no difference for

action. In sections (6) and (7) I provide reasons researchers have for randomizing, reasons that hold *even if* we were to grant Urbach and Worrall's claim that any U is as good as randomizing for inference.

6. Randomization is more reliable.

I will now argue that randomization is a more reliable process for assigning subjects to experimental groups than using any U. That is, randomization is less prone to form unbalanced groups. This is a good reason for *choosing* randomization over U, and so it directly challenges Urbach and Worrall's conclusion. Moreover, against Urbach and Worrall's main premise, I will also argue (later in this section) that randomization makes a difference for *inference*.

In which sense is randomization less prone to failure? I think there are two main ways in which an assignment method might fail in a particular instance. First, as discussed above, even if we assemble the experimental groups using a variable that is not systematically related to covariates, we might get baseline imbalance in our trial just by pure bad luck. This holds for randomization, but also for any U. A variable is unsuspicious for a researcher only because the researcher has no reason to think there is a systematic relation (present in general, beyond the study population) between the U and *Y*'s covariates. To use Urbach's example, perhaps we do not expect the variable whether-their-hair-is-parted-on-the-left-or-the-right to correlate in general with any significant covariate of, say, wages. However, even

if we are correct, we might be "unlucky," so that in our particular study population they happen to correlate. Just like with randomization, using a U does not guarantee a non-systematic (due to pure bad luck) connection that is enough to produce baseline imbalance in the study population. So, with respect to this first way of failing, there is a tie.

The second way in which an assignment method might fail in a particular instance is when there is a systematic connection between the assignment variable and *Y*'s covariates. I can think of two plausible mechanisms triggering this failure when using a U. First, the researcher for whom a variable is unsuspicious (say, the second and third character of a subject's surname), may be *ignorant* of the fact that this variable is systematically related to some covariates. Our researcher conducting the job training evaluation in India, plausibly, say, from Boston, might just not know that Indian surnames relate to caste divisions (within areas). Urbach's argument implies that this researcher has *no* reason to prefer randomization over using names. This seems controversial enough. But Urbach's conclusion holds *even if the researcher knows* that she is not familiar with how names in India are formed. As long as she has "no evidence that the selection procedure will produce unbalanced groups" (which, presumably, she won't have if she knows little about Indian surnames), according to Urbach, she has no reason to randomize instead of using a U. In contrast, it seems plain that this researcher has a strong reason to randomize: her U might be systematically related with some unobserved covariates, which would make imbalance in those covariates very likely in the study population, thus threatening baseline balance. And this risk is not one she runs with

randomization. A random generated number is, in a sense, designed and created for not being systematically related to anything that could be a covariate in a study.

You may think this is not a strong enough reason for researchers who know (and know that they know) many details about a particular U. Perhaps these researchers would not use, say, the parted-hair variable: although it is a U for them, they know nothing about what drives variations of it. They will rather choose a U they are knowledgeable about. It is true that choosing U is less risky for these knowledgeable researchers than for the ignorant-about-U-researcher. But the larger point still holds for them—the possibility of failure due to ignorance of a systematic relation remains, whereas it is not present for these knowledgeable researchers if they randomize. So they also have a reason to prefer randomization.

Importantly, this way of failing is not a mere theoretical possibility, but something that happens in science. I cannot provide straightforward examples because I know of no interventional study that used a U instead of randomizing (which is, perhaps, telling of the issue at stake). However, there *are* examples where a particular variable was thought to play a role in an observational study analogous to what a U would play in an interventional study, and then (later on) shown to be correlated in general with covariates. I illustrate with observational studies that use the instrumental variable method for causal attribution.

Imagine I am interested in the impact that years of schooling (T) have on wages (Y). The problem I face has already been explained: those with more schooling differ in other

(observed and unobserved) characteristics. We can control for the observed characteristics, but not for the unobserved ones. What the instrumental variable method proposes is the following. If you find another variable (*Z*) that affects *T* and does not affect *Y* by a route other than *T*, then you can use *Z* to estimate the causal impact of *T* on *Y* (see details in Angrist and Pischke 2009). More specifically, for *Z* to be a valid instrumental variable, *Z* must have no effect on *Y* other than the one it has through *T,* nor be correlated with other unobserved covariates of *Y.* That is, conditional on the value of *T*, *Y* has to be independent of *Z.* Here we see the connection with U: *Z* has to be a variable that, when used to compose groups for comparison, produces groups that differ on *T* but do not differ in other covariates of *Y*. Thus, when a researcher believes that a variable is a *Z,* that variable is a U for that researcher. Of course, the researcher might be wrong about it being a *Z*.

There are many instances in which researchers thought a variable was a *Z* and they were then proved wrong. For one, Angrist and Krueger (1991) famously thought they had found a *Z* for estimating the impact of schooling on wages in the US—the quarter of birth. Because of laws regarding compulsory age for schooling, quarter of birth and schooling years are related in the US population. This, and the belief that quarter of birth is not related to other covariates of wages, led Angrist and Krueger to believe that quarter of birth is a good instrument for schooling. Thus, for them, quarter of birth was a U. Importantly, many other reputed researchers thought this as well (see Buckles and Hungerman 2013, 711). Alas, two decades later, Buckles and Hungerman (2013) showed that the season of birth is not as orthogonal to

covariates as one would require it to be for the instrumental variable strategy to work. Plainly, we are not talking of mere theoretical possibilities here.

A second plausible mechanism is *unconscious* (also called "implicit") bias. A researcher might consciously consider a variable as a U, but this might just be that she is not aware of the reasons she had for picking it. The researcher, for instance, might think to herself that the time subjects arrived at the survey center is uncorrelated with any covariate, making that variable a U. However, she might be unaware that her preference for this variable derives in fact from her wanting to 'give a good chance to the study,' and thinking this variable indicates punctuality, an attribute which might be important for the treatment to have any effect. In such cases, of course, the researcher would be choosing an allocation method that, though unsuspicious for her, is systematically related to unobserved covariates (punctuality). Just like with ignorance, this risk is not run with randomization.

How relevant is this consideration? Bear in mind that I'm not imputing to the researcher a conscious intent to rig the study. This is only about implicit bias. And it would be hard today to think of implicit biases as something too uncommon to consider (Brownstein and Saul 2016). If the question at stake, (D), is meant for us (non-ideal epistemic agents), this possibility of failing is relevant.

Randomization, then, is more reliable in the following sense. There are two ways in which an assignment method might fail in a particular instance. Both randomizing and using a U are

susceptible to the first way, namely, getting baseline imbalance just by pure bad luck. The other way is researchers picking a U that systematically relates to some unobserved covariates—due to ignorance or unconscious bias. This connection makes imbalance in those covariates in the study population very likely. Since this risk arises only by using a U—whereas the risk of getting baseline imbalance by pure bad luck is common across methods—researchers have a strong reason to *choose* randomization over using a U. Whatever the probability of wrongly believing there is balance (and to that extent attributing causation) when the only risk is that of pure bad luck, it can only increase by choosing a U versus randomization.

So far we have directly challenged Urbach and Worrall's overall conclusion—we found a reason for *choosing* to randomize. What about their premise, that any U is as good as randomizing for *inferring* causation?[14] The degree to which *E* confirms *H* depends on the plausibility of alternative hypotheses. The less likely the alternative hypotheses, the more *E* confirms *H*. That is why, Howson and Urbach insist, "*the chief concern when designing a clinical trial should be to make it unlikely that the experimental groups differ on factors that are likely to affect the outcome*" (2006, 259 their emphasis). One general alternative hypothesis, H*, says: (2') is sufficiently larger than 0 due to an unbalanced distribution of some unobserved covariates or others.

[14] I thank a referee for suggestions here.

Randomization, I argued above (Section 4), makes H* unlikely (in large trials). More importantly for comparative claims, however, this section's argument entails that H* is more plausible if the assignment is made with a U than with randomization: (2') might be sufficiently larger than 0 due to a systematic connection with unobserved covariates (about which researchers are ignorant, or consciously unaware). In contrast, learning that the allocation was randomized gives us evidence against H* by ruling out some of the mechanisms that could produce baseline imbalance (namely, any systematic connection). Since H* is more plausible under a U, *E* confirms *H* to a lesser degree under a U. Thus, it is not true, even by Bayesian lights, that any U is as good as randomizing for causal inference.

To be clear, this conclusion is comparative—*E does* confirm *H* under a U, but to a lesser degree than with randomization. Now, we can say something more fine-grained here, and in this way start developing a unified framework for assessing the quality of studies with regards to balance. How much of a difference does randomization versus a U make? This depends on how plausible H* is under the particular U. Arguably, not all U's are equal. Following Urbach, we defined U as a variable, not randomly generated, about which researchers have no positive evidence that it will produce baseline imbalance. But absence of positive evidence about imbalance is compatible with different amounts of evidence (or theoretical reasons) *for* balance. As hinted above, somebody ignorant about the particular U may have no reason to think that groups are unbalanced when subjects are assigned by that U. But this lack of reason, mostly driven by ignorance, should *not* generate ample confidence

in the assignment. In contrast, somebody knowledgeable about the particular U, who knows about its causes and effects, can reject some mechanisms linking U and covariates of *Y*. This knowledge gives her some reason to be more confident in assigning subjects according to that U—she knows that her lack of evidence about U's connection with covariates is not due to a general lack of evidence in the matter.[15] In this way, by ruling out mechanisms linking U and covariates of *Y*, H* becomes less plausible and *H* better confirmed.

Take an extreme (and impossible) case for illustration—somebody who knows everything about a U. In particular, someone who knows how that variable varies in the overall population with any covariate of *Y*. This person, then, knows that the variable at stake is not systematically correlated with *Y*'s covariates (otherwise it would not be a U). She is in the same epistemic position with that U as with randomizing—in both cases a systematic connection between the variable used for assigning subjects and covariates of *Y* can be rejected. Thus, for this imagined researcher *E* confirms *H* equally under U than under randomization. The upshot should be clear: the more researchers know about a U, the less of a difference randomization makes for them. Still, as argued, for *real-life* researchers,

[15] This is analogous to the case where absence of evidence provides evidence of absence (Sober 2009).

randomizing should make even the very knowledgeable researcher more confident in the assignment.

Since one of the goals of this paper is to (partially) conciliate practitioners and philosophers of science, I close this section showing how my arguments help reduce some of the current disagreement regarding the value of randomization. I focus on two positions by practitioners that philosophers of science (partly influenced by Urbach and Worrall's arguments) have criticized. First, it is not difficult to see among some practitioners a *general* suspiciousness about observational studies. Second, in some areas, researchers show almost a fixation with randomization, to the point that they seem to consider randomization necessary for (statistical) causal inference. Both Urbach (1985; Howson and Urbach 2006) and Worrall (2002, 2007) quote several researchers exemplifying these attitudes.

For starters, note that it is common practice to estimate causal impacts from observational studies in social science. Even those scientists best known for their randomization activism conduct observational studies where they attribute causation (e.g., Banerjee et al. 2010, Duflo 2004). Several research designs have been developed for this purpose—instrumental variables, differences-in-differences, regression discontinuity design, etc. (see Angrist and Pischke 2009).[16] All these methods aim to make the case for valid causal inference, and the

[16] Though widely taught in econometrics courses, these methods have not been systematically used in biomedical research (Deaton and Cartwright 2018, 17).

above discussion clarifies *how* they make that case. Observational studies make a good case when the variable that accounts for the distribution of subjects across comparison groups is a U more or less well-understood. To illustrate with a previous example, what researchers have to argue in order to make a good case, is that quarter of birth is uncorrelated with other unobserved covariates of income. This argument usually draws from empirical evidence—for example, showing a lack of correlation between quarter of birth and many observed covariates of income, performing so-called "over-identification tests" (French and Popovici 2011)—and theoretical reasons (e.g., the factors thought to affect quarter of birth are not thought to affect income).

Where, then, is the general suspicion coming from? The suspiciousness has, in many cases, little to do with dogmatism,[17] and much to do with the fact that many observational studies *fail* to make a good case for comparability. Due to scarcity of data, and thus the inability to

[17] For a clear expression of non-dogmatism from authors well-known for their endorsement of randomization, here are Angrist and Pischke: "[Edward Leamer] also argued that randomized experiments differ only in degree from nonexperimental evaluations of causal effects, the difference being the extent to which we can be confident that the causal variable of interest is independent of confounding factors. We couldn't agree more." (2010, 6) They then add: "Indeed, we would be the first to admit that a well-done observational study can be more credible and persuasive than a poorly executed randomized trial." (9)

control for covariates, many such studies provide few empirical or theoretical reasons to be confident in the assignment (see French and Popovici 2011). Nevertheless, this suspiciousness should (and does) decrease when the variable that accounts for the distribution of subjects across comparison groups is a U more or less well-understood. Indeed, observational studies that make a good case that groups are comparable are continuously published in top journals (e.g., Duflo 2004).

Regarding the second type of disagreement, the crucial point to note is the following. For some of the questions in which researchers are interested, background knowledge strongly suggests that a good case for causal inference is not forthcoming using observational data. This occurs when background knowledge strongly suggests the presence of imbalance in covariates that typically are unobservable. A classic example: economists' reluctance to attribute differences in income between participants and non-participants in job training programs to the programs even after controlling for education, age, and gender. This reluctance is *not* due to a blind fixation with randomization. Rather, economic theory suggests there are *reasons* why some took the program and others did not. Economists expect a systematic connection between participation and unobserved covariates even after having controlled for observable variables. First, people more eager to get a job (more "motivated") are likely to be over-represented among the participants. Moreover, according to economic theory, those who took the program are more likely to be those who believe they would benefit the most from taking it. If this expectation of a larger benefit is correct on average, or

if having this expectation is correlated with unobserved covariates (like ability, motivation, family networks, etc.), then researchers have further positive reasons for not attributing the whole difference to the program.

All said, it is background theory (e.g., about people's decision-making processes), and not an unwarranted belief in "the special power of randomization" (Worrall 2002, S319), which drives (in many cases) researchers' apparent fixation for randomized studies. That is why when background theory suggests the absence of a systematic connection, researchers (even well-known "randomistas") go on and draw causal inferences from good quality observational studies. This point was largely missed by Urbach and Worrall's discussion, and, with it, the opportunity to better conciliate research practice with sound epistemological principles.

7. Further Reasons for Randomizing.

Even if we were to grant Urbach and Worrall's premise that any U is as good as randomizing for causal inference, there are further reasons in favor of randomizing in (D).

7.1 Epistemic Reason: Rational Stability

One such reason lies in the stability of our rational beliefs. Randomizing fares better in this regard in two ways. First, there is a difference between rationally assigning a credence of ½ to the proposition that a coin will come out tails because (i) I have no reason to think that the

coin is biased versus (ii) I have good evidence that it is fair. The difference is not seen in the credence—either way is ½. It is seen in how I should rationally react to *further* evidence. Learning that the coin landed tails in each of a good number of consecutive tosses should have a greater impact on my credence in the case of (i) than in the case of (ii). That is, in (ii), my credence is more "resilient" (Skyrms 1977).

Something somewhat analogous seems to hold with respect to a U versus randomization. Neither way of assigning subjects gives me a reason to think that one group is better endowed in terms of covariates than the other. In the case of a U, this is because I *don't have* reasons to think the groups are unbalanced. As discussed above, not all U's are equal, and some will make me more rationally confident about baseline balance. But in real-life cases, none will make me as confident as randomization. In this sense, my (rational) causal attribution will be more resilient, and thus more stable, with randomization. In particular, further evidence questioning the causal attribution (e.g., a contradictory result observed afterwards) should make me react more if I used a U than if I randomized.

A second reason why randomization produces more stability is as follows. If we use a U, there is a kind of undercutting evidence that can be presented to us and that, if presented, will force us to withdraw the causal attribution: evidence that the variable used is systematically correlated with a significant covariate of *Y*. This is what happened with the case of quarter of birth—further evidence showed it to vary systematically with significant covariates. Because random numbers are not systematically correlated with significant covariates, no such

evidence can be provided if we randomize. Moreover, it is not only empirical evidence that can be presented to us and that may lead us to withdraw the causal attribution when using a U. Theoretical arguments might be developed later on, arguments which suggest that the U at stake is systematically correlated with covariates of *Y*. Again, this is not something that can occur if we randomize. We have found, then, a second way in which our causal attribution will be more stable if we randomize—not because of the resilience of our credences towards *rebutting* evidence (as before), but because of the difference in potential *undercutting* evidence (or theoretical reasons) regarding baseline balance.

Randomizing may produce more rational stability than using a U through the two mechanisms mentioned. Arguably, rational stability is important. When knowledge guides action, it is in the agent's interest that such guidance is more (rationally) stable, more robust to new evidence. The action referred to might well be intellectual (e.g., decisions about further research), or practical. When political institutions request scientific reports on controversial policy topics, what they are looking for is valuable information. This surely includes information that is trustworthy, in the sense that (i) the opinion roughly reflects the scientific consensus, but also (ii) in the sense that it is stable, that it is unlikely to change (unless surprising new evidence is collected). Why wouldn't scientists choose a method of assignment that makes their rational conclusions more stable?

7.2 Epistemic Reason: Rational Agreement.

Is there any further reason that a *community* of researchers might have for asking each researcher to randomize, which is, perhaps, not immediately a reason for the individual researcher? Take the case of our job-training researcher, and imagine she assigns subjects according to a variable (call it *V*). *V* is a U for that researcher. There might be another researcher within the same field that disagrees—*V* is not a U for her. This second researcher thinks that subjects for which *V*=1 (versus those for which *V*=0) systematically differ in relevant unobserved covariates, and thus that assigning subjects according to *V* makes condition (2') unlikely to hold. This could be because the second researcher just knows more about *V*, in which case it is a bad thing for science that the first researcher chose to use *V* instead of randomizing. But the disagreement can also be due to differences in rational beliefs the two researchers have about *V*'s possible relations to some covariates of *Y*. That is, not due to the second researcher *knowing* that *V* is systematically related to some covariates, but rationally believing so.[18] This rational disagreement, however, cannot occur if the first researcher randomized. Nobody has a reason to believe that the random variable used to assign subjects is systematically related to some covariates.

Since rational agreement and consensus among scientists seems highly valuable, we found a collective reason for randomizing: The community should ask the individual researcher to

[18] Disagreement of this kind is not a mere theoretical possibility, but rather a common feature of research communities. For examples in economics, see French and Popovici (2011).

randomize *even if* there is another variable that she considers to be a U, since some colleagues might not (rationally) agree. Randomizing reduces for the community, without cost, disagreement that comes from researchers' different takes on potential U's.[19]

7.3 Non-epistemic Reasons.

For completeness, I briefly discuss non-epistemic reasons for randomizing in (D). Part of Urbach's and Worrall's criticisms of randomized studies relate to well-known ethical and pragmatic considerations. As mentioned above, though valid and important, these criticisms do not bear on (D). In contrast, there are other ethical and pragmatic reasons—relevant in (D)—in favor of randomizing.

Pragmatically speaking, researchers will not always have potential U's in their database. They could collect them, but generating a random number is less costly. Some variables that researchers usually have that might potentially be considered U's are not straightforwardly unsuspicious. Names, for instance, can be correlated with socioeconomic status (in our example above). The time at which the subject arrived at the survey center can be correlated with punctuality—hardly a variable one wants unbalanced in a job training program evaluation. And so on. Thus, U is not always a feasible (not costlier) alternative.

[19] This argument draws from (Suppes 1982; La Caze 2013).

On the ethics side, the literature has discussed the moral value of using lotteries for political decisions for some decades. As Stone (2010) argues, standard arguments for using lotteries only show that random selection is as good as "picking"—selecting an option for no particular reason. For instance, if the alleged reason to select randomly is that all the options are equally good (or bad) according to the relevant criteria, then random selection provides no benefit over picking. However, there are cases in which we have a reason to prefer random selection over picking. These are cases where fairness is at stake. The contrasting position is presented here by Elster, in the context of a physician who has to allocate organ transplants in accordance with need and faces the question of how to allocate them once need has already been considered.

> To say that we might as well use a lottery is not to say, however, that a lottery is rationally or morally required. If there is no detectable, relevant difference among the candidates, all are equally worthy and hence it might appear that no wrong is done by using other methods of allocation. Thus it has been argued that one might as well select the most beautiful, the ugliest, the tallest (and presumably the shortest) people in the pool. (1989, 109)

Stone (2010, 154) correctly rejects that random allocation is not morally preferable here. Assigning according to beauty violates the impartiality requirement of justice. Justice is not exhausted by taking good reasons into account (need, in this case). It also requires avoiding *bad* reasons (beauty). This holds for interventional studies also—whenever a U embodies bad

moral reasons (such as when U=beauty), researchers have a moral reason to use randomization over a U.

A further moral reason for randomizing over using a U: it is preferable to choose a method that *subjects perceive as fairer*. Subjects might perceive randomization as fairer than entrusting the assignment to, say, "a person without any axe to grind." This might be because U embodies a morally bad reason, or merely because subjects distrust more assignments made with non-random variables. Indeed, this perceived-fairness factor is considered by practitioners[20]. This is why researchers sometimes do the randomized assignment publicly. In short, as long as the subjects perceive as fairer a random assignment, researchers have a moral reason to randomize in (D).

In sum, assigning by a U may be unjust or be perceived as unjust by the subjects relative to randomization. In any of these two cases, researchers have a moral reason to randomize. Granted, these considerations about justice and perceived fairness might not apply in *all* cases—some treatments in the social sciences are not significant enough to trigger justice concerns, and some researchers might happen to have available U's for which subjects have

[20] See David McKenzie's blog: https://blogs.worldbank.org/impactevaluations/should-we-require-balance-t-tests-baseline-observables-randomized-experiments. (Retrieved May 6th, 2019)

no fairness complaint. Nevertheless, it seems to me, these considerations would apply in many cases.

8. Conclusion.

Pace Urbach and Worrall's insistence that, after controlling for observed covariates, randomizing can do "no further epistemic good" (Worrall 2007, 463) and thus that "there's no cause to randomize," I have provided epistemic (and non-epistemic) reasons to prefer randomizing over its alternative, U, in interventional studies. Moreover, the greater reliability of randomization over U for achieving baseline balance provides a sound justification to the current practice of giving a positive value to randomization when *assessing* studies. This does not entail rejecting observational studies' capacity to achieve baseline balance, and thus to warrant causal inference. When assessing studies, evidence-appraisers need to judge how plausible the alternative hypothesis H* is, given what we know of the way the comparison groups were formed. When comparison groups are formed by a U about which we understand little (well), the alternative hypothesis H* is more (less) plausible, which decreases (increases) the quality of the causal evidence. And, when background knowledge strongly suggests there is a systematic connection between being in one comparison group and unobserved covariates of *Y*, ceteris paribus, observational studies provide little justification for causal attribution. Thus, not all observational studies are on a par with regards to baseline balance. This more nuanced and unified view on baseline balance

provides a better framework for developing evidence-appraisal tools that do not treat all non-randomized studies on a par.

References

Angrist, Joshua, and Alan Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106:979-1014.

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Banerjee, Abhijit, Esther Duflo, Gilles Postei-Vinay, and Tim Watts. 2010. "Long-run Health Impacts of Income Shocks: Wine and Phylloxera in Nineteenth-century France." *The Review of Economics and Statistics* 92:714-28.

Blackburn, McKinley, and David Neumark. 1995. "Are OLS Estimates of the Return to Schooling Biased Downward? Another Look." *The Review of Economics and Statistics* 77:217-30.

Brownstein, Michael, and Jennifer Saul. 2016. *Implicit Bias and Philosophy*. Oxford: Oxford University Press.

Bruhn, Miriam and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1:200-32.

Buckles, Kasey, and Daniel Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *The Review of Economics and Statistics* 95:711-24.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210:2-21.

Duflo, Esther. 2004. "The Medium Run Effects of Educational Expansion: Evidence from a Large School Construction Program in Indonesia." *Journal of Development Economics* 74:163-97.

Duflo, Esther, and Abhijit Banerjee, eds. 2017. *Handbook of Field Experiments.* Vol. 1. Amsterdam: North Holland.

Elster, Jon. 1989. *Solomonic Judgements: Studies in the Limitations of Rationality*. Cambridge: Cambridge University Press.

French, Michael, and Ioana Popovici. 2011. "That Instrument is Lousy! In Search of Agreement when using Instrumental Variables Estimation in Substance use Research." *Health Economics* 20:127-46.

Fuller, Jonathan. 2019. "The Confounding Question of Confounding Causes in Randomized Trials." *The British Journal for the Philosophy of Science* 70:901-26.

Howson, Colin, and Peter Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. 3rd ed. Chicago: Open Court.

La Caze, Adam. 2013. "Why Randomized Interventional Studies." *The Journal of Medicine and Philosophy* 38:352-68.

Senn, Stephen. 2013. “Seven Myths of Randomisation in Clinical Trials.” *Statistics in Medicine* 32:1439–50.

Skyrms, Brian. 1977. “Resiliency, Propensities, and Causal Necessity.” *The Journal of Philosophy* 74:704-13.

Sober, Elliott. 2009. “Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads.” *Philosophical Studies* 143:63-90.

Steele, Katie. 2013. “Persistent Experimenters, Stopping Rules, and Statistical Inference.” *Erkenntnis* 78:937-61.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Stone, Peter. 2010. “Three Arguments for Lotteries.” *Social Science Information* 49:147-63.

Suppes, Patrick. 1982. “Arguments for Randomizing.” In *PSA 1982,* vol. 2, ed. Peter Asquith and Thomas Nickle, 464-75. East Lansing: Philosophy of Science Association.

Urbach, Peter. 1985. “Randomization and the Design of Experiments.” *Philosophy of Science* 52:256-73.

Worrall, John. 2002. “*What* Evidence in Evidence-Based Medicine?” *Philosophy of Science* (Proceedings) 69:S316-S330.

---. 2007. “Why There’s No Cause to Randomize.” *The British Journal for the Philosophy of Science* 58:451-88.