Overcoming Frege's Curse: Heuristic Reasoning as the Basis for Teaching Philosophy of Science to Scientists

Till Grüne-Yanoff, KTH Royal Institute of Technology Stockholm
Teknikringen 76, 10044 Stockholm, Sweden, gryne@kth.se

ABSTRACT A lot of philosophy taught to science students consists of scientific methodology. But many philosophy of science textbooks have a fraught relationship with methodology, presenting it either a system of universal principles or entirely permeated by contingent factors not subject to normative assessment. In this paper, I argue for an alternative, *heuristic* perspective for teaching methodology: as fallible, purpose- and context-dependent, subject to cost-effectiveness considerations and systematically biased, but nevertheless subject to normative assessment. My pedagogical conclusion from this perspective is that philosophers should aim to teach science students heuristic reasoning: strategies of normative method choice appraisal that are sensitive to purposes, contexts, biases and cost-effectiveness considerations; and that we should do so by teaching them exemplars of such reasoning. I illustrate this proposal at the hand of three such exemplars, showing how they help students to appreciate the heuristic nature of both methods and methodology, and to normatively assess method choice in such circumstances.

Keywords: Methodology, heuristics, method choice, justification, teaching philosophy of science, philosophy of science for scientists

> *There is, indeed, very little fixed method; but, with all due respect to those who suffer from the curse of Frege, there is objectivity enough. (Levi 1980, 430)*

1. *Introduction*

Science students expect normative guidance from methodology. But philosophy of science, at least as represented by the majority of textbooks used for this purpose, too often are swayed by Frege's Curse: they assume that a normative methodology either consist of a fixed system of universal principles, or that scientific method choice is influenced by psychological and other contingent factors and thus not subject to normative control at all. This makes for uneasy company in the classroom, where philosophers tasked with instructing science students often teach highly abstract methodological principles without clear bearings on the actual scientific practices students engage in, or alternatively resort to teaching history of science without any normative input.

In this paper, I offer a way out of this cursed dichotomy. Based on the idea that scientific methods are heuristic in nature (Bechtel & Richardson 1993; Wimsatt 2007), I argue that teaching methodology should focus on the justification of method choice as fallible, purpose- and context-dependent, subject to cost-effectiveness considerations and systematically biased. In other words, methodology itself is heuristic in nature (Hey 2016; Grüne-Yanoff 2021b). This precludes methodology being a fixed system of universal principles, but it does not prevent methodology from being normative. There are good and bad heuristics for particular contexts and purposes. Yet instead of providing a more localized recipe book, I argue, philosophers should teach methodology as *heuristic reasoning*, by training students at the hand of exemplar method choices. These exemplars illustrate that there are multiple methods for a given purpose, that the fallibility and bias of each can be amplified or curtailed through contextual features, and that inadvertent or intentional mismatch between method and context can lead to method misuse. They further show that purposes, contextual factors and method success conditions are often vague and uncertain, so that one must revert to rules of thumb for choosing between them. Through these exemplars, students learn to justify method choice, without taking recourse to a system of universal principles.

The paper is structured as follows. Section 2 sketches Frege's curse, and how a heuristic methodology can avoid it. Section 3 draws some general pedagogical conclusion from the preceding argument, and then illustrates teaching methodology as heuristic reasoning at the hand of three exemplars. Section 4 concludes.

2. Frege's Curse

Scientific methodology is about the justification of method choice in science.[1] Like with any human choice, the subject of methodology is thus closely connected to questions about human cognitive abilities and their limitations. For this reason, many traditional philosophers of science have had a fraught relationship with methodology. Following Frege's rejection of psychologism, they assume that "the objectivity of scientific inquiry is made to stand or fall with the existence of a fairly powerful and fixed system of principles applicable to all agents on all occasions" (Levi 1980, 427). This assumption is shared by those defending a universal methodology, for example Carnap or Popper, as well as those rejecting it, for example Feyerabend. Only if there was a universal normative methodology could the objectivity of scientific inquiry be upheld; if there wasn't, then one must concede that scientific inquiry is exposed to psychologically, sociologically or historically contingent factors and thus not subject to serious critical control. The belief in this purported dichotomy Levi called the *curse of Frege*.

Granted, philosophy of science might have been gradually freeing itself from the curse. Since the 1990s, there is a palpable shift from generalist philosophy of science to a philosophy of the special sciences, and there are successful organizational efforts, e.g. in the wonderful *Society for the Philosophy of Science in Practice* (SPSP), to refocus on scientific practices as the primary locus of philosophical analysis. However, there are some caveats. First, a lot of philosophy of the special sciences is *not* about methodology, but about conceptual or foundational issues. Second, amongst those who focus on practices in the special sciences, many avoid an explicitly normative analysis. There are of course important exceptions - e.g. Bechtel & Richardson (1993), Wimsatt (2007) - who appreciate the heuristic nature of scientific methods and draw explicit methodological conclusions from it.[2] In any case, these views have so far not percolated into philosophy of science textbooks, which most often carry on presenting the competing universalist positions (see the review of recent textbooks in Grüne-Yanoff 2014); hence I suspect that it hasn't percolated into teaching practices either. On the question of methodology, philosophers of science thus divide into traditionalists beset by Frege's curse (either by defending a universal method or by rejecting a normative methodology altogether), and particularists who try to free themselves from it. When it comes

[1] In this paper I focus on the challenges of teaching methodology. But I do not deny that science students can benefit from philosophers teaching other subjects as well (cf. Laplane et al. 2019).

[2] My impression from attending many SPSP conferences, however, is that many eschew normative conclusions, instead constraining themselves to *descriptions* of scientific practices. That strikes me as them being stuck on the other extreme of the cursed dichotomy.

to teaching philosophy of science to science students, it seems the traditionalists still have the edge.

I notice these conflicting positions also amongst my students.[3] On the one hand the "universalists", who take the textbook version of science at face value and assume that sufficient attention to the facts has and will fuel the steady progress of science. To them, methodology is merely a technical question of acquiring the skills needed to handle the ever-more sophisticated machineries for recording and process the facts.[4] To give but two examples, such students tend to consider statistics as providing universal algorithms and embrace the evidential hierarchy. From such a perspective, a discussion of method choice seems often redundant. Instead, the difficulty lies in mastering the algorithm or the procedure that produces "best" evidence.

On the other hand the "particularists": They sense that the tools and practices of scientific inquiry are too multifaced, have too many purposes and are too uncertain in their application conditions and their successes to be controlled by a universal norm. They seem to grasp that many practices they acquire in their studies are conventional. Yet they also learn about the admired mavericks who advanced the field with their unconventional approaches. They might begin to realize how much depends on funding, hierarchies and the review lottery; and if they look at other disciplines, they see that core features like evidence standards, experimental practices and modelling strategies often look entirely different from those of their own discipline.

Neither the universalists nor the particularists benefit from philosophy in the thrall of Frege's curse. The universalists think they already have their technological solution to any methodological problem, and are unlikely to listen to the technologically unskilled philosophers for advice. The particularists intuitively understand that there are no universal recipes for doing science, and they have no need for philosophers lecturing on them.[5]

Nor do either of these groups have much patience with philosophers holding forth on the other pole of the cursed dichotomy: that science, because it is infused with psychologically, sociologically or historically contingent factors, is not subject to critical control. This is obviously the case for the universalists. But even particularists, in my experience, are attuned to the need of such normative assessment. They are aware that to challenge conventional and dominant theories and practices in their field, they need to produce arguments that stand critical scrutiny; and they know that in order to collaborate interdisciplinarily, practices and standards independent of the historical and sociological contingencies of the involved disciplines must be agreed. Therefore they seek normative critical guidelines, albeit not in the form of a universal methodology.

This sketch of what scientific methodology is or should be raises the question: How to lift Frege's Curse? How to teach science students to justify, assess and critique method choice

[3] This is an intuition formed through teaching more than 1000 MSc and PhD science students per year over the last 10 years. I do not have empirical data to support this intuition, nor do I know of any relevant study that could support it.

[4] Such students are often supported in their views by (senior) scientists defending universalist understanding of methodology (the many lip-services to falsification come to mind here, as well as the references to 'the' scientific method, for instance in debates about replication).

[5] Smith (2017) and Grüne-Yanoff (2014) offer additional ideas what might make science students resistant to philosophy courses.

without falling back on the chimera of a universal methodology? My proposal is to convey to students the heuristic nature not only of methods but also methodology, and teach them how to reason heuristically for the use of specific heuristic methods suitable for their research purposes and -contexts.

The notion of heuristic has been widely used in cognitive science and decision theory, starting with Simon in the 1950s, and gaining wider recognition with the work of Tversky, Kahneman and Gigerenzer, amongst others (for an overview, see Chow 2015). In the first place, the notion has been used for explanatory purposes. But importantly, its normative assessment has been increasingly discussed: there are good and bad heuristics for specific purposes and contexts; and sometimes, a simple heuristic can outperform a more substantial universal algorithm (Gigerenzer & Sturm 2012; Arló-Costa & Pedersen 2013). For scienctific methodology, it has been Feyerabend and then in particular Wimsatt who stressed the importance of heuristics.[6] According to Wimsatt, many scientific methods are heuristics in that they (i) are fallible, (ii) are motivated by cost-effectiveness considerations, (iii) are systematically biased (and show this bias in some contexts but not in others) and (iv) transform the original problem into a manageable but non-identical one (Wimsatt 2007, 76-77).

Methodology thus studies the justification of heuristic method choices. It must consequently consider purposes and contextual factors - for example: Is the context of use similar to the heuristic's past successes? Does the context provide resources to make this heuristic effective? Does the context fuel or starve the heuristic's biases? Answers to all these questions will typically be highly uncertain, and where uncertainty can be resolved, it will often require highly local, experiential or tacit knowledge. In order to deal with these uncertainties, methodology itself must resort to heuristics (Hey 2016; Grüne-Yanoff 2021b). There still is a norm for each specific problem and context. But because context and purpose conditions are now so fine-grained, it is often difficult to precisely determine them and to check whether a specific research context satisfies them. This uncertainty and the resulting interpretation problems often force scientists to revert to simple, fallible and biased meta-heuristic rules.

This contextualization and localization have important implications for teaching approaches. Even though some philosophers have delved deeply into some specific purposes and contexts that scientists face, philosophers can hardly be expected to know these purposes and contexts as well as a scientist trained in a specific discipline. Nor do they have the scientists' practical skills and know-how. Finally, the knowledge and skill requirements would become even more daunting when philosophers face an interdisciplinary audience, as they often do. To teach students a large number of concrete recipes of the kind "if you aim at X and you face conditions Y, do Z" would be pointless, given philosophers' comparative ignorance of the relevant Xs and Ys (here the universalists amongst my students might have a point). What I propose instead is to train students in *heuristic reasoning*, by systematically examining *exemplars* of heuristic method choice in different domains of scientific practice.

[6] Interestingly, Feyerabend exempted such heuristics from his crusade against method: "Andererseits ist gar nichts gegen Faustregeln einzuwenden....sie fordern [den Forscher] auf, kindliche Dinge, wie die Logik, hinter sich zu lassen und auch erkenntnistheoretische Regeln nie zu ernst zu nehmen" (Feyerabend 1986, 377). By this, he also seemed to have exempted them from normative control altogether. By thus implicitly equating normative methodology with universal principles, Feyerabend reveals himself to still suffer from Frege's curse.

3. *Teaching Heuristic Reasoning by Exemplar*

While there are a number of proposals for a normative theory of heuristic reasoning, they remain fragmented, incomplete and controversial. Furthermore, as I argued above, any such theory useful for scientists would likely to be so domain-specific that it could not be fruitfully taught to an interdisciplinary audience; and the context-specificity of the required knowledge would make philosophers not the ideal teachers anyway. Instead, I teach heuristic reasoning by exemplar.

This proceeds as follows. I start by identifying a popular method and identify potential alternative methods that could be used in its stead. This makes clear that scientists must make a choice, which in turn requires a justification. Given the heuristic nature of the method, this justification depends both on the research purpose and various contextual features. What purposes does the method serve? and: are there purposes that this method cannot serve? are my questions here. Then, for achievable purposes, what are the contextual conditions that further or prevent the successful attainment of the goal with this method? And what features might indicate the presence of such success or failure conditions?

In the following, I present my teaching approach at the hand of discussing three methods: Fisherian significance testing, experimental randomization and massive simulation modelling. Each case speaks against universalist intuitions: the methods discussed cannot be seen as general algorithms. Instead, each case shows the need for normative assessment: to justify the method choice in a simple, fallible and biased – i.e. heuristic – way. Each also exemplifies ways how to heuristically reason for such a justification: by identifying relevant success and failure conditions to the best of one limited knowledge and cognitive abilities. I chose these cases because they represent different domains of research practice, and because they tend to work for an interdisciplinary student body (ranging from students of the natural sciences, engineering to the social sciences). They also put the spotlight on slightly different aspects of heuristic approaches. In the first two cases, the problem is to determine whether a method offers a desired function or not, while in the third case, the problem is to find an acceptable trade-off between different desiderata. Furthermore, in the first case, the problem is at least partly solved by being more specific about the goal in using a certain method; while in the second case, a better understanding is required of what the method actually does. Importantly, these cases are exemplars: they do not teach methodological recipes for specific purposes and contexts. Instead, they help students understand and apply procedures of justifying heuristic method choice.

3.1 Significance testing

In *p*-value significance testing, one chooses a model (the null hypothesis) and a threshold value, called the significance level of the test, traditionally 5% or 1%. One then calculates the *p*-value, the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. If the *p*-value is less than the chosen significance level, that suggests that the observed data is sufficiently inconsistent with the null hypothesis and that the null hypothesis may be rejected.

P-value significance testing is a statistical method widely used in the sciences. But it also has attracted a fair amount of controversy, with some authors claiming that scientists should stop using it (Amrhein et al. 2019) and some journals banning its use in submitted manuscripts (Trafimow & Marks 2015). What are the reasons of this controversy, and what can science students learn from it for their on methodological choices?

That statistics promulgates lies is a hackneyed platitude. More interesting is the claim that statistics *can* be misused, as the popular textbook *How to Lie with Statistics* (Huff 1954) tongue-in-cheek explains. *P*-value significance testing methods can help ill-intentioned scientists to make false claims about how well their hypotheses are supported, as the recent replication crisis in psychology and medicine shows (Open Science Collaboration 2015; Begley & Ellis 2012). The fact that statistics can be misused shows that it does not provide a universal algorithm that for each problem assigns a unique method and yields an unambiguous result. Instead, statistics offers a *toolbox* - a collection of methods that can be put to different uses, good and bad. Choosing a proper tool for a given problem requires *statistical thinking* (Gigerenzer 2004). Users must argue why employing a certain statistical method is justified given their purpose and the application context.

If statistics can be misused, one might wonder what its legitimate uses are. To drive this point home, I ask students: Why use statistics at all?[7] After all, they already encountered explications of testing procedures that did just fine without reference to statistics, like Popper's falsification and Hempel's logic of confirmation. For good measure, I also introduce them to Austin Bradford Hill, who reported that in many of his studies "[t]he evidence was so clear cut … that no [statistical] tests could really contribute anything of value to the argument. So why use them?" (Hill 1965, 299). Evidently, statistical methods are not *necessary* for making scientific inferences – although they might sometimes be useful.

But what are they useful for? I offer three answers: first, stochastic hypotheses only imply distributions, and thus can be tested only with statistical descriptions of the data; second, probabilities help quantify the degree of confirmation that a certain piece of evidence confers on hypothesis; and third, statistics helps to quantify the probability of error when one infers a systematic pattern from noisy data. While these alternatives are not meant to be comprehensive, they show the heterogeneity of purposes for which statistical methods are used. Furthermore, not every method can satisfy every purpose: p-value significance testing, for example, cannot be used to quantify the degree of confirmation that a certain piece of evidence confers on hypotheses.

Applying a method to a purpose it cannot satisfy constitutes a form of statistics misuse. Such misuse is surprisingly common, typically based on a misunderstanding of the method in question. The ASA's statement on p-values corrects a number of such misunderstandings, for example: that significance testing proves a tested hypothesis to be false; or that one can infer the probability of hypotheses from significance tests (Wasserstein & Lazar 2016). Statistical thinking thus requires not only an answer to "why use statistics at all for this problem?", but also "which statistical method(s) can satisfy my goal in investigating this problem?".

Differentiating purposes, however, is not the only demand of statistical thinking. The purpose of Fisherian significance testing is to quantify the probability of erroneously rejecting H_0. The p-value is the probability of obtaining observations like the ones made, or more extreme ones if the H_0 were true. By quantifying this error, and thus making it comparable to accepted standard thresholds, significance testing helps ensure that hypotheses are subjected to a *severe test* – i.e. one that maximally confronts a claim's potential flaws with the available

[7] Although that is not the main focus of this paper, I hope it becomes clear that my teaching approach aims to maximally engage the students through having them answer questions, discuss with and present to their peers and apply novel concepts and arguments by analyzing scientific texts or by making explicit the method choices in their own MSc or PhD thesis. For more on the 'how' of such teaching, see Grüne-Yanoff (2014).

data (Mayo 2018). Most of the researchers who produced non-reproducible studies presumably knew that this was the purpose of significance testing and used it in order to show that they performed a severe test. They thus used the testing procedure for the right purpose, but they used it – either inadvertently or intentionally – under conditions that resulted in the procedure not satisfying this purpose. Instead of the procedure satisfying the intended purpose, their use of the procedure only made it *appear* that it satisfied the procedure, while in fact it did not.

This insight allows us to identify a second kind of statistics misuse: to use the many decisions that researchers need to make in performing a statistical analysis as means of generating the mere appearance of a severe test. These include strategic hypothesis formulation, biased selection and operationalizion of independent variables, *ad hoc* discarding of participants, failing to specify the sampling plan, non-random assignment, correcting data during data collection in a non-blinded manner, intermediate significance testing, *ad hoc* data cleaning, and many more (Wicherts et al. 2016).

Didactically, I ask students to participate in a *p-hacking competition*, making use of these 'tricks' with the goal is to adjust the testing procedure based on a given data set in such a way that the p-value ends up below the 1% threshold, thus giving the appearance of severity. Students are asked to test a simple hypothesis (that a coin is fair) with a simulated data set (tossing a coin *n* times, where – unknown to the students - the probability of the coin showing heads at each toss is indeed 50%). Students can manipulate a number of test features: the formulation of the null hypothesis, the number *n* of tosses, whether it is a one or a two-sided test, whether a certain part of the sequence should be discarded as an outlier and whether to use a normal or non-normal distribution of errors. These manipulations can be performed throughout the exercise, while students can continuously observe the resulting p-value. Their task is to draw their conclusion (either that the coin is fair or that it is not) at a significance level of 0.01. The ease with which students can generate such support for either conclusion gives a good impression of the pitfalls in using this method. I then reveal the true data generation mechanism and review the test procedure with the students, discussing various safeguards (e.g. preregistration, sampling plan, data-cleaning rules) against such misuses.

To conclude, significance testing is a heuristic method that can successfully serve particular purposes, under particular conditions. It is a good exemplar for teaching heuristic methodology to science students, because it is widely applied in science practices across disciplines, and most students have basic technical skills in statistics. Universalist intuitions run high here - many students seem to believe that statistics must always be applied to craft an empirical argument, and that it can be applied like an algorithm. Heuristic methodology, to the contrary, shows that statistics is fallible even when used for the right purpose, under the right conditions, and that it is often hard to determine both purpose & conditions precisely for a given problem. Nevertheless, given that success is likely in certain conditions, to call for a ban of this method is misplaced. Significance testing furthermore is a good exemplar, because scientists have begun to actively participate in methodological debates relating to it, thus deflating the apperance of any fundamental opposition between philosophy and science. What discussing this exemplar shows it that scientists need to acquire statistical thinking competences: the ability to formulate clear goals pursued with significance testing (severe test) and the conditions necessary to ensure that provides a severe test.

3.2 Randomization

An experiment is called "randomized" when the assignment of experimental subjects or objects to treatment and control groups proceeds with the help of a randomization device. Today, randomization is often considered to be the property that distinguishes highest-quality experimental evidence from lower quality one (e.g. USPSTF 1996). However, a growing number of authors has voiced criticism against attributing such general quality improving capacities to randomization (e.g. Rawlins 2008 for medicine, Ravallion 2020 for economics, and Deaton & Cartwright 2018 for a cross-disciplinary perspective). Consequently, I encourage students to explore the reasons for this high regard for randomization.

The most commonly cited reason for randomizing experiments is that it supposedly helps control the influence of background conditions. Such control is of course important to make valid inferences from experimental observations: when comparing the effect of an intervention with a control group, one must ensure that the observed difference is not affected by those background conditions. But 'control' here is ambiguous, referring to multiple different practices and results. One controlling practice is the elimination of a background factor, leading to a state where the background factors have no influence at all on the variable of interest - for example falling experiments in a vacuum, or electrical experiments in a Faraday cage. Another practice is the homogenization of background factors, leading to a state where the background factors have the same influence on the variable of interest in both treatment and control groups. This can be done, for example, by manipulating the background variables in laboratory settings, or by matching members of the treatment group with similar members of the control. Either way, these practices are demanding in multiple ways: both epistemically, as one need to know what the relevant factors to control are, as well as productively, as one must have the abilities to eliminate, manipulate or match these factors. Clearly, this knowledge or these abilities are sometimes unavailable.

Randomization does not generate any of the above states: it does not eliminate or homogenize background factors. Instead, because of its random assignment, it might generate equal distributions of background factors in both treatment and control group, so that the *expectation* of difference in their background factors - calculated over a large number of trials - is zero. This is called *perfect balance* (Schulz 1996). Perfect balance ensures that the difference between the mean effects observed in treatment and control group, respectively, is exactly equal to the average of the treatment effects among the treated. In this sense, randomized experiment can yield an *unbiased* estimate. Randomization thus provides a version of control that can support valid inference of the average treatment effect from an experiment, if they satisfy the condition for perfect balance. But that is a big if. It needs to be recognized that randomization is a particular version of control, and that it is subject to specific success conditions. Thus, researchers must choose their method of control, considering at least four features of randomization.

First, for any one trial, randomization will most likely not generate perfectly balance. Instead, random assignment is likely to over-represent a known relevant background factor in one arm over the other. Inferring that a difference between the means of the two groups is caused by the treatment would be a mistake, as the imbalance in the arms likely contributed as well. Mere randomization therefore is not sufficient for balance, and thus not a full-fledged substitute of other means of control that do guarantee unbiasedness.

Second, randomization is more likely to achieve balance when the sample size is large. As the sample size tends to infinity, the means of background factors in treatment and control groups converge. Yet of course, sample size is never infinite, so the question is how big finite samples need to be in order to ensure balance. This in turn depends on how many factors are known to be relevant: the more factors that need to be included, the larger the sample must be. It is obvious that for many experiments, the number of relevant factors is too large to hope to secure balance through sample size.

Third, there are strategies to help randomization achieve balance even with smaller sample sizes. At the very least, one must check for imbalances amongst the known factors after randomization. Treatment and control group factor distributions should be considered with respect to their known influence and the size of any imbalance that occurred (Altman & Doré 1990). If imbalances are too large, one can redo the randomized assignment until the imbalance is eliminated (Morgan & Rubin 2012). Alternatively, one can try to *stratify* the sample. By categorizing the sample into relevant background conditions and randomly assigning equal numbers from each category to the experimental arms, one can ensure balance of all the categorized factors. However, stratification is strictly limited by the number of factors and their possible realizations. The number of strata rises exponentially in both, so that it quickly outruns even large sample sizes. Finally, another alternative is to adjust for imbalanced factors by running a covariance analysis, checking for possible interactions between treatment and explanatory variables. Such estimates are biased in final samples (Freedman 2008). While such a strategy might be reasonable, it compromises on unbiasedness, one of the purported promises of randomization.

Fourth, it is sometimes claimed that randomization also can control unobserved or unknown factors (compare e.g. the quotes in Deaton & Cartwright 2018, 5). If this were true, it would be an important positive difference to elimination or homogenization methods of control, which cannot achieve this. But it isn't true. Mere randomization does not ensure balance, thus it does not ensure balance of unknown factors in particular. Furthermore, none of the mentioned strategies to achieve balance can be applied to unknown factors. Adjusting the sample size requires knowledge of the number of relevant variables. Post-randomization check requires knowledge of the identity of relevant variables. Finally, what isn't known cannot be stratified. Consequently, randomization does not help with directly controlling unknown factors.

Further arguments for randomization include that it helps eliminate bias in the control/treatment assignments, and that it is a useful instrument to implement blinding. However, randomization is not necessary for either of these results, nor is it sufficient for blinding.

Didactically, I ask students to design experimental assignments and check for balance between treatment groups. The (simulated) scenario involves testing a drug on a random sample of participants, where 15 known and 3 unknown background factors mediate the drug’s effect. Students must choose the sample size and must decide how to assign participants to control and treatment groups. Manual assignment is an option, as is randomization and stratified randomization. The drug has a medium effect size, but without checking for post-assignment balance (via visual comparison of pie charts), and adjusting the assignment procedure, simulated experimental results will likely not exhibit any discernable difference between control and treatment group. After the exercise, I reveal the true effect size and review reliable assignment procedures.

Randomization thus is a particular control strategy with its own advantages and disadvantages, that play out in the specific conditions of the experimental study at hand. One of these conditions is knowledge: if one doesn't know enough to implement the other methods of control, randomization is a viable alternative. But if one has this knowledge, other methods might yield better results (Savage 1962; Ziliak 2014). Further conditions are sample size, and whether one can repeat the assignment process.

Researchers thus face a genuine choice whether to randomize or not. This choice should be informed by the above-discussed conditions and whether they are satisfied in the given problem. Because these conditions in practice are often uncertain, researchers must make this choice relying on rules of thumb. This makes randomization is a good exemplar for teaching heuristic methodology to science students. It is often presented as an unquestioned methodological ideal for all experiment designs. The basic idea is easily explained, and the strong claims associated with randomization are *prima facie* plausible. However, it is a simple case where these *prima facie* impressions (in particular regarding balance and control of unknown factors) turn out to be unambiguously wrong, and can easily be shown to be so. Finally, it is an opportunity to showcase how a method might serve multiple objectives; that each of these objectives are reached only under specific conditions; and that some of these goals might be irrelevant for a researcher's specific purpose.

3.3 Massive Simulation Modelling

Modelling studies investigate a representation of a target instead of the target itself. Agent-based simulations allow the construction and manipulation of detail-rich models of agents and their interaction. Their amount of detail, and the complexity of manipulations increases with the ever-growing computational capacities of the newest machines. This raises the question whether computational limitations should be the *only* constraint on the detail and complexity of agent-based models, or whether there are good reasons to limit one's models for other reasons. I will call models constrained only by current computational capacities *Massive Simulation Models* (MSMs). Models whose detail is also constrained by other considerations (of simplicity, of transparency, etc.) I call *Abstract Simulation Models* (ASMs). To make this discussion more manageable, I focus here on simulation models for policy purposes in epidemiology and economics (Grüne-Yanoff 2021a).

Based on the richness in realistic detail, MSMs are sometimes claimed to offer a highly accurate picture of their targets. Their structure is designed to allow a mapping from the model to the target without taking recourse to mediating models. The Eubank et al. (2004) model, for example, is introduced as a direct representation of the city of Portland. ASMs, in contrast, can hardly ever claim to represent a real system directly — their level of detail is not sufficient. At best, they are able to represent *stylized facts* about a system, which have been prepared through an abstraction or idealization procedure from the real system. The Burke et al. (2006) model, for example, explicitly claims to represent an “artificial city” that shares some properties with real cities, but is different otherwise (Burke et al. 2006, 1142).

For a policymaker, MSMs often seem the *prima facie* preferable choice, for two reasons. First, the more detail is represented, the better a model user can assess the model - and if the represented detail is accurate, the more confidence she might have in the model (Dawid and Fagiolo 2008, 354). Second, if the model is a sufficiently close and complete representation of the real system, the policymaker might take it as a “virtual universe” in which the effects of interventions simulated in the model, are taken as accurate forecasts of the results of such interventions in the real system (Farmer and Foley 2009, 686). Against these intuitions, I seek to convince students that *both* are viable modelling methods, and that for some purposes and contexts, MSMs indeed are superior.

First, MSMs have higher number of free parameters, in comparisons to ASMs. As is well known in the model-selection literature, models with more free parameters have a larger *potential* to fit the target well; but the larger number of free parameters often yields an *actually* lower fit than the one achieved by a model with fewer parameters. To see this, distinguish two steps in the process of fitting a model to data. The first step consists in selecting a model — i.e., in specifying the number of parameters. Here, increasing the number of parameters indeed increases the model's *potential* to accurately represent the target. The second step consists in calibrating or estimating the parameters based on a data *sample* drawn from the population. Increasing the number of parameters increases the model's fit to the sample — but this is not the ultimate goal. Rather, increasing the model's fit *to the target* is. Fitting the model "too closely" (i.e., by including too many parameters) to the sample will pick up on the inevitable random error in the sample, and thus leads to an increase in the divergence between model and target. As various studies have shown, if the sample size is large, adding more parameters above a certain threshold will not substantially increase fit to target; if sample size is medium or small, adding more parameters even decreases fit to target (Zucchini 2000; Gigerenzer and Brighton 2009). Because MSMs have higher number of free parameters than ASMs, there are more prone to this source of error.

Another problem is MSMs' higher number and higher complexity of represented mechanisms, in comparisons to ASMs. The more complex a model, the more subcomponents it has. Furthermore, when running a simulation on a complex model, these model components are run together and in parallel. But they do not all independently contribute to the model result. Rather, the components, in the course of a simulation, often exchange results of intermediary calculations among one another — so that the contribution of each component to the model result is in turn influenced by all those components that it interacted with. The resulting interactivity between mechanisms, dubbed "fuzzy modularity" (Lenhard and Winsberg 2010), prevents the separation of mechanisms and their contributions to the variables of interest. Yet without knowing how individual mechanisms contribute, epistemic goals like deign or explanation are hard to satisfy. The designer needs to know *where* to intervene, and an explanation requires to identify the specific difference-makers of the *explanandum*. Because of MSMs' higher number and higher complexity of represented mechanisms, they are more likely to amplify each other or cancel each other out, and therefore less likely to further goals like explanation or design.

Didactically, to illustrate the problem of overfitting, I ask students to fit a model curve to a two-variable data set, generated by an underlying true process confounded by some error. Students can choose between five functions of increasing polynomial degree; the higher the polynomial, the better the fit. However, if they choose functions with degree 4 or 5, despite better data fit, the model will be a worse representation of the true process than functions of lower polynomial degree. This will become obvious after I reveal the true process and the error component. I then review strategies to avoid overfitting.

Model choice for agent-based simulation is a good exemplar for teaching heuristic methodology to science students, because the choice appears in many scientific disciplines, and it is easily illustrated with interesting examples. Each of the different choice options, here represented in the simple binary distinction MSM vs. ASM, have their specific advantages and disadvantages. In particular, it is not true that MSM are generally superior to ASM; instead, ASM sometimes can do the job as well as MSM, and often conditions are such that they can do the job better because they are not as prone to specific errors as ASM. Whether any particular investigation demands these conditions, however, is often highly uncertain. This exemplar thus showcases how competent method choice requires analyzing the objectives of one's own research project, and deciding which advantages and disadvantages of the respective options are the most important ones.

4. Conclusion

My main goal in teaching philosophy of science to science students is to train them in heuristic reasoning. This includes appreciating the heuristic nature of scientific methods and the particular kind of justification for the associated method choice. The core competence that this training aims to create is the ability to ask the right questions when crafting such a justification – both concerning the research purpose and the contextual success and failure conditions of the respective heuristics, adapted to the limited knowledge about the actual research context. Teaching by exemplar conveys the basics of such reasoning with respect to different research practice domains, without having to require or appeal to highly specific contextual knowledge. It shows that there is a normative question to be answered about each method choice, and illustrates how one can reason towards such a justification, while acknowledging that for most actual method choices that students will face, *they* will have to perform this evaluation, based on specific knowledge of purpose and context. My approach thus allows teaching normative method choice appraisal, without taking recourse to a universal system of principle, and thus without submitting to Frege's curse.

Declarations

I have no conflicts of interests or competing interests to declare.

REFERENCES

Amrhein, V., S. Greenland, and B. McShane. 2019. Retire statistical significance. *Nature* 567: 305-307.

Altman D.G. and C.J. Doré. 1990. Randomisation and baseline comparisons in clinical trials. *Lancet* 335:149-53.

Arló-Costa, H. and A.P. Pedersen. 2013. Fast and frugal heuristics: rationality and the limits of naturalism. *Synthese* 190(5): 831-850.

Bechtel, W. and R.C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton, NJ: Princeton University Press.

Begley, C.G. and L.M. Ellis. 2012. Raise standards for preclinical cancer research. *Nature* 483(7391): 531-533.

Burke D.S., J.M. Epstein, D.A. Cummings, J.I. Parker, K.C. Cline, R.M. Singa and S. Chakravarty. 200). Individual-Based Computational Modeling Of Smallpox Epidemic Control Strategies. *Academic Emergency Medicine* 13(11): 1142-9.

Chow, S. J. (2015). Many meanings of 'heuristic'. *The British Journal for the Philosophy of Science* 66(4): 977-1016.

Dawid, H. and G. Fagolio. 2008. Editorial. *Journal of Economic Behaviour & Organization* 67: 351-354.

Deaton, A. and N. Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210: 2-21.

Eubank, S., H. Guclu, V.S.A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczcai and N. Wang. 2004. Modelling Disease Outbreaks In Realistic Urban Social Networks. *Nature* 429: 180-184.

Farmer, J. D. and D. Foley. 2009. The economy needs agent-based modelling. *Nature* 460(7256): 685-686.

Feyerabend, P. K. 1986. *Wider den Methodenzwang*, Frankfurt a.M.: Suhrkamp.

Freedman, D. A. 2008. Randomization does not justify logistic regression. *Statistical Science* 23(2): 237-249.

Gigerenzer, G. 2004. Mindless statistics. *The Journal of Socio-Economics* 33(5): 587-606.

Gigerenzer, G. and H. Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1(1): 107-143.

Gigerenzer, G. and T. Sturm 2012. How (far) can rationality be naturalized? *Synthese* 187(1): 243-268.

Grüne-Yanoff, T. 2014. Teaching Philosophy of Science to Scientists: Why, What and How. *European Journal for Philosophy of Science* 4(1): 115-134

Grüne-Yanoff, T. 2021a. Choosing the right model for policy decision-making: the case of smallpox epidemiology. *Synthese* 198: 2463-2484. 1-22.

Grüne-Yanoff, T. 2021b. Justifying Method Choice: A Heuristic-Instrumentalist Account of Scientific Methodology. *Synthese* 199: 3903-3921.

Hey, S.P. 2016. Heuristics and meta-heuristics in scientific judgement. *The British Journal for the Philosophy of Science* 67(2): 471-495.

Hill, A.B. 1965. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* 58(5): 295-300.

Huff, D. 1954. *How to lie with statistics*. New York, NY: WW Norton & Company.

Laplane, L., P. Mantovani, R. Adolphs, H. Chang, A. Mantovani, M. McFall-Ngai, and T. Pradeu. 2019. Opinion: Why science needs philosophy. *Proceedings of the National Academy of Sciences* 116(10): 3948-3952.

Lenhard, J. and E. Winsberg. 2010. Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41(3): 253-262.

Levi, I. 1980. *The Enterprise of Knowledge, An Essay on Knowledge, Credal Probability, and Chances*. Cambridge, Mass: MIT Press.

Mayo, D.G. 2018. *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Morgan, K.L. and D.B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *Ann. Stat.* 40(2): 1263–1282.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: 6251.

Ravallion, M. 2020. Should the randomistas (continue to) rule? *NBER Working Paper* 27554. DOI 10.3386/w27554.

Rawlins, M. 2008. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet* 372: 2152–2161.

Savage, L.J. 1962. Subjective probability and statistical practice. In: Barnard, G.A. and G.A. Cox (eds.), The Foundations of Statistical Inference. Methuen, London, England, pp. 9–35.

Schulz, K. 1996. Randomised Trials, Human Nature, and Reporting Guidelines. *The Lancet*, 348: 596-598.

Smith, S. 2017. Why philosophy is so important in science education. *Aeon*. https://aeon.co/ideas/why-philosophy-is-so-important-in-science-education. Last accessed 18/01/2022.

Steele, J. M. 2005. Darrell Huff and Fifty Years of How to Lie with Statistics. *Statistical Science*, 20 (3), 205–209.

Trafimow, D. and M. Marks. 2015. Editorial. *Basic and Applied Social Psychology* 37(1): 1-2.

USPSTF - US Preventive Services Task Force, United States. Office of Disease Prevention, & Health Promotion. (1996). *Guide to clinical preventive services: report of the US Preventive Services Task Force*. US Department of Health and Human Services, Office of Public Health and Science, Office of Disease Prevention and Health Promotion.

Wasserstein R.L. and N.A. Lazar 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70 (2): 129–133.

Wicherts, J.M., C.L. Veldkamp, H.E. Augusteijn, M. Bakker, R. Van Aert and M.A. Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* 7, article #1832. doi: 10.3389/fpsyg.2016.01832

Wimsatt, W.C. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Ziliak, S.T. 2014. Balanced versus randomized field experiments in economics: why W. S. Gosset aka ‘Student’ matters. *Review of Behavioral Economics* 1: 167–208.

Zucchini, W. 2000. An introduction to model selection. *Journal of Mathematical Psychology* 44(1): 41-61.