The Epistemic Consequences of Pragmatic Value-Laden Scientific Inference

Adam P. Kubiak and Paweł Kawalec

March 9, 2021

Abstract

In this work, we explore the epistemic import of the value-ladenness of Neyman-Pearson's Theory of Testing Hypotheses (N-P) by reconstructing and extending Daniel Steel's argument for the legitimate influence of pragmatic values on scientific inference. We focus on how to properly understand N-P's pragmatic value-ladenness and the epistemic reliability of N-P. We develop an account of the twofold influence of pragmatic values on N-P's epistemic reliability and replicability. We refer to these two distinguished aspects as "direct" and "indirect". We discuss the replicability of experiments in terms of the indirect aspect and the replicability of outcomes in terms of the direct aspect. We argue that the influence of pragmatic values is beneficial to N-P's epistemic reliability and replicability indirectly. We show that while the direct influence of pragmatic values can be beneficial, its negative effects on reliability and replicability are also unavoidable in some cases, with the direct and indirect aspects possibly being incongruent.

1. Introduction

The value-free ideal of science (VFI) assumes that collecting evidence and formulating scientific conclusions can be undertaken without making pragmatic value judgments,[1] and states that scientists should attempt to minimize the influence of these pragmatic values on scientific reasoning (see e.g., Douglas 2009), and in particular on the methods of statistical inference. Among the general issue of the role of values in science, there is a lively discussion about whether it is beneficial that a decision to accept or reject a hypothesis considers the potential pragmatic consequences of error (see e.g., Elliott, Richards 2017). There are few general and paradigmatic approaches to statistical inference (see. e.g., Royall 1997; Bandopadhyay, Forster, 2011; Romeijn 2017). However, the approach that is relevant for the issue of the influence of pragmatic factors on the process of the scientific acceptance of a hypothesis, through taking into account the potential pragmatic consequences of errors, is J. Neyman and E. Pearson's (see e.g., Neyman 1952) conception of hypothesis testing (N-P hereafter), which itself is a classic method of frequentist statistics. There are two types of error distinguished in N-P: falsely rejecting the tested hypothesis and falsely accepting the tested hypothesis. N-P theory captures these pragmatic factors by explicitly articulating the pragmatic preferences for avoiding one type, and one size, of errors more so than others. The important question is whether or when the value-ladenness of statistical inference based on N-P is epistemically adverse, neutral, or perhaps beneficial.

The issue of the influence of pragmatically-driven (uneven) differentiation of error probabilities in N-P (PDDEP hereafter) on epistemic reliability can be related to the problem of replicability. This reference seems to be important as long as the replicability of outcomes

[1] This phrase embraces the notion of the influence of non-epistemic and non-cognitive values on scientific inference. A basic division of values that can influence the process of formulating scientific cognition as well as an explanation of how we use the term "pragmatic" is made in the subsequent part of this section.

is deemed the "gold standard for science" (Grant 2012). The problem of frequentist statistics meeting this standard has been debated in established discussions (see e.g., Stahel 2017). Many authors explicitly analyze frequentist hypothesis testing from the perspective of the replication crisis (Rubin 2019; Open Science Collaboration 2015; Johnson 2013; Amrhein et al. 2017). The role of the two types of error, and the unproportionate concern about one of them, has also been discussed from certain replication perspectives (e.g., Fiedler et al. 2012; LeBel et al. 2017), but this discussion lacks a deepened analysis of the influence of the pragmatic value-ladenness which is the root of this unproportionate concern.

An important defense of the pragmatic value-ladenness of scientific inference has recently been propounded by Daniel Steel (2010), who discusses the so-called *argument from inductive risk* (Rudner 1953).

Steel has paraphrased the argument from inductive risk as follows (2010, 17):

> "1. One central aim of scientific inference is to decide whether to accept or reject hypotheses.
>
> 2. Decisions about whether to accept or reject a hypothesis should depend in part on nonepistemic value judgments about the costs of accepting the hypothesis when it is false and rejecting it when it is true.
>
> 3. Therefore, nonepistemic values should influence scientific inference."

We draw upon those claims of Steel which can be interpreted as referring to N-P, or whose cogency can be directly verified in reference to N-P. We examine if whether these claims comply with and are validated by the main tenets of N-P. We check to see if Steel's

arguments can be amended or extended to remain consistent with N-P.[2] Finally, we analyze the feasibility, necessity, and consequences of the arguments on replication issues, which we build from Steel's ideas and N-P.

The paper is structured as follows. We begin by presenting the main tenets of N-P. Further, we argue that the epistemic reliability of N-P can be assessed despite its decision-theoretic character. Next, we explicate Steel's claims and apply them in examining the epistemic consequences of N-P's PDDEP. Then, we embark on an investigation of Steel's intuition that it is epistemically better to accept a false hypothesis that is close to the truth than to accept a false hypothesis that is very far from the truth. After the argument for the discussed type of value-ladenness of N-P is formed, we briefly discuss the subjectivity of value judgments and the possible avoidance of value judgments. We then devote attention to the consequences of our argumentation upon the replicability of experiments and research outcomes. In the concluding section, we situate our results in a more general philosophical context.

We proceed with a brief exposition of the basic differences between the types of values that may differently inform each of the research process's main stages. Manifold values can influence the formulation of a research problem, gathering of evidence, acceptance of a hypothesis, and analysis and theoretical interpretation of an outcome. These values may affect the content of evidence, hypothesis, and theory, as well as the features of the methods used, like, for example, the margin of acceptable error. According to the classic approach, these values follows the threefold classification (see e.g., Laudan 2004): (i) *cognitive values* that constitute the criteria of scientific understanding but do not directly relate to the goal of

[2] Steel elaborates his original arguments in subsequent publications, e.g. (Kaivanto, Steel 2019), but nevertheless the basic claims applicable to N-P remain as presented in his original paper (2010) which we discuss here.

discovering the truth (e.g., simplicity, explanatory power, falsifiability, and consistency with other theories); (ii) *epistemic values* whose implementation is related to successful truth acquaintance (in the classical sense). The relevant aspects thereof include the possibility of getting to know new truths at all, the low risk of committing an error, and the accuracy of conclusions; and (iii) *social values* that are neither epistemic nor cognitive (under this category fall, for example, economic, ethical, cultural, political, and religious values). The distinction between (i) and (ii) has been contested and disputed (see e.g., Steel 2010; Douglas 2013). While we do not discuss how cognitive values may be truth conducive, we assume that, as a set, cognitive and epistemic values jointly exercise a positive influence on reaching the attainment of truth.[3] By the latter, we mean the positive contribution of these values towards reaching the goal of the assertion of new theses that are close to the truth and avoid the assertion of theses that are far from the truth (cf. David 2001). Hereafter, this is referred to as the *truth goal*. We call the aforementioned set of values "epistemic" and contrast them with social values which we call "pragmatic" to stress that these pragmatic values aim at reaching pragmatic goals without a direct connection to reaching the goal of discovering the truth, understood classically as the assertion of claims that mirror reality. The use of this term also stresses that the strict association of N-P's imbalance of error probabilities with these types of values is connected to pragmatistic epistemology (Chiffi, Pietarinen 2019). In this paper, we undertake a philosophical analysis of the influence of the pragmatic values presumed by N-P regarding their potential epistemic effect, i.e., their influence on the attainment of truth. We distinguish two aspects of reaching the truth goal as defined above.

The first aspect concerns what contributes to an increase in a researcher's potential in asking new and fruitful questions that can lead to asserting new hypotheses that are close to

[3] Steel (2010) would call these values, respectively, *extrinsically* and *intrinsically* epistemic values.

the truth. This includes, in particular, the acquiring of new epistemic resources, maintaining these resources, and increasing the effectiveness of their use to contribute to the goal of the attainment of truth. These resources are human resources, money, time, and access to material resources necessary to conduct research. The epistemic reliability of N-P within this aspect we call *indirect epistemic reliability*. The adjective "indirect" is used to distinguish this term from N-P's reliability in obtaining the truth goal as straightforwardly indicated by probabilities of error of a given type and size. This straightforward reliability refers to the second aspect of the truth goal, which is the avoidance of the assertion of false hypotheses, especially those that are distant from the truth. This straightforward reliability we call *direct epistemic reliability*. A test has a higher direct epistemic reliability the better it protects itself from committing errors, primarily those distant from the truth. A test has a higher indirect reliability, the better it can secure maintenance and effective use of resources that are the perquisite for asking and answering new research questions.

2. Neyman-Pearson Hypothesis Testing

While we assume that the reader is already familiar with the tenets of N-P, here we review some rudimentary facts with an emphasis on the uneven pragmatic setting of the importance of errors.

2.1. The Two Types of Error

Two possible unsatisfactory cases may result from the application of a statistical test: H is true whereas the action taken is a rejection of H, or the complement H^C is true while the

action is the acceptance of H (see Neyman 1950, 261). There are two kinds of random errors[4] associated with the two types of unsatisfactory cases:

(a) the error of the I[st] type $\alpha = P(reject\ H|h\ is\ true)$, and

(b) the error of the II[nd] type $\beta = P(accept\ H|h'\ is\ true)$[5]

where h is a simple hypothesis – a particular instance of H or equivalent to H (likewise h' is a particular instance of H^C or equivalent to H^C), and $1-\beta$ is the probability that, given h', the sample point will fall in the rejection region specified for h. This latter point concerns the probability that a test will detect the falsehood of H (reject it) when the hypothesis h' is true—this is what Neyman called the *power* of a test (Neyman 1952, 55; Neyman 1950, 267-268).[6] As a function of the possible point hypothesis, power is the essential category in deciding which test to choose.[7] What is important is that based on power analysis, N-P advocates as standard the choice of the so-called *likelihood ratio test* according to which the decision to accept one of the hypothesis is essentially based on which of them is more supported by the evidence obtained (see Neyman, Pearson 1928, 1933; Neyman 1950, 340-341).

Keeping the first type of error at the desired nominal level is unproblematic as the researcher determines the significance level (α) of a test procedure during the research design

[4] The verdict on taking a particular action is random. This is because such a verdict depends on the random variable(s) which determine the position of the sample point. Due to this, there is no inconsistency in considering the probability of the verdict having a certain property, such as being erroneous (Neyman 1950, 56-57).

[5] Today it is standard to use "β" to represent the probability of making an error of the II[nd] type, but in N-P's original version the notation for this error was "$1-\beta$."

[6] The presented concept of errors is the most general approach that covers various test cases (e.g. cases with different distributions) and interval estimation. This does not cover errors of a different nature like errors of measurement instruments, biases, and model assumption errors (e.g. false normality, the assumption of independence).

[7] This issue, however, is beyond the scope of this paper; the reader can consult e.g. (Neyman 1950, 258-268).

stage. The β nominal error rate depends in turn on the chosen value of α that determines the rejection region, a fixed instance of an alternative hypothesis, and on the distribution of a test statistic, which is thus determined by the value of a sample size and by population variance. One important consequence of this is that an increase of α results in a decrease of β for any particular parameter value in H^C provided a test is unbiased.[8]

2.2. The Importance of Errors

It needs to be stressed that these two types of error rates do not necessarily have equal importance.[9] For Neyman, this is consistent with the context of applications, where the importance of avoiding these two types of error turns out to be strikingly unequal:

> "The adoption of hypothesis H when it is false is an error qualitatively different from the error consisting of rejecting H when it is true. This distinction is very important because, with rare exceptions, the importance of the two errors is different, and this difference must be taken into consideration when selecting the appropriate test" (Neyman 1950, 261).

The importance of an error is related to the pragmatic costs of making it. To illustrate the difference in the relative importance of the two aforementioned kinds of error, Neyman examined an example of testing drug toxicity, in which the random variable is the number of deaths of experimental animals that receive a particular dosage. Suppose that the new drug is a slightly cheaper generic substitute for a commonly used drug. In the case of the hypothesis

[8] A test is unbiased when its power against any alternative point hypothesis is at least as high as a type I error (see Neyman, Pearson 1936, 210-211).

[9] Of course, it is possible to set these error rates as equal. Once the values necessary to calculate the probability of the IInd type of error for the desired value of h', which is sometimes regarded as representing the "effect size" that is relevant from the perspective of the theory of the subject of research, are known (population variance can be estimated based on the sample), one can increase the probability of the Ist type of error in order to decrease the probability of the IInd type of error to an appropriate level.

that the new drug is toxic, it would be natural to think that falsely rejecting the hypothesis, asserting that the drug is safe when it actually is toxic, would be potentially much more pragmatically harmful than in the case of falsely asserting that the drug is toxic. Thus, the first type error would be more important to avoid in contrast to the error of the second type (Neyman 1950, 262-263). The concept of a more important error is not uniquely connected to errors of the first kind. Which of the two types of error is more important depends on the context of research and how a hypothesis is formulated. There may be cases in which one would prefer to avoid an error of the first type and where, "(…) the desirable property of the test of *H* is as high a power as practicable, perhaps with some neglect of the probability of rejecting *H* when true" (Neyman 1971, 4).

2.3. The Epistemic-Pragmatic Nature of N-P

Given this exposition of hypothesis testing, it can be tentatively assumed that N-P, partially and in a specific way, can abide by the truth goal. This is in a minimal sense reflected by the two types of error's nominal probabilities, which can be understood as the most basic indicators of the level of a method's epistemic reliability in the two aspects of error risk.[10] This, together with the idea of the size (magnitude) of error and the ability to increase potentiality to undergo more independent tests, are the most basic and natural aspects in which N-P can be considered as being somehow related to the realization of the truth goal. These aspects are the basis of our argumentation, but they do not preclude further analysis similar to ours using other, perhaps more specific measures of the epistemic reliability of a hypothesis test, particularly observation-dependent measures (e.g., Spielman 1973; Mayo,

[10] This understanding of reliability resembles process reliability as formulated in the epistemological debate over the justification of beliefs where an assertion (in the case of a propositional formulation) or belief (in the case of a doxastic formulation) is justified if it is formed by a reliable process (Goldman 2008).

Spanos 2006) and non-observation-dependent measures (e.g., Ioannidis 2005; Rochefort-Maranda 2013; Kubiak et al. 2020).[11]

While acknowledging that N-P can have an epistemic reliability, it is an inherent trait of the method that the outcomes of performing N-P procedures also depend on the researcher's pragmatic goals. This epistemic-pragmatic blend is a tenet of N-P explicitly acknowledged by Neyman: "(…) the theory was born and constructed with the view of diminishing the relative frequency of errors, particularly of 'important' errors" (Neyman 1977, 108). He thus succinctly articulated his method's twofold ambition: epistemic—to avoid errors, and pragmatic—to discern the important errors conditional on the user's pragmatic considerations which, in consequence, influence the outcome of statistical inference. To recognize this influence, it suffices to consider, for example, a case in which avoidance of the Ist type of error is strongly favored, and a false *H* is asserted. If, instead, the IInd type of error had been treated as sufficiently more pragmatically important than the Ist type of error, and thus was avoided with greater stringency, then a false *H* would not be asserted.

Neyman interpreted statistical hypothesis testing as a special case of Abraham Wald's (1950) more general theory of making pragmatic decisions (Neyman 1957; 1971). So, the question arises if whether the method in question's epistemic performance becomes irrelevant, given its ultimate goal is to make pragmatic decisions. The question of how well a method of making pragmatically favorable decisions conforms to some epistemic norms seems analogous to how the technique of launching a rocket conforms to some criteria of

[11] Error probabilities in N-P are indicators of the level of outcome-independent epistemic reliability; following Rochefort-Maranda: "One way to distinguish both concepts (level of evidential support and level of its credibility) is to realize that the credibility of the support does not depend on the actual output of the instrument whereas the degree of support does" (2013, 11).

quietness, despite the goal of the technique not being to be quiet. As such, a particular method can serve one goal and simultaneously conform, to a greater or lesser degree, to some other criteria possibly unrelated to that goal. Therefore, N-P can potentially conform to some epistemically reliable criteria, although the method's focal goal is to be pragmatically reliable. The latter is apparently more successful if there is also the epistemic aspect present in N-P. Namely, the success rate in avoiding pragmatically unfavorable decisions is a function of the risk of making false assertions (expressed by error probabilities), and concerns the epistemic aspect and of the pragmatic loss associated with such assertions, which is itself the pragmatic aspect.

3. Steel's Arguments for Acceptability of a Pragmatically-Driven Uneven Setting of Error Risks

Steel (2010) has recently presented a convincing counterargument against the doubt that the implementation of pragmatic preferences within procedures of statistical inference promotes pragmatic goals at the expense of an ameliorated accomplishment of epistemic goals. He challenges the view that pragmatic values impede the attainment of truth and argues that these values can promote the truth. Steel's arguments considered here may be divided into two categories which correspond to how error rates reckon with pragmatic values. One category, which we call the *inter-test* aspect, corresponds to the pragmatic value-driven differentiation of error probabilities, particularly the threshold for the error of the first type, in different testing situations. The second category, which we call the *in-test* aspect, corresponds to the pragmatic value-driven differentiation of error probabilities found between the two

aforementioned types of error in a particular testing situation.[12] We apply these categories in examining Steel's arguments. First, we discuss an argument regarding the *inter-test* aspect which claims that when testing is performed in different research contexts, it is epistemically good that the assigned error levels are differentiated for pragmatic considerations. The second argument considers the *in-test* aspect to the effect that the stringency of avoiding an error of one of the two aforementioned types does not need to be, from the epistemic perspective, equal to the stringency of avoiding an error of the IInd type.

3.1. The Argument for the Inter-Test Presence of PDDEP

Steel's argument runs as follows. It is a truism that our cognitive resources are limited and therefore need to be somehow allocated. Some investigated hypotheses are more useful as the foundation for further research than others. Thus, from the epistemic perspective, more cognitive resources should be devoted to those momentous research questions whose error rates should be set lower than for other questions. One also needs to take into account that while setting very stringent standards for avoiding errors in promoting the truth goal, at the same time this setting of standards blocks this truth goal on a more fundamental level. Very exacting standards suspend the drawing of conclusions. By continuing to devote more resources to making a research outcome more accurate, a researcher suspends completion of the research process and deprives themself of potential cognitive resources that could have been used to resolve other research questions. Moreover, it may be more important for some research questions to be answered more correctly than others, depending on their importance in further research programs. Therefore, a balance must be reached between the need to avoid mistakes and the need to effectively scrutinize hypotheses in finite time and with limited resources. Setting a threshold for being wrong but acceptably close to the truth allows a

[12] Steel does not offer such a distinction himself, but this distinction becomes indispensable for the purpose of our analysis.

researcher to continue her efforts in testing other hypotheses. PDDEP between tests reflects such a balance between a test's stringency and the need to test a hypothesis in a finite time and with respect to available resources in a given research context. For different research contexts, this balance can be set differently due to the diversity of available resources in the particular pragmatic contexts of research processes (Steel 2010, 27-28).[13]

3.2. The Argument for the In-Test Presence of PDDEP

The argument in favor of the in-test epistemic neutrality of treating different types of errors on unequal terms due to pragmatic reasons is the following. Steel compared two cases, one of which is the acceptance of a false assertion of the value of a parameter that is of some distance from the true value, where the true value of the parameter is greater than the falsely asserted parameter value. The other case concerns a false assertion of a value that is of the same distance from the true value, but this time is greater than the true value. The true value is 0.01, while the two false statements claim 0.005 and 0.015, respectively. Steel stated that in these two cases of committing an error, there is no epistemic reason for favoring the avoidance of one kind of error over the other and that this does not mean that there is an epistemic reason against such a preference (Steel 2010, 29). This means that there is no evident reason to state that it would be adverse epistemically to treat the avoidance of one of the two aforementioned errors as more important for some pragmatic reasons.

4. Extending and Applying Steel's arguments to N-P

[13] An example of a universal pragmatic factor that determines the distribution of cognitive resources is the cost of a particular research process. Setting a lower standard of accuracy for an expensive research project may open the possibility of answering a number of other questions which are less expensive to settle. Of course, by extending the body of knowledge, this is a benefit from a purely epistemic perspective. Nonetheless, a level of standards must also consider expectations related to the pragmatic contexts of error risks.

Two conclusions are supported by the inter-test version of Steel's argument. First,

> (C1) for different research processes carried out in different pragmatic contexts, a pragmatically driven setting of different error rates of α,[14] or β, or both, is epistemically favorable.[15]

Of course, (C1) holds when this difference is adequately balanced in accordance with pragmatic reasons that are correlated with the increase of the cognitive effectiveness in a particular research context.[16]

Second,

> (C2) excessive minimization of errors can occur in a statistical test and is epistemically unfavorable.[17] In particular, an effort to avoid infinitely small errors is epistemically bad.

The consequence of the in-test argument would be that

> (C3) an additional pragmatic reason to favor the avoidance of one kind of error over some other error is legitimate when these errors are both equally bad from an epistemic perspective.

Steel did not examine how his in-test argument applies to the case of N-P testing and the two types of error assumed by N-P. Nevertheless, an analogy to testing may be helpful. There is

[14] This means that it may be epistemically optimal to set, for example, the Ist type of error as 0.05 in one context, and 0.01 in another.

[15] This claim refers to a between-test difference and does not favor the setting of different error rates for the two types of error in a particular test.

[16] Such a balance of the risk of error is prevalent in the applied sciences, such as conservational biology, where the pragmatic goal of research guides the decision concerning the appropriate level of research standards and includes a proper balancing of error risks (Baumgaertner, Holthuijzen 2017, 49-51).

[17] Committing errors needs to be optimally balanced against limited resources, as described in Section 3.1.

no epistemic reason to favor the wrong acceptance of H over the wrong acceptance of H^C (or the other way around). This does not mean that there are epistemic reasons against such a favoring. Therefore, having some additional pragmatic reason to set a lower error rate for the wrong acceptance of H^C than for the wrong acceptance of H is legitimate. At the same time, given Steel's example, it is clear that he had in mind a case in which the compared errors are of the same distance from the true value. Applied to N-P, this would entail that Steel's argument is restricted to cases where the wrongly accepted H is considered to have the same distance from a true value as H^C would have if it was wrongly accepted.

Now, the questions arise; how to compare two possible errors that are not of the same size, and what is the subsequent epistemic consequence of PDDEP? Although Steel did not explicitly consider this question, he formulated a general principle concerning the epistemic weight of errors of different sizes:

> (A1) "it is epistemically better to accept a false hypothesis that is close to the truth than to accept a false hypothesis that is very far from the truth" (Steel 2010, 29).

The consequence of (A1) is that

> (C4) from the epistemic perspective, avoidance of one error should be favored over avoidance of another if the first is meaningfully farther from the truth than the second.

Statement (A1) should be viewed as the completion of (C3). When errors are of the same size, then the difference of stringency in avoiding both is irrelevant epistemically. However, when errors are of different sizes, it is epistemically more important to avoid a bigger error (a false conclusion that is farther from the truth) than a smaller one (a false conclusion closer to the truth). However, when we applied (A1) to N-P, this exposed interesting tensions concealed in Steel's original argument. These tensions prompted a more detailed analysis, and as such, in

the next section, we will be devoted to examining them. Before moving to this topic, we must strengthen the conclusions of Steel's arguments and analyze how they justify the claim that PDDEP in N-P is, or can be, epistemically beneficial.

It is important to note that the types of positive epistemic influences of PDDEP in N-P, considered in the inter-test argument, do not make N-P more reliable by positively contributing to the avoidance of the assertion of false hypotheses, especially those that are distant from the truth. Instead, these influences increase the epistemic potential of researchers in the first aspect of the epistemic goal of the attainment of truth as defined in Section 1. So far, we have explicated how Steel's arguments can be specifically applied to N-P. Below we expand his line of reasoning and draw a generalized conclusion.

The plausibility of (C2) is strongly justified by the fact that it is thoroughly entrenched in research practice and even pushed to its extreme form in the case of incredibly small departures from the truth, i.e., false theses that only slightly depart from the truth are accepted with premeditation. Such actual practices confirm that (C2) is indeed applied in scientific research. Striking examples of research policy that deliberately assert false hypotheses when it comes to errors of small size include the following practices:

(1) Rounding up mathematical values in statements, as in the case of the value of π. In building theories (e.g., predictive ones), people assert π to be, for example, 3.1416. This is because people think that providing a more accurate value, say 3.14159 26535 89793 23846 26433 83279, even if possible, would be irrelevant epistemically to a theory. Providing the true value of π is essentially impossible, thus the acceptance of a false statement about π with precision dictated by pragmatic considerations is a necessity.

(2) A model of the hydrological processes in a big river's catchment area will typically ignore garden ponds. Such a fine-grained scale is epistemically irrelevant in modeling large-scale hydrological phenomena and can even hinder this as too precise models are possibly less accurate in their predictions, as aptly illustrated by Gigerenzer and Brighton (2009).

(3) Similarly, an engineer in her calculation of the load capacity of a new prototype car will presumably ignore the Moon's gravitational pull.[18]

Steel argued that pragmatically-driven, balanced, and diverse inter-test allocation of finite epistemic resources in the form of diverse risk levels could yield more epistemic successes. We have supplemented this argument with the above examples. There is a duality between N-P tests and confidence intervals (see e.g., Bickel, Doksum 2001, 241-248), and changes in error probabilities are interrelated with the changes of a confidence interval's accuracy. Therefore, these examples indicate that a pragmatically-driven setting of the rate of error is sometimes inevitable. This thus spares epistemic resources and uses them more efficiently, as in the case of a superior prediction of a phenomenon at a given scale in which too precise models are possibly less accurate. It is important to note that a possible exaggeration applies to particular situations of performing a statistical test. This means that conclusion (C2) does not state that the amalgamation of outcomes should not lead to a continuous and progressive minimalization of errors as far as meta-analysis is concerned. For example, a hypothetical extreme exaggeration would be the idea to avoid infinitely small errors in a particular situation of performing a statistical test. Obviously, this does not mean that a continuous increase of accuracy or decrease of an error risk over time and upon acquiring new research resources is adverse epistemically.

[18] For some other arguments in favour of the thesis that a false assertion can serve an epistemic advantage, see Wimsatt (2007, 93-132).

The above examples show that it can be indirectly epistemically beneficial to determine the different risks of error of definite error sizes—avoid the risk of big errors and neglect small errors—in different testing situations, depending on the research context (including the goals of application) of a testing situation. Statement (C2) also epistemically justifies a significant in-test difference in the probability of the IInd type of error dependent on the error's size. The acceptance of a false hypothesis in the case when it departs only slightly from the truth is less harshly avoided than acceptance of this hypothesis if the truth would be far from it. Alas, this does not justify that in-test PDDEP in N-P can positively influence epistemic reliability, as the in-test argument's conclusion does not address this statement.[19] We supplement Steel's argument by providing a rationale for such a claim by demonstrating that more pragmatically harmful[20] errors can indirectly cause higher epistemic damage than errors that do little pragmatic damage. This stems from the fact that democratic society controls science by way of funding it. If a research team would exceed the threshold of societal acceptance of the consequences of harmful pragmatic errors, the continuation of the whole research program would simply be suspended or stopped by democratic institutions. Thus, pragmatically harmful errors may turn out to also be harmful epistemically. The simple example of research on new drugs funded by a democratically governed agency makes this clear. By adopting the pragmatically motivated focus of avoiding the false assertion that the drug is non-toxic, the research program will presumably avoid pragmatically harmful errors. This, in turn, will also lead to positive epistemic consequences, as the research team will be able to continue their research activities and extend the existing body of knowledge rather than face a shortage of funding or even be banned if they would fail to avoid, and in effect commit, errors that may come to be unwanted by donors. This means that a small error of the Ist type may have, all things considered (such as the context of research and the way a

[19] What can be derived so far is that this can be neutral in some specific cases.

[20] This means bringing about a pragmatic loss.

hypothesis is formulated), more grave epistemic consequences because of its indirect negative epistemic consequences than a bigger error of the IInd type.

Hence, the general conclusion concerning the question of the proper approach to the assessment of epistemic reliability is that

> (C5) the indirect epistemic reliability of a method of scientific investigation rises if the method can utilize pragmatic considerations to properly: (a) differentiate error risk levels in different research processes and (b) differentiate the risk of different types and sizes of errors in a particular research process.

For in-test PDDEP in N-P, under limited cognitive resources, the in-test distribution of the risk of error has to be adequately balanced. This is due not only to the direct epistemic weight of potential errors of different sizes but also to the following:

> (C6) the in-test distribution of the risk of error has to also be adequately balanced in terms of the pragmatic preference for one type of error over another, and the distribution that would set the same error risk for the Ist kind of error and the IInd kind of error of the same size is not optimal from the perspective of the level of indirect epistemic reliability.

In section 5, we examine Steel's assumption (A1) and show that it leads to strains in presenting Steel's arguments and when applied to N-P. We argue that these issues can be resolved in a way consistent with Steel's original arguments and the original views of Neyman and Pearson. We also address the potential tension found between the direct and indirect epistemic reliability of N-P.

5. Is Committing a Big Error Epistemically Graver Than Committing a Small Error?

Intuitively, (A1) seems to be a true assumption. Though, it apparently is in tension with Steel's inter-test argument. Moreover, (A1) is also undermined by our in-test argumentation. We discuss this in more detail below.

The first problem concerns the inter-test aspect. Consider two cases of research regarding a certain quantitative characteristic of a population, such as a certain economic determinant which, unknowingly to the investigators, happens to have the same value. In both cases, the statistical models are also the same, but the pragmatic contexts differ. Imagine a quantity is sought out for in two countries with different policies regarding research funding and concern towards the relevance of the quantity in question. In accordance with (C1), and because of the difference between the pragmatic contexts, it may be favorable epistemically to reflect this difference in setting error risks. So, in one of these cases, a risk of, say, α error of size s may be rightfully set lower than the risk of the α error of a size much greater than s in the second case. In other words, in the first case, the pragmatic context may require the setting of a more stringent (smaller) error risk for a small error than in the second case in which the error considered is large. This possibility is entailed by (C1) and is captured in the conclusion (C5a). The discussed inconsistency follows from the fact that (A1) entails (C4). This states that it would be principally favorable epistemically to set a lower error risk for the error considered in the second case than the one considered in the first case. This means that the consequence of (A1) described above is contrary to the consequence of (C5a), although (A1) is Steel's own statement and (C5a) is a crucial conclusion that follows from Steel's argument.

The second problem with (A1) as seen from the in-test aspect is that (A1) supports (C4) which, in turn, is undermined by the conclusion (C5b) of our argument for the epistemic profitability of in-test PDDEP. If both direct and indirect aspects of epistemic reliability are at stake, then it may happen that focusing on the severe avoidance of a small error of one type will yield greater net epistemic reliability than focusing on the firm avoidance of a big error of

the other type. The aforementioned example of research on drug toxicity illustrates this. In case the level of toxicity of the drug slightly oversteps the threshold of dangerous toxic reactiveness[21], and we assert that this drug does not exceed the threshold, the overall unwelcome epistemic consequences may be grave because of the indirect consequence of being dangerously reactive, while the direct epistemic consequence (the error of small size) remains insignificant. In contrast, a slightly greater error in overestimating the toxicity of the drug, as well as the almost insignificant indirect epistemic consequence of this mistake, may lead to a less unwelcome epistemic result, all things considered. In such a case, it would be epistemically better to more stringently avoid the possible error of the smaller size.

Such cases occur in the standard application of N-P. Therefore (A1) seems to be repudiated when it comes to comparing the size of the risk in the two aforementioned types of error. Consider a t-test and a composite hypothesis H: $\mu \leq \mu_0$. The risk (β) of asserting a false H, when the hypothetical true value (h') is a little lower than the critical value representing the rejection threshold, equal to slightly more than 0.5. Meanwhile, the risk of falsely asserting H^C when the true value (h) amounts to or is slightly less than μ_0 is equal to or slightly less than α, which can be fixed at the level of, say, 0.01. Thus, the N-P testing framework allows for cases in which the risk of committing an error of a smaller distance from the truth (the second case) is lower than the risk of committing an error of greater distance from the truth (the first case). In general, when two errors belonging to two types thereof are compared in terms of their "magnitude" (the distance of the asserted hypothesis—H or H^C respectively—from the true value), the following possibility emerges: the distance d from a wrongly accepted hypothesis to a hypothetical true value may happen to satisfy

$$d_1(H^C, h) < d_2(H, h'), \quad (1)$$

[21] In chemistry a change in the reactivity of a compound with increasing some factor associated with it is often of a qualitative character.

where $d_1(H^C, h)$ stands for the distance d_1 between the accepted H^C and the true value h, and $d_2(H, h')$ for the distance d_2 between the accepted H and the true value h', while

$$P(accept\ H^C|h) < P(accept\ H|h'). \quad (2)$$

Such a possibility does not conform to the restriction entailed by Steel's statement (A1) because, as in the above example, the possible mistake that is epistemically worse is less studiously avoided than the possible mistake that is epistemically better in the sense of (A1).

So far, we have argued that (A1) is inconsistent with Steel's inter-test argument, conclusion (C5), and the N-P scheme. Granting all this, assumption (A1) may be considered as too strong a requirement. If (A1) holds, it only does so within the particular pragmatic context of a test and for one type of error. Therefore, (A1) should be restated in a restricted form. Neyman and Pearson explicitly considered an intuition similar to (A1) and restricted to the second type of error:

> "(…) if wrong judgments cannot be avoided, their seriousness will, at any rate, be diminished if on the whole Hypothesis A is wrongly accepted only in cases where the true sampled population, Π', differs but slightly from Π" (Neyman, Pearson 1928, 177).[22]

Granted (C5), an adjusted version of (A1) could be formulated thus:

> (A1') in a single testing situation and for one type of error, it is epistemically better to accept a false hypothesis that is close to the truth than to accept a false hypothesis that is very far from the truth.

(C4) has to be reformulated accordingly:

[22]See also (Neyman 1950, 277-278).

(C4') in a single testing situation and for one type of error, the avoidance of a bigger error should be, from the epistemic perspective, favored over the avoidance of a smaller error. This restriction of (A1) and (C4) to one testing process and one type of error depends on argumentation that appeals to the aspect of indirect epistemic reliability. Still, (A1) appears intuitively to be correct when one considers only direct epistemic reliability. Therefore, it seems plausible to admit that it would be better if the original (A1) would be satisfied by N-P. Putting aside the inter-test aspect, N-P would satisfy (A1) if probabilities for both types of error were the same and thus expressed no preference for avoiding one type of error more stringently than the other. Satisfaction of (A1) would be directly beneficial epistemically if this would increase the overall potential of avoiding large errors. This turns out not to be the case because while PDDEP increases the risk of committing large errors of one type, it is compensated by a decrease in the risk of committing large errors of another type. So, while the prior probability of H is assumed in N-P to be unknown, one cannot claim that obeying (A1) would principally increase the general potential of avoiding large errors. This may differ from case to case depending mainly on truth value (or probability) of H.

In the case of inter-test PDDEP, the increase or decrease of N-P's direct potential of avoiding large errors also depends on the respective situation. This potential has to be evaluated as pertaining to all testing cases considered jointly. Suppose, given a fixed amount of research resources, that in two instances of a research community testing different research questions, or the same one, the error risks are higher in the second case because of the setting of lower error risks in the first case. In this situation, the general potential of avoiding a large error is not decreased by inter-test PDDEP. What happens is only a *local* decrease of this in the second case and a local increase thereof in the first case. Therefore, the increase or decrease of the direct epistemic reliability as the resultant of the inter-test PDDEP effect on the reliability of all cases jointly considered can again go in either of the two aformentioned

directions and depends mainly on how changes in particular testing processes compensate each other, on the truth-value (or the probability) of the hypotheses tested. Thus, the upshot is that

(C7) the violation of (A1) and implementation of pragmatic values in accordance with PDDEP:

(C7a) does not necessarily make N-P less directly epistemically reliable if in-test PDDEP is considered, and

(C7b) does not necessarily make N-P less directly epistemically reliable, if inter-test PDDEP is considered.

6. The Feasibility and Necessity of Pragmatic Value-Ladenness

Before drawing the consequences of our analysis for replication issues, we first touch upon two concerns. The first concern is the subjectivity of the pragmatic importance of errors and subsequent pluralism, and the second concern is the possibility of disentangling the discussed pragmatic influences from scientific inference. It is important to identify and address the first concern because, if sufficiently objective and controllable PDDEP is not possible, our conclusions concerning PDDEP drawn thus far may seem inapplicable. The second concern is important because if PDDEP could be avoided and replaced by a more efficient solution while N-P is in use, our conclusions could be found to be irrelevant.

6.1. The Subjectivity and Plurality of Pragmatic Value Judgments

Neyman pointed out that there may be several different points of view regarding the importance of both types of error. Sometimes this may take the form of a conflict. For

example, from the viewpoint of a jury who may grant or refuse a lady's claim of having the ability to distinguish by taste if whether milk or tea was first poured in a cup (and perhaps award her for having this distinguishable skill), it seems natural to consider that the more important error to avoid is granting the claim when in fact it is false. On the other hand, for the lady, it seems natural to recognize the error of falsely asserting that she has no ability as more important to avoid (Neyman 1950, 274). According to Neyman, this relative importance is not an issue for the method itself, as "this subjective element lies outside of the theory of statistics" (Neyman 1950, 263). Nonetheless, this problematizes the indirect aspect concerning what from one pragmatic perspective can have a positive indirect influence on epistemic reliability and can negatively influence epistemic reliability when seen from another pragmatic perspective. It seems that the perquisite for PDDEP is the possibility of finding a basis for standards for determining PDDEP in the light of people's pluralistic and often conflicting pragmatic and ethical views.

John (2015) argues that scientific knowledge based on value-laden statistical inference is inapplicable in the public communicative context. This means such knowledge cannot contribute to "public knowledge" (Kitcher 2011, 85)—the body of shared scientific knowledge from which people draw in pursuing their own ends. The reason for this is that the standards for determining PDDEP are floating. If standards are floating, value-laden outcomes should vary depending on the identified audience and the pragmatic consequences related to the communication of the outcome to a particular audience. Due to this author, such an approach cannot govern much scientific assertion as it would strip scientific knowledge of its public character, which is not acceptable.

But not all pragmatic standards for PDDEP are necessarily floating. Like the practical limits of accuracy in a given type of research, some of them could even be seen as a pragmatic standard anchored in a scientific discipline itself. In other cases, where if among the variation

of advertised values, none is common enough, scientists could base their standards on those audiences who matter most to them both pragmatically and epistemically, and upon whom the prolificacy and limitations of scientific resources depend on the most. Finally, the responsibility for implementing solutions that would become binding standards for scientists and that could serve to infer the best PDDEP for particular research cases can be relegated to the democratic process that sets societal preferences (see Kitcher 2001).[23]

It appears that objectivizing and coordinating the standards of scientists, which may be particularly important in regards to the discussion on replication, is not utterly impossible. Still, in analyzing the question of whether the value-ladenness of scientific inference can be epistemically justified, we find a slightly different question than the question of when and how, in detail, this can be put to work. We argued that value-ladenness is necessary and is partially epistemically profitable when it can be successfully applied. How successful it is applied is another question. In particular, a successful application thereof assumes the truth-directedness of science and scientific policy. If this basic condition is not satisfied, the abolishment of the truth-goal can take the form of the paradoxical effect of researchers being provided with more epistemic resources if their discoveries, although false, are in line with the donor's untruthful expectations. Perhaps this is less likely if the source of the risk standard adopted is controlled by a democratic society and not, for example, by private organizations. Such a conjecture might be supported by the observation that the whole of society is presumably more likely to request audits, to detect fraud, and eventually minimize error as compared to a small, less diversified, and more prone to bias subgroup (like a private organization) (see Page 2007).

[23] Pointing at such a possibility is sufficient for argumentative purposes. We do not aim to offer any particular sociological solution, and thus refrain from analyzing pathologies such as a scientific establishement's masking of the pragmatic motives behind decisions about error rate standards, etc.

It appears to us that any philosophical or methodological argument regarding the epistemic characteristics of statistical methods is vulnerable to the existence of scientific fraud but taking up this topic is beyond the scope of goals of this paper. Our analysis can be seen as relating to the case of the proper application of values in the minimal sense, by which we mean such that do not reject the truth-goal in the first place.

6.2. Avoidance of Value Judgments

In Section 4, we argued that some value judgments of the inter-test character in the setting the risks of different sizes of error are inevitable. But perhaps some value judgments that influence the outcome of the testing method of N-P can possibly be methodologically replaced by a qualitatively different solution, or perhaps be detached from the method of scientific inference and implemented in the form of standards concerning the post-research deliberation of decision-makers.

An interesting argument that value-laden scientific decisions (conclusions) can be systematically avoided by a change of methodology was introduced by Betz (2013). Betz's position is based on demonstrating that policy-relevant scientific statements can be avoided by stating "hedged" hypotheses that are weakened enough to be established beyond a reasonable doubt and which avoid substantial inductive risk. These hypotheses are intended to be represented by "ranges of observational values" as opposed to "plain" or "unequivocal" hypotheses. This assumes that what should dictate such statements for pragmatically relevant policy advice is (the level of) uncertainty (see Betz 2013, 215). In N-P, in turn, it's the policy that determines which statements are of interest, or relevance, from the pragmatic perspective and what should be the level of a method's uncertainty in deriving the true acceptance of a statement of interest. Hypotheses in N-P don't need to be "plain" or "unequivocal" but can also state ranges of values (be composite). Still, it's the pragmatic reason which determines a

certain range of interest. It appears to us that from the perspective of policy that N-P's approach is competitive, as having some pragmatic control over the level of uncertainty, instead of being deprived of this, seems to be profitable from the perspective of policy. From the epistemic perspective adopted in this paper, it seems that adopting Betz's proposal is unnecessary and its profitability over N-P's approach becomes less clear. Why should we avoid value-laden scientific conclusions if this value-ladenness present in a form of PDDEP generally brings about indirect epistemic gains and does not necessarily cause direct epistemic loss (as we concluded in C5 and C7 respectively)? Betz's program of avoiding this type of value-ladenness could be found to be an unequivocally epistemically better solution than sticking to PDDEP if PDDEP generally would lead to negative epistemic consequences. We have argued to the contrary. Moreover, hypotheses, when broadened to the extent that they can be accepted beyond any doubt, may turn out to be too broad to be useful in practical applications. Finally, Betz's critique does not fully apply to N-P, in which PDDEP does not entail stating "plain" hypotheses (for they can be defined as ranges of values as well).

The second mentioned approach to avoiding PDDEP involves delegating risk consideration to post-research deliberation. There could then be a first stage, a scientific investigation, with a uniform distribution of errors, and a second stage of deliberation in which decision-makers decide on whether and how to implement outcomes. We argue that if a certain outcome is pragmatically vulnerable (i.e., it is pragmatically important that the outcome not be falsely adopted), the decision-expert committee will want to be more sure (than in the case of a less vulnerable outcome) that this outcome has not been asserted mistakenly before they agree to base their decisions on it. Therefore, this committee will request the scientific community to continue with more stringent scrutiny of the sensitive outcome (let us call it the acceptance of H') to eliminate the possibility of error with a greater force. This can be translated by researchers into the lowering of the risk of the competing

hypothesis (H in this case) being falsely rejected when compared to original standards. The latter possibility is unlikely if the outcome were the opposite of the one obtained. The acceptance of H would not prompt decision-makers to ask for more stringent supplementary scrutiny as to whether the rejection of H' is not mistaken. Thus, in effect, this would be tantamount to the phenomenon of the uneven avoidance of different types of error, which is encapsulated in N-P in the form of setting an uneven importance to errors before research. The conclusion from the above is that the pragmatic, value-laden, and uneven weighing of errors in scientific investigation is a fact which is independent of the choice of the method of investigation (e.g., frequentist or Bayesian), and it seems that this value-ladenness is at least to some extent expected by those who make important decisions based on scientific outcomes and those who financially enable research to be made.

7. Consequences for Replicability

There are many definitions of the terms 'replicability'/'reproducibility' and 'replication'/'reproduction' (Laraway et al., 2019). While replicability is usually related to the potential to reproduce the same result by the same team, reproducibility is related to the reproduction of the same result by a different team (Plesser 2018). Nevertheless, no consensus seems forthcoming on how to measure this potential (Open Science Collaboration 2015). Moreover, when speaking about replicability/reproducibility, one may refer not only to the potential of obtaining the same (experimental) result or conclusion but may also regard the potential of re-running the same experiment (using the same methods) instead (see Plesser 2018). Machery (2020) calls replication a repetition of the same type of experiment, broadly understood, without distinguishing who does this experiment. But some also propose to distinguish conceptual replication (as opposed to direct replication) when speaking of asking

the same type of question but finding a different type of experiment (Schmidt 2009). Observing this variety of meanings of replication and replicability and the disagreement on how these notions should be operationalized, it might be fruitful to conceive of replicability from a broader, multi-faceted perspective.

Thus in this section, we propose a dual approach to understanding replicability as adopted by our investigation. We concentrate on the distinction between methods and results. We speak of experiment replication in the broadest sense (similar to Machery 2020) and result replication in the sense of drawing the same qualitative conclusion (Goodman et al., 2016).[24] Subsequently, we propose to understand replicability from a broader sense as having two complementary and interrelated components: replicability of the result as the direct aspect of replicability, and replicability of the experiment as the indirect aspect of replicability. This is analogous to how we understand the epistemic reliability of N-P in this paper, namely as a broader characteristic that can be described from the position of two complementary and interrelated aspects— both indirect and direct. Accordingly, we propose to treat the influence of PDDEP on experiment replicability as an indirect type of influence while the influence of PDDEP on outcome replicability as a direct type of influence. The term "indirect" not only indicates that the influence considered is not related to a result but, moreover, that this influence is not restricted to affecting a particular experiment to which a test with PDDEP is applied. PDDEP, as applied to one experiment, may affect the replicability of a different experiment that belongs to a certain set of experiments under consideration. In the following subsections, we discuss the indirect and direct type of influence of PDDEP on replicability and how these two types of influence are related. When discussing the effect of PDDEP, we take into account the discrimination of in-test and inter-test PDDEP.

[24] It needs to be stressed that a result of the application of N-P is the acceptance of a hypothesis, not a *p*-value, which itself is subject to certain problems regarding the notion of replication (see e.g. National Academies 2019, 74).

7.1. PDDEP and Experiment Replicability

To consider the influence of in-test and inter-test PDDEP on experiment replicability, an argumentation analogical to the one presented in Section 4 can be applied. Suppose the influence of pragmatic values discussed by us allows one to better dispose of existing, or potential, cognitive resources. In that case, it indirectly positively affects the prospective of repeating one or another research experiment in the future and in this sense, it positively influences replicability potential by enhancing and securing the possibility of the re-conduct of research in the first place. This stands for both inter- and in-test PDDEP.

The positive indirect influence of in-test PDDEP on replicability can be depicted with the help of the following simplified example. A research institution underwent a number of experiments with tests, each concerning a different research topic. In every case, a false rejection of H was considered pragmatically very unfavorable and, if applied to social life, would end with the deprivation of all future research resources. Given the fixed research resources, the institution could neutrally balance error probabilities so that both types of error were fairly possible. In another scenario, with the same experiments and resources, the institution applied in-test PDDEP by securing the I^{st} type of error as almost impossible to commit at some cost of power. Imagine that one of the outcomes is implemented in social life. In the second scenario, there is almost no risk of losing future resources (due to the implementation of a false outcome) which may be used for additional replications of the particular types of experiments conducted thus far. In the first scenario, the risk of committing a pragmatically critical error is higher, and thus, there will also be a higher risk of implementing a critically false outcome; this means an increase of the risk that this same institution will not get further resources and potentially use them for replication of the experiments they have performed. Therefore in-test PDDEP as applied to every experiment increases this institution's potential to replicate these experiments.

We argued in Section 3 and concluded in C5a that inter-test PDDEP positively influences indirect epistemic reliability. This is because of the positive influence of inter-test PDDEP on obtaining epistemic resources, maintaining them, and increasing their usage effectiveness. This gain entails an increase of the potential to perform more experiments. Therefore, inter-test PDDEP increases the potential to replicate all experiments to which inter-test PDDEP was applied. Therefore, inter-test PDDEP has a positive indirect influence on replicability.

This positive influence might be seen as a bit problematic if inter-test PDDEP is applied to a set of experiments concerning the same subject matter. If one assumes that a change in error probability in repeated research is an essential change of the method used, such research might not constitute a replication of an experiment. Suppose the change of error probabilities, or interrelated change of precision, obstructs the idea of direct replications in consecutive replications. In that case, the inter-test PDDEP can still be profitable for performing conceptual replications, which can subsequently be regarded as generalizability tests (see Nosek, Errington 2020).

7.2. PDDEP and Outcome Replicability

The direct influence of in-test PDDEP can also be distinguished and assessed. This requires a further explication of the adopted understanding of outcome replicability. An aspect of N-P's outcome replicability can be the probability of the result of the rejection of H (the acceptance of H^C) in an identical follow-up experiment led by an independent research team (see Miller 2009). Assuming that neither hypothesis is more probable than the other, for the increase of replicability so defined, the avoidance of the IInd type of error is more important. Such a PDDEP that expresses a concentration on the Ist type of error decreases the outcome (the

acceptance of H^C) replicability.[25] But N-P is a symmetric approach: a possible result of the acceptance of H is in N-P equally as statistically a relevant outcome as the acceptance of H^C is. Therefore, the probability of reproducing the result of the acceptance of H should be equally important in the assessment of the influence of PDDEP on outcome replicability. The value of the probability of reproducing the result of the acceptance of H will increase with the decrease of the Ist type of error at the cost of an increase of the IInd type of error. If outcome replicability is understood as a resultant of both values of replicability considered jointly, then in-test PDDEP does not necessarily decrease outcome replicability (probability of replicating an outcome without specifying the type of outcome) when assessed before a possible outcome is obtained.[26] Still, if only the outcome of the acceptance of H^C is to be considered, the application of in-test PDDEP may happen to increase the replicability of such an outcome. This will occur if the more important error to avoid coincides with the error of the IInd type, which is admissible in N-P (Neyman 1971, 4). The conclusion is that assessing the influence of in-test PDDEP on outcome replicability may vary dependently case by case and would require specific calculations every time.

What about the direct influence of inter-test PDDEP? Here, the reasoning from Section 5 concerning the direct influence of inter-test PDDEP on epistemic reliability can be referred to. If the decrease of epistemic reliability—and thus of outcome replicability—in a single testing situation is balanced by an increase of replicability in some other testing situation, then

[25] For example, if both types of error are at the level of 0.15, then this probability will be equal to 0.734. But if the Ist type of error is seen as more important to avoid, at the level of 0.1, at the cost of the increase of the IInd type of error to the level of 0.2, then this probability will equal 0.717. In this example we assumed a more neutral (uninformative) initial probability of a tested hypothesis, but as Lash (2017, 629-630) argues, the influence of a more stringently avoided error of the first kind on an outcome's replicability is more grave in the case of innovative research, where the hypothesis tested has a lower prior probability.

[26] An increase of outcome replicability understood as a resultant of both values would also take place naturally if H was sufficiently more probable than an alternate hypothesis.

the combined direct influence of this inter-test PDDEP on outcome replicability in these two situations considered jointly is not negative. However, if a decrease of outcome replicability in one of these situations is unrelated to the increase of outcome replicability in the other, in the way that spared epistemic resources are not (to be) transferred to a second research situation but may be used in not yet specified future research, then the inter-test PDDEP can be regarded as having a negative influence on outcome replicability. This is because inter-test PDDEP thus lowers the aforementioned replicability in one of the two situations considered but does not increase the other situation's replicability.

The same reasoning can be applied in determining the influence of inter-test PDDEP on not combined outcome replicability in a single research situation: whether this influence is positive or negative depends on which of the two situations is under scrutiny. For these reasons, it can be claimed that the direct influence of inter-test PDDEP on outcome replicability is, in principle, neither negative nor positive, as it turns out to be negative in some cases while positive in others depending on the reference class of the analysis of inter-test PDDEP influence.

7.3. The Relation between the Indirect and Direct Influence of PDDEP on Replicability

We argued in Subsection 7.1. that the indirect influence of PDDEP on replicability is positive: PDDEP indirectly increases replicability by enhancing and securing the possibility of reconducting a research process. This possibility is a necessary condition for the replication of an outcome. Therefore, PDDEP might be considered to contribute to the replicability of a result (the direct aspect of influence) in this specific sense.

However, the argument from Subsection 7.2. indicates that there is no general rule regarding the effect of PDDEP on the replication of a result. There are possible situations in which PDDEP will increase the replicability of experiments but decrease the replication of the

results thereof. This, in turn, means that the influence of PDDEP on replicability as understood in the broad sense proposed here is ambivalent as related to its direct and indirect aspects. This ambivalence indicates that for some cases the ideas of the replicability of the experiment and the replicability of results may be incompatible. Below we provide an example of such incompatibility.

Imagine several new, different modifications of a certain drug are being synthesized by a research institution and tested for toxicity in order to find the one that will replace the old version of the drug in question, which itself became toxic. Assume that the hypothesis tested (in every case) states that the synthesized version is toxic and PDDEP is applied with an emphasis placed on the avoidance of error of the first type (i.e., detecting toxicity is pragmatically more important than detecting non-toxicity) (see Neyman 1950, 263). If the result of such a test is that the synthesized version is toxic, further research would be devoted to testing other modifications, which means posing a new research question and no experiment replication. If the result would be that the tested version is not toxic, then it would be a promising outcome, and the institution might want to run experiment replication on that synthesized version, presumably with the hope that the initial result will be confirmed. Under such a scenario, from two possible results, the one under which replication of the experiment will be of great interest, and the replication of the result hoped for, is the one in which an incorrect acceptance is more pragmatically important to avoid. Setting the importance of error probabilities this way lowers the replicability of this type of result (non-toxicity), if it is a true discovery, by lowering the test's power to detect this true non-toxicity. In other words, PDDEP, in this case, lowers outcome replicability for a situation in which the replication of an experiment would be of interest and the outcome is true. Simultaneously, PDDEP secures the possibility of obtaining the result of non-toxicity on one of the versions to be tested and further replicating the experiment related to testing that version when it is truly a non-toxic

version. This is because PDDEP lowers the risk of falsely asserting the non-toxicity of a toxic version and this lowers the risk of the institution in question being deprived of epistemic resources that could be used to discover the truly non-toxic version and then replicate the experiment related to testing that version.

The result is that a PDDEP-driven increase of an experiment's replicability for the case of a promising outcome (the situation in which replication is desired) that is true takes place at the cost of the decrease of an outcome's replicability for this case. Similarly, an increase of outcome replicability for the case of a desired (and true) discovery (of the non-toxic version) takes place at the cost of not giving privilege to the first type of error, which means an increase of the risk of being deprived of resources needed for the realization of the scenario of performing an experiment with a true discovery as its outcome and the subsequent replication of this experiment.

8. Conclusions

Science is influenced by social preferences to a considerable extent. We claim that this is not only a matter of contingency in selecting research questions (Hacking 1999). The outcomes of scientific research contribute to the "common pool of knowledge" upon which subsequent practical applications crucially depend (Stiglitz, Greenwald 2015, 121). Thus, it can be granted that an important role of science is to also serve as the basis for making the best decisions and actions pragmatically. The N-P solution faces this challenge by implementing pragmatic factors into inferential scientific procedures. This implementation is present in the form of the mechanisms concerning the uneven attribution of risks of error. Neyman and Person's theory identifies these various factors as an inherent part of the research process and, importantly, that their proper methodological consideration can cause the

epistemic reliability and replicability potential entailed by a method to increase. This is far from being one-dimensional and is not always achievable. We have found that the pragmatic value-ladenness of N-P expressed by PDDEP has a twofold influence on N-P's epistemic reliability and replicability. Both in-test and inter-test PDDEP enhances indirect epistemic reliability of N-P and, associated with this aspect, replicability of experiments. The influence of PDDEP on direct epistemic reliability of N-P and related to it replicability of outcomes is ambiguous and depends on the case in question. For some cases, PDDEP can simultaneously enhance and weaken the epistemic reliability and replicability relative to the aspect considered (indirect or direct). These critical cases, presumably, require individual analysis and cautious balancing of epistemic and pragmatic reasons.

While Neyman and Pearson's methodology provides an important complement to pragmatistic epistemology and the currently developing post-Kuhnian theories of science, this methodology shows how the general philosophical stance may be applied in precise methodological solutions which maintain the validity of the epistemic perspective. N-P helps one control and make better epistemic use of irremovable pragmatic factors, which can thus be epistemically beneficial. Moreover, the pragmatic value-laden weighing of errors is not merely an inherent peculiarity of N-P that could be eliminated by scientific inference by using some other methodology that relegates this aspect to the stage of making value-laden decisions based on value-free scientific judgments. The fact that the pragmatic value-ladenness of N-P is indirectly advantageous when considered from the perspectives of epistemic reliability and replicability, and that this value-ladenness' direct influence is not in principle disadvantageous to these aspects strengthens the argument from inductive risk and partially undermines the value-free ideal of science. This fact indicates that for epistemic reasons, scientists should not attempt to always minimize the influence of pragmatic values on scientific reasoning but rather make use of it in an informed, moderate way.

References

Amrhein, Valentin, Korner-Nievergelt, Franzi, Roth, Tobias. 2017. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ* 5: e3544.

Baumgaertner, Bert, Holthuijzen, Wieteke. 2017. On nonepistemic values in conservation biology. *Conservation Biology* 31: 48–55.

Betz, Gregor. 2013. In defence of the value-free ideal. *European Journal for the Philosophy of Science* 2: 207–220.

Bickel, Peter, J., Doksum, Kjell, A. 2001. *Mathematical Statistics. Basic Ideas and Selected Topics.* Vol. 1. 2nd ed. Ney Jersey: Prentice Hall.

Chiffi, Daniele. Pietarinen, Ahti-Veikko. 2019. Risk and Values in Science: A Peircean View. *Axiomathes* 29: 329–346.

Collins, Harry, M., and Evans, Robert. 2002. The third wave of science studies: Studies of expertise and experience. *Social Studies of Science* 32: 235–296.

David, Marian. 2001. Truth as the Epistemic Goal. 2001. In *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, ed. M. Steup, 151–169. Oxford: Oxford University Press.

Elliott, Kevin, C., Richards, Ted, (eds). 2017. *Exploring inductive risk: case studies of values in science*. Oxford: Oxford University Press.

Fiedler, K., Kutzner, F., & Krueger, J. I. 2012. The Long Way From α-Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate. *Perspectives on Psychological Science*, 7(6): 661–669.

Forster, Malcolm, R., Sober, Elliott. 2011. AIC Scores as Evidence: A Bayesian Interpretation. In *Handbook of the Philosophy of Science.Vol. 7:*

Philosophy of Statistics, ed. D.M. Gabbay, P. Thagard, J. Woods, P.S. Bandyopadhyay, and M.R. Forster, 535–549. Amsterdam: Elsevier.

Gigerenzer, Gerd., Brighton, Henry. 2009. Homo Heuristicus: Why Biased Minds Make Better Inferences. Topics in Cognitive Science, 1: 107-143.

Goldman, Alvin I. 2008. Immediate justification and process reliabilism. In Epistemology*: New Essays*, ed. Q. Smith, 63–82. New York: Oxford University Press.

Goodman, Steven N., Fanelli, Daniele., Ioannidis, John P.A. 2016. What does research reproducibility mean? *Science Translational Medicine 01 Jun 2016* (8) 341: pp. 341 ps 12.

Grant, Bob. 2012. Science's reproducibility problem. *The Scientist*, 18 December 2012.

Hacking, Ian. 1999. *The social construction of what?* Cambridge: Harvard University Press.

Ioannidis, John, P.A. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2 (8): e124.John, Stephen. 2015. Inductive risk and the contexts of communication. *Synthese* 192: 79–96.

Johnson, Valen E. 2013. Revised standards for statistical evidence. In *Proceedings of the National Academy of Sciences Nov 2013*, 110 (48): 19313-19317.

Kaivanto, Kim, Steel, Daniel. 2019. Adjusting Inferential Thresholds to Reflect Nonepistemic Values. *Philosophy of Science* 86, 2: 255-285.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kitcher, Philip. 2011. *Science in a democratic society*. New York: Prometheus Books.

Kubiak, Adam, Kawalec, Paweł, Kiersztyn, Adam. 2020. Epistemic Reliability of Neyman-Pearson Hypothesis Testing and Its Pragmatic Value-Laden Unequal Error Risk Setting. Preprint. http://philsci-archive.pitt.edu/id/eprint/18594

Laraway, Sean, Snycerski, Susan, Pradhan, Sean, Huitema, Bradley E. 2019. An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis. *Perspect Behav Sci* 42:33–57.

Lash, Timothy L. 2017. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *American Journal of Epidemiology*, 186 (6): 627–635.

Laudan, Larry. 2004. The Epistemic, the Cognitive, and the Social. In *Science, Values, and Objectivity*, ed. Peter Machamer and Gereon Wolters, 14–23. Pittsburgh: University of Pittsburgh Press.

LeBel, Etienne P., Campbell, Lorne, Loving, Timothy J. 2017. "Benefits of open and high-powered research outweigh costs." *Journal of Personality and Social Psychology*, 113(2), 230–243.

Machery, Edouard. 2020. "What Is a Replication?" *Philosophy of Science*, 87, 545-567.

Mayo, Deborah, Spanos, Aris. 2006. Severe Testing as a Basic Concept in Neyman-Pearson Philosophy of Induction. *The British Journal for Philosophy of Science*, 57: 323–357.Miller, Jeff. 2009. What is the probability of replicating a statistically significant effect? *Psychon Bull Rev*, 16 (4): 617–640.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press.

Neyman, Jerzy. 1950. *First Course in Probability and Statistics*. New York: Henry Holt and Co.

Neyman, Jerzy. 1952. *Lectures and conferences on mathematical statistics and probability*. Washington: U.S. Department of Agriculture.

Neyman, Jerzy. 1957. "'Inductive Behavior' as a Basic Concept of Philosophy of Science." *Revue De L'Institut International De Statistique* 25, 1/3: 7–22.

Neyman, Jerzy. 1971. Foundations of Behavioral Statistics. In *Foundations of Statistical Inference*, ed. V. P. Godambe and D. A. Sprott. Toronto: Holt, Rinehart and Winston.

Neyman, Jerzy. 1977. Frequentist probability and frequentist statistics. *Synthese* 36: 97–131.

Neyman, Jerzy, Pearson, Egon, S. 1928. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I *Biometrika* 20A: 175–240.

Neyman, Jerzy, Pearson, Egon, S. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A* 231: 289-337.

Neyman, Jerzy, Pearson, Egon, S. 1936. Contribution to the theory of testing statistical hypotheses." *Statistical Research. Memoirs* 1:1-37. Reprinted in: *Joint Statistical Papers. J. Neyman and E. S. Pearson.* Cambridge 1967, 203-239.

Nosek Brian, A., Errington, Timothy M. 2020. What is replication? *PLoS Biol* 18(3): e3000691.

Open Science Collaboration. PSYCHOLOGY. 2015. Estimating the reproducibility of psychological science. *Science*. 2015; 349(6251): aac4716.

Page, Scott E. 2007. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University.Rochefort-Maranda, Guillaume. 2013 Statistical Power and P-values: An Epistemic Interpretation Without Power Approach Paradoxes. Manuscript. http://philsci-archive.pitt.edu/14220/

Romeijn, Jan-Willem. 2017. Philosophy of Statistics. In *The Stanford Encyclopedia of Philosophy (Spring 2017 Edition)*, ed. Edward N. Zalta. https://plato.stanford.edu/entries/statistics/

Plesser, Hans E. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology." *Front. Neuroinform*. 11: 76.

Royall, Richard. 2007. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall.

Rubin, Mark. 2019. What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*. https://doi.org/10.1007/s11229-019-02433-0

Rudner, Richard. 1953. The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20: 1–6.

Schmidt, Stefan. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2): 90–100.Spielman, Stephen. 1973. A refutation of the Neyman-Pearson theory of testing. *British Journal for the Philosophy of Science* 24(3): 201–222.Stahel Werner A. 2016. Statistical issues in reproducibility. In Atmanspacher H., Maasen S., eds. Reproducibility: principles, problems, practices, and prospects, 87-114. Hoboken: Wiley.

Steel, Daniel. 2010. Epistemic Values and the Argument from Inductive Risk. *Philosophy of Science* 77: 14–34.

Stiglitz, Joseph, E., Greenwald, Bruce, C. 2015. *Creating a learning society: a new approach to growth, development, and social progress*. New York: Columbia University Press.

Wald, Abraham. 1950. *Statistical Decision Functions*. New York: John Wiley and Sons.

Wimsatt, William, C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.